TECHNISCHE UNIVERSITÄT DARMSTADT

LARGE-SCALE MULTI-AGENT REINFORCEMENT LEARNING VIA MEAN FIELD GAMES

Vom Fachbereich Elektrotechnik und Informationstechnik der
TECHNISCHEN UNIVERSITÄT DARMSTADT

zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von

KAI CUI, M.SC.
geboren am 5. April 1996 in Frankfurt am Main.

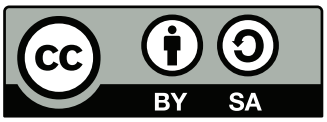
Referent: Prof. Dr. techn. Heinz Koeppl
Korreferent: Prof. Dr. Mathieu Laurière

Tag der Einreichung: 24. Juni 2024
Tag der mündlichen Prüfung: 17. Oktober 2024

D17
Darmstadt 2024

Die Arbeit von Kai Cui wurde durch die LOEWE Initiative des Landes Hessen im Rahmen des LOEWE-Zentrums emergenCITY gefördert.

Cui, Kai: *Large-Scale Multi-Agent Reinforcement Learning via Mean Field Games*
Darmstadt, Technische Universität Darmstadt
Jahr der Veröffentlichung der Dissertation auf TUprints: 2024
Tag der mündlichen Prüfung: 17. Oktober 2024



KURZFASSUNG

In dieser Dissertation diskutieren wir mathematisch-rigorose Multiagenten-Lernmodelle basierend auf Mean Field Games (MFG) und Mean Field Control (MFC). Dynamische Multiagenten-Kontrollprobleme und ihre spieltheoretischen Analoga finden in der Praxis viele Anwendungen, können aber schwer auf viele Agenten hochskaliert werden. MFGs und MFC ermöglichen die skalierbare Modellierung großer dynamischer Multiagentenkontroll- und Spielprobleme. Im Wesentlichen werden hier die Interaktionen zwischen unendlich vielen homogenen Agenten auf ihre anonyme Verteilung – das so genannte Mean Field – reduziert. Dies vereinfacht viele praktische Probleme auf die Betrachtung eines einzigen repräsentativen Agenten und – durch das Gesetz der großen Zahlen – dessen Wahrscheinlichkeitsverteilung. In dieser Arbeit stellen wir verschiedene neue Lernalgorithmen und theoretische Modelle in MFGs und MFC vor. Wir addressieren existierende algorithmische Limitationen, und erweitern außerdem MFGs und MFC über ihre Beschränkung auf (i) schwach interagierende Agenten, (ii) allwissende und rationale Agenten oder (iii) Homogenität der Agenten hinaus. Abschließend werden einige praktische Anwendungen kurz betrachtet, um die Nützlichkeit der entwickelten Algorithmen zu demonstrieren.

Zunächst betrachten wir den selbstinteressierten Fall der MFGs. Dort zeigen wir, dass im einfachsten Fall endlicher MFGs die bestehenden Algorithmen starke Einschränkungen haben können. Insbesondere zeigen wir, dass die übliche Annahme kontraktiver Fixpunktoperatoren schwer zu erfüllen ist. Anschließend werden approximative Lernalgorithmen für MFGs vorgestellt und analysiert, die auf Regularisierung basieren und einen Kompromiss zwischen Optimalität und Konvergenz ermöglichen. Weiter erweitern wir die Ergebnisse zu MFGs auf Graphen und Hypergraphen, um die Beschreibungsfähigkeit der MFGs zu erhöhen und die Einschränkung der Homogenität zu umgehen. Schließlich übertragen wir die Ergebnisse auch auf die Präsenz sowohl stark interagierender als auch vieler schwach interagierender Agenten, um Skalierbarkeit für Fälle zu erreichen, in denen einige Agenten nicht unter die Mean-Field-Approximation fallen.

Zweitens untersuchen wir den kooperativen Fall des MFC. Zunächst betrachten wir eine Erweiterung auf Umweltzustände unter der vereinfachenden Annahme statischer Mean Fields. Die annähernde Optimalität einer MFC-Lösung über Lösungen des endlichen Problems wird gezeigt. Anschließend und allgemeiner wird MFC auf stark interagierende Agenten ausgedehnt, ähnlich wie im MFG-Szenario. Unsere letzte Erweiterung berücksichtigt partielle Informationsstruktur, bei der dezentralisierte Agenten auf der Grundlage begrenzter, verfügbarer Informationen handeln. Hier wird eine Optimierung über Lipschitz-Klassen von Strategien eingeführt. Für die beiden letztgenannten Szenarien erhalten wir außerdem Garantien für die Approximation der Strategiegradienten. Die Modelle werden theoretisch verifiziert, indem eine approximative Optimalität der MFC-Strategien gezeigt wird, sowie experimentell verifiziert, indem eine Performanz demonstriert wird, die im Vergleich zu modernsten Multiagenten-Verstärkungslernalgorithmen gleichwertig oder besser ist.

Abschließend werden einige mögliche Anwendungen von MFGs und MFC in Szenarien mit großen Agentenpopulationen untersucht und erläutert. Dazu gehören Anwendungen in den Bereichen verteiltes Rechnen, cyber-physische Systeme, autonome Mobilität und Routing, sowie Natur- und Sozialwissenschaften. Wir werfen auch einen genaueren Blick auf zwei spezielle Anwendungen der UAV-Schwarmkontrolle und des Edge Computing. Im ersten Fall betrachten wir die Auswirkungen der Kollisionsvermeidung für MFC mit physischen Roboterschwärmen. Im zweiten Fall vergleichen wir die MFG- und MFC-Ergebnisse für die Auslagerung von Berechnungen.

Insgesamt untersuchen wir in dieser Arbeit die Eignung von MFG und MFC Methoden für großskaliges Multiagenten-Verstärkungslernen. Wir formulieren neue Lernmethoden und theoretische Approximationsmodelle, und untersuchen einige Anwendungen. Im Großen und Ganzen stellen wir fest, dass MFGs und MFC erfolgreich für die Analyse großer kontroll- und spieltheoretischer Probleme eingesetzt werden können, und zwar mit hoher Allgemeinheit und besserer Leistung als einige existierende Lösungen.

ABSTRACT

In this dissertation, we discuss the mathematically rigorous multi-agent reinforcement learning frameworks of mean field games (MFG) and mean field control (MFC). Dynamical multi-agent control problems and their game-theoretic counterparts find many applications in practice, but can be difficult to scale to many agents. MFGs and MFC allow the tractable modeling of large-scale dynamical multi-agent control and game problems. In essence, the idea is to reduce interaction between infinitely many homogeneous agents to their anonymous distribution – the so-called mean field. This reduces many practical problems to considering a single representative agent and – by the law of large numbers – its probability law. In this thesis, we present various novel learning algorithms and theoretical frameworks of MFGs and MFC. We address existing algorithmic limitations, and also extend MFGs and MFC beyond their restriction to (i) weakly-interacting agents, (ii) all-knowing and rational agents, or (iii) homogeneity of agents. Lastly, some practical applications are briefly considered to demonstrate the usefulness of our developed algorithms.

Firstly, we consider the competitive case of MFGs. There, we show that in the simplest case of finite MFGs, existing algorithms are strongly limited in their generality. In particular, the common assumption of contractive fixed-point operators is shown to be difficult to fulfill. We then contribute and analyze approximate learning algorithms for MFGs based on regularization, which allows for a trade-off between approximation and tractability. We then proceed to extend results to MFGs on graphs and hypergraphs, in order to increase the descriptiveness of MFGs and ameliorate the restriction of homogeneity. Lastly, we also extend towards the presence of both strongly interacting and many weakly-interacting agents, in order to obtain tractability for cases where some agents do not fall under the mean field approximation.

Secondly, we investigate cooperative MFC. Initially, we consider an extension to environmental states under a simplifying assumption of static mean fields. Approximate optimality of an MFC solution is shown over any finite agent solution. More generally, we proceed to extend MFC to strongly interacting agents, similar to the MFG scenario. Our final extension considers partial observability, where decentralized agents act only upon available information. Here, a framework optimizing over Lipschitz classes of policies is introduced. We obtain policy gradient approximation guarantees for the latter two settings. The frameworks are verified theoretically by showing approximate optimality of MFC, and experimentally by demonstrating performance comparable or superior to state-of-the-art multi-agent reinforcement learning algorithms.

Finally, we briefly explore some potential applications of MFGs and MFC in scenarios with large populations of agents. We survey applications in distributed computing, cyber-physical systems, autonomous mobility and routing, as well as natural and social sciences. We also take a closer look at two particular applications in UAV swarm control and edge computing. In the former, we consider the effect of collision avoidance as an additional constraint for MFC in embodied robot swarms. In the latter, we compare MFG and MFC results for a computational offloading scenario.

Overall, in this thesis we investigate the suitability of methods based on MFC and MFC for large-scale tractable multi-agent reinforcement learning. We contribute novel learning methods and theoretical approximation frameworks, as well as study some applications. On the whole, we find that MFGs and MFC can successfully be applied to analyze large-scale control and games, with high generality and outperforming some state-of-the-art solutions.

PUBLICATIONS

The following publications were produced during the course of the doctoral candidacy:

PEER-REVIEWED CONFERENCE PROCEEDINGS

[1] K. Cui, C. Fabian, and H. Koeppl, "Major-minor mean field multi-agent reinforcement learning", *Proc. ICML*, 2024.

[2] K. Cui, S. Hauck, C. Fabian, and H. Koeppl, "Learning decentralized partially observable mean field control for artificial collective behavior", in *Proc. ICLR*, 2024, pp. 1–40.

[3] K. Cui, G. Dayanıklı, M. Laurière, M. Geist, O. Pietquin, and H. Koeppl, "Learning discrete-time major-minor mean field games", in *Proc. AAAI*, vol. 38, 2024, pp. 9616–9625.

[4] K. Cui, L. Baumgärtner, M. B. Yilmaz, M. Li, C. Fabian, B. Becker, L. Xiang, M. Bauer, and H. Koeppl, "UAV swarms for joint data ferrying and dynamic cell coverage via optimal transport descent and quadratic assignment", in *Proc. LCN*, 2023, pp. 1–8.

[5] K. Cui, M. Li, C. Fabian, and H. Koeppl, "Scalable task-driven robotic swarm control via collision avoidance and learning mean-field control", in *Proc. ICRA*, 2023, pp. 1192–1199.

[6] K. Cui, M. B. Yilmaz, A. Tahir, A. Klein, and H. Koeppl, "Optimal offloading strategies for edge-computing via mean-field games and control", in *Proc. GLOBECOM*, 2022, pp. 976–981.

[7] K. Cui and H. Koeppl, "Learning graphon mean field games and approximate Nash equilibria", in *Proc. ICLR*, 2022, pp. 1–31.

[8] K. Cui, A. Tahir, M. Sinzger, and H. Koeppl, "Discrete-time mean field control with environment states", in *Proc. CDC*, 2021, pp. 5239–5246.

[9] K. Cui and H. Koeppl, "Approximately solving mean field games via entropy-regularized deep reinforcement learning", in *Proc. AISTATS*, 2021, pp. 1909–1917.

[10] A. Tahir, K. Cui, A. Rizk, and H. Koeppl, "Collaborative optimization of the age of information under partial observability", *to appear in IFIP Networking 2024, arXiv:2312.12977*, 2023.

[11] A. K. Sreedhara, D. Padala, S. Mahesh, K. Cui, M. Li, and H. Koeppl, "Optimal collaborative transportation for under-capacitated vehicle routing problems using aerial drone swarms", in *Proc. ICRA*, IEEE, 2024, pp. 8401–8407.

[12] M. Li, K. Cui, and H. Koeppl, "A modular aerial system based on homogeneous quadrotors with fault-tolerant control", pp. 8408–8414, 2024.

[13] C. Fabian, K. Cui, and H. Koeppl, "Learning mean field games on sparse graphs: A hybrid graphex approach", in *Proc. ICLR*, 2024, pp. 1–39.

[14] C. Fabian, K. Cui, and H. Koeppl, "Learning sparse graphon mean field games", in *Proc. AISTATS*, 2023, pp. 4486–4514.

[15] A. Tahir, K. Cui, and H. Koeppl, "Learning mean-field control for delayed information load balancing in large queuing systems", in *Proc. ICPP*, 2022, pp. 1–11.

[16] R. Ourari, K. Cui, A. Elshamanhory, and H. Koeppl, "Nearest-neighbor-based collision avoidance for quadrotors via reinforcement learning", in *Proc. ICRA*, 2022, pp. 293–300.

PEER-REVIEWED JOURNAL ARTICLES

[17] K. Cui, W. R. KhudaBukhsh, and H. Koeppl, "Hypergraphon mean field games", *Chaos*, vol. 32, no. 11, 2022.

[18] K. Cui, W. R. KhudaBukhsh, and H. Koeppl, "Motif-based mean-field approximation of interacting particles on clustered networks", *Phys. Rev. E*, vol. 105, no. 4, p. L042301, 2022.

[19] C. Fabian, K. Cui, and H. Koeppl, "Mean field games on weighted and directed graphs via colored digraphons", *IEEE Control Syst. Lett.*, vol. 7, pp. 877–882, 2022.

SUBMITTED / PREPRINTS

[20] K. Cui, A. Tahir, G. Ekinci, A. Elshamanhory, Y. Eich, M. Li, and H. Koeppl, "A survey on large-population systems and scalable multi-agent reinforcement learning", *in preparation for AI Review, arXiv:2209.03859*, 2022.

[21] A. Tahir, K. Cui, and H. Koeppl, "Sparse mean field load balancing in large localized queueing systems", *submitted to MobiHoc 2024, arXiv:2312.12973*, 2023.

ACKNOWLEDGMENTS

This thesis is the result of multiple years of work, during which I had the pleasure to work with many exceptional colleagues. Despite the COVID-19 pandemic during the first years of this thesis, I am happy to have visited various conferences and to have met many great people, in both virtual and in-person conferences in the latter half of my doctoral candidacy.

First and foremost, I would like to thank my supervisor Prof. Heinz Koeppl for providing great guidance and important topics of research during my work on this thesis. He positively influenced my research interests and made them productive. I am especially thankful for having the opportunity to freely pursue my work on mean field games and to be able to collaborate with many clever people of diverse backgrounds from all over the world.

In particular, I would also like to thank all of my great collaborators, without which many works would not have happened: Akash Kopparam Sreedhara, Ahmed Elshamanhory, Amr Rizk, Anam Tahir, Anja Klein, Bastian Alt, Benjamin Becker, Christian Fabian, Deepesh Padala, Gizem Ekinci, Gökçe Dayanıklı, Lars Baumgaertner, Lin Xiang, Mark Sinzger, Mathieu Laurière, Matthieu Geist, Maximilian Bauer, Mengguang Li, Mustafa Burak Yilmaz, Olivier Pietquin, Sascha Hauck, Shashank Mahesh, Wasiur R. KhudaBuhksh and Yannick Eich.

Further, I want to thank everyone else I met in the Self-Organizing Systems lab (formerly Bioinspired Communication Systems) for maintaining a positive working atmosphere and providing a variety of inputs: Alina Kuzembayeva, Anja Engel, Bin Ji, Christian Wildner, Christiane Hübner, Christine Cramer, Christoph Reich, Dominik Linzner, Eike Mentzendorff, Ekaterina Solyus, Erik Kubaczka, Felix Reinhardt, Fengyu Cai, François Lehr, Gamze Dogali, Hongfei Liu, Irem Ergenlioglu, Jacob Christian Mejlsted, Jérémie Marlhens, Julia Detzer, Klaus-Dieter Voss, Maik Molderings, Maleen Hanst, Markus Baier, Matthias Schultheis, Maximilian Gehri, Megan Bailey, Melanie Mikosch-Wersching, Nicolai Engelmann, Nikita Kruk, Özdemir Cetin, Philipp Fröhlich, Sandip Saha, Sebastian Wirth, Sikun Yang, Sofia Startceva, Stanislav Stepaniuk, Tim Prangemeier and Yujie Zhong.

Moreover, I would like to thank everyone in the overarching emergenCITY project – too many to list at this point – for providing another great working environment with focus on important practical applications. The collaborative work on demonstrators and presentations, the many joint events and the technical presentations of everyone have broadened my horizon beyond what one typically experiences inside a single group.

Finally, I want to thank my friends and family for their unlimited support!

Darmstadt, June 24, 2024

Contents

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

The study of Mean Field Games (MFGs) considers games with an infinitude of agents, each of which acts independently and in accordance with its own interests. Closely related, the recent area of Mean Field Control (MFC) instead investigates the case of full agent cooperation, in order to maximize a single global objective. Historically, MFGs were pioneered by [22, 23] in the context of controlled stochastic differential equations, and were since extended to discrete-time and learning literature [24, 25]. In essence, the interactions between agents are decomposed to the interaction between any single agent and the mass of all other infinitely many agents – the Mean Field (MF). See also Figure 1.1 for a visualization. Despite considering infinitely many agents, MFGs and MFC can be more conducive to analysis or learning than dynamic games or control with a large finite number of agents, since the complexity no longer scales with the number of agents. As the infinitude of agents approximates systems with sufficiently many agents, the tractability motivates MFGs and MFC, since large-scale games and control are useful in many applications mentioned hereafter.

In this thesis, we consider novel learning algorithms and theoretical frameworks for solving dynamic games and control problems using both MFGs and MFC. To be precise, learning here refers to both the iterative finding of game-theoretic Nash equilibria in MFGs, as well as sample-based finding of optimal collaborative behavior via MFC. Our results allow scaling up Multi-Agent Reinforcement Learning (MARL) to arbitrarily high numbers of agents in highly general scenarios, which is verified and discussed for various realistic applications. First however, we expand upon the motivation for learning MFGs and MFC, and present the structure of the thesis with its research questions.

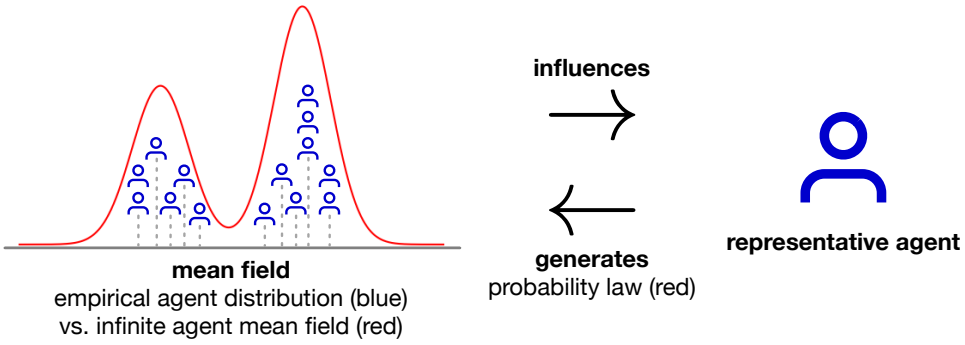


FIGURE 1.1: MFG and MFC interaction model. The interaction between agents is given by an anonymous interaction between any single agent and the MF distribution of all agents. In the infinite agent limit, the MF is replaced by the probability law of any representative agent by law of large numbers.

1.1 MOTIVATION AND RESEARCH QUESTIONS

The motivation of this thesis is the learning of control or decision-making in large-population systems, which has a wide range of applications. In recent years, sequential decision-making via Reinforcement Learning (RL) [26] has prominently found application in a variety of areas, including but not limited to robotics [27] with many examples such as autonomous cars [28], stratospheric balloons [29] or teams of Unmanned Aerial Vehicles (UAVs) [30–32], and also highly complex strategic games [33–36], economics and finance scenarios [37, 38], as well as more recently Large Language Models (LLMs) [39].

While standard single-agent RL continues to be an important and active area of research, many practical applications include more than one agent or decision-maker. As a result, one must also distinguish between different problem scenarios to consider. These multi-agent control problems are typically considered by the extended subject area of so-called MARL [40]. In general, one can consider a multitude of settings, ranging from fully competitive two-player zero-sum games to fully cooperative decentralized control problems. And although MARL approaches are sometimes applicable due to their generality – especially for systems with a few to a dozen agents [36, 41, 42] – their application often remains difficult due to various challenges. For example, the precise definition of learning scenario, simultaneous learning of multiple non-stationary agents, and common partial information structure in many-agent systems all provide difficulty in finding a general solution to MARL. See also many extensive surveys [40, 43, 44].

The main reason of existence for MFGs and MFC in learning literature is given by the notorious intractability challenge of general multi-agent control, both in the competitive scenario of self-interested agents [45], and in the cooperative scenario where agents must oftentimes coordinate, e.g., under partial information [46]. As a result, in general it becomes difficult to solve all problems with more than a few agents. MFG and MFC methods avoid the otherwise intractable direct optimization of the finite games and control problems while still remaining a general class of problems. This is done by seeing that the empirical distribution of an arbitrarily large number of homogeneous agents converges to a limiting, deterministic marginal state distribution – the MF – by the Law of Large Numbers (LLN). The solution of limiting MFGs and MFC becomes easier, and rigorously approximates an optimal solution in the difficult but finite large-scale problem. See also Figure 1.2.

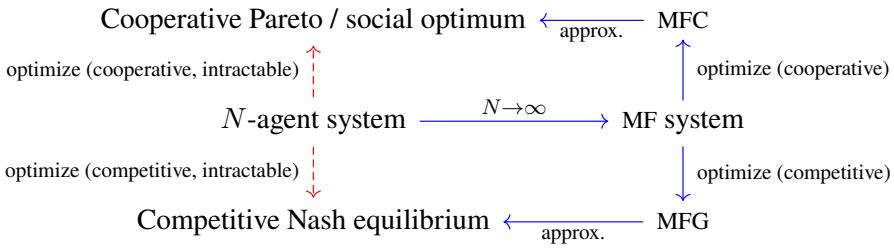


FIGURE 1.2: Pictorial scheme of approximation for MFGs and MFC. The finite N-agent system is first approximated by a MF system, which is then solved through learning algorithms, giving an approximately optimal solution in sufficiently large finite systems. The difficult finite problem is thereby circumvented.

In fact, despite considering only a subset of all possible problems, the MF model assumptions are general to a certain degree, as applications of MFGs or MFC are manifold. Even prior to considering controlled systems with defined agent objectives, MF theory has long found application in as diverse branches of science as statistical physics [47], chemistry [48], epidemiology [49], computer science

[50] and social science [51] through the tractable analysis of interacting particle systems on complex networks [52]. With the advent of MFGs, the idea of controlled MF systems has led to the application of both cooperative MFC and competitive MFGs, e.g. in smart heating [53], traffic engineering [54, 55], large-scale batch processing [56, 57], peer-to-peer networks [58], epidemics [59], or crowds of people [60] to mention a few examples.

For instance, we could consider scenarios from automated driving, epidemics control or finance as depicted in Figure 1.3. Automated driving or driving assistance such as Google Maps could offer the possibility to engineer traffic in a socially optimal manner, or to analyze the behavior of many uncontrolled, self-interested vehicles via MFGs [54, 61]. Meanwhile, in epidemics control such as for COVID-19, one could use MFGs to guide the design of optimal incentives for many rational, self-interested people [62]. Lastly, in finance, the usage of MFG models can guide financial choices such as optimal portfolio liquidation in a large market with many other rational agents [63].

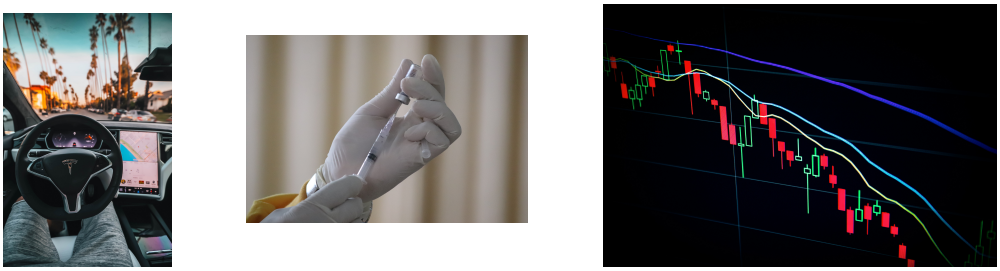


FIGURE 1.3: Many-agent applications, ranging from autonomous driving to epidemics control to finance.

Many more applications can be found in surveys of MFG applications for economics or finance [64] and engineering [65], see also our survey [20]. We will further consider some applications to UAV swarms and networking later in this thesis. Still, MFG *learning algorithms* remain limited and require certain assumptions such as contractivity or monotonicity, while in MFC the algorithms are not analyzed as MARL algorithms on finite multi-agent systems. In addition to algorithmic limitations, standard MFG and MFC *models* restrict the general space of scenarios that can be modelled, since agents must be

- weakly-interacting, i.e. each agent alone has only negligible effect on any other agent;
- all-knowing and rational, such that the agents know e.g. their true own state and the initial MF in the system (and by deterministic propagation, possibly also the MF at all other times);
- homogeneous, in the sense that all agents interact with all other agents in the same way, without additional structured interactions or regularity to consider.

However, in practice, learning algorithms should also handle applications where standard assumptions are not fulfilled. Further, (i) agents are not always weakly-interacting, such as under the presence of a few agents that affect all other agents directly; (ii) agents cannot always know their own true state or the initial MF in the system, especially in large-scale decentralized systems where such coordination becomes difficult; and (iii) agents may not be completely exchangeable, but instead could be of certain types or interconnected according to some graph structure. For this reason, the preceding motivation leads us to the following Research Questions (RQs):

I First, how can we perform tractable MFG and MFC learning to begin with?

II Second, is it possible to overcome the aforementioned limitations of MF models?

III Third, where and how can MFGs and MFC potentially be applied in practice?

1.2 OVERVIEW AND CONTRIBUTION

In this thesis, we develop theory and algorithms in the intersection of MFGs and MARLs. We split the discussion of RQs I and II into the competitive MFG case and the cooperative MFC case. RQ III is addressed by experiments on realistic problems throughout the thesis, and explicitly at the end. Our contributions can be briefly summarized as:

1. In competitive MFGs, we provide approximate learning algorithms for cases where existing algorithms have no guarantee of convergence. Existing algorithms are shown to fail in most finite MFGs, and our regularized approaches succeed against previous approaches.
2. We extend MFGs and associated learning frameworks to graphs and more general agents, in order to improve the generality of the framework. Our results ameliorate the strong usual assumption of many globally weakly-interacting agents, to incorporate interaction through neighborhoods and few more generally-interacting agents. The equations are also extended to hypergraph structures for non-pairwise interactions.
3. In cooperative MFC, we similarly extend towards general agents, and also formulate first results towards (i) policy gradient approximations in finite MFC systems, and (ii) the optimality of MFC solutions over heterogeneous policies in the static case. In a predecessor setting with environment states and static MF, the sufficiency of identical policy sharing between agents is verified.
4. As our last theoretical and learning contribution, we consider the first discrete-time MFC learning framework under partial information, in order to learn collective behavior as potentially simple local interaction rules fulfilling a global objective by emergence. The results are verified by learning common swarming behavior in Kuramoto or Vicsek models.
5. Finally, we briefly survey general application areas that may profit from MF-based solutions, as well as challenges in the fields of robotic swarms and communication networks. We combine MFC with collision avoidance for realistic large-scale control of UAV swarms, and consider MFG and MFC-based resource allocation problems in edge computing.

A list of all publications performed during the course of the thesis can also be found on Pages vii to viii. Note that we do not discuss all the listed publications in full detail in this thesis, but may give some brief references in the following, whenever appropriate.

Overall, the aforementioned contributions can be organized as shown in Table 1.1, and are presented in this thesis according to the following detailed structure:

TABLE 1.1: An overview of how we address learning and MF limitations.

Sec. / Ch.	Refs.	*Learning Framework* (RQ I)	*Model Generality* (RQ II)	*Case*
Sec. 3.1	[9]	FPI hardness, optimality trade-off	–	MFG
Sec. 3.2	[7]	reduction to MFG & guarantees	graph interaction	MFG
Sec. 3.3	[17]	as above	higher-order interaction	MFG
Sec. 3.4	[3]	extension of fictitious play	strong interaction	MFG
Sec. 4.1	[8]	MFC MDP policy gradients	environmental states	MFC
Sec. 4.2	[1]	as above + MARL approximation	strong interaction	MFC
Sec. 4.3	[2]	as above + MARL approximation	partial information	MFC
Ch. 5	[5, 6, 20]	– (addresses RQ III)	– (addresses RQ III)	both

CHAPTER 2. In this initial chapter, we give a background on optimal control and RL, game theory and multi-agent RL, as well as scalable models of standard MFGs and MFC. We recap the primary mathematical models for the above, together with some of their basic properties and well-known algorithms. For the interested and unfamiliar reader, we also give introductory references.

CHAPTER 3. In this chapter, we first study the competitive case of MFGs as a framework for tractable equilibrium learning in large-scale games. Here, we mainly focus on the general setting of evolutive MFGs, where agents are endowed with dynamic states and the MF evolves over time.

To begin, we consider the simplest setting of MFGs, i.e., finite MFGs with finite state and action spaces, that are played over a finite discrete time horizon. While the majority of literature either considers monotonicity conditions or assumes contraction in the fixed point iteration, we show that even in our simplest setting, the contraction condition cannot be fulfilled in non-trivial problems. As an alternative, we propose methods for finding approximate equilibria by employing entropy regularization and related Boltzmann policy schemes. Iterating relative entropy regularization can further improve solutions, and deep techniques with particle filters allow solutions in otherwise intractable settings. Section 3.1 is based on the conference publication

[9] K. Cui and H. Koeppl, "Approximately solving mean field games via entropy-regularized deep reinforcement learning", in *Proc. AISTATS*, 2021, pp. 1909–1917.

In the following Section 3.2, we extend MFGs to non-homogeneously connected and interacting agents. In particular, we assume that agents are connected and interact according to a dense graph. In order to analyze MFGs on graphs, graphons from graph limit theory are used to obtain Graphon Mean Field Games (GMFGs). We give one of the first discrete-time GMFG models together with theoretical foundations and learning algorithms. A reduction to standard MFGs is proposed and its error is analyzed. The work in Section 3.2 is based on the conference publication

[7] K. Cui and H. Koeppl, "Learning graphon mean field games and approximate Nash equilibria", in *Proc. ICLR*, 2022, pp. 1–31.

We go one step further and consider higher-order interaction through hypergraphs in Section 3.3. The theoretical and algorithmic framework is extended from GMFGs, and allows for consideration of effects such as cliques in social networks. We have since also extended the work on GMFGs in collaborations [13, 14, 19] towards increasingly sparse and weighted or directed graphs. In the separate work [18], we also consider some heuristic motif-based approximations in the continuous-time and more difficult bounded-degree graph setting. The above references are only discussed briefly at the end of this chapter. The extended hypergraph MFG is based on the journal publication

[17] K. Cui, W. R. KhudaBukhsh, and H. Koeppl, "Hypergraphon mean field games", *Chaos*, vol. 32, no. 11, 2022.

Lastly, addressing both weak interaction and homogeneity of agents, in Section 3.4 we extend MFGs also towards "major" agents, through the first framework of discrete-time major-minor MFGs: So-called major agents can affect the usual "minor" agents from standard MFGs in an arbitrary direct manner, providing more flexibility for MFG models. We provide some basic theoretical properties and a learning algorithm. The work in Section 3.4 is based on the conference publication

[3] K. Cui, G. Dayanıklı, M. Laurière, M. Geist, O. Pietquin, and H. Koeppl, "Learning discrete-time major-minor mean field games", in *Proc. AAAI*, vol. 38, 2024, pp. 9616–9625.

CHAPTER 4. In contrast to the competitive setting in Chapter 3, in Chapter 4 we consider the cooperative setting of MFC, where agents are not assumed to be self-interested, rational agents.

As an initial contribution, we show in Section 4.1 that in a simplified static MF scenario, MFC with stochastic environment states allows us to improve beyond basic MARL by MFC-based RL. The optimality of identical over heterogeneous policies is shown, and results are verified on a realistic load balancing scenario. The work is based on the conference publication

[8] K. Cui, A. Tahir, M. Sinzger, and H. Koeppl, "Discrete-time mean field control with environment states", in *Proc. CDC*, 2021, pp. 5239–5246.

In Section 4.2 we proceed to extend general discrete-time MFC to major agents as major-minor MFC, similar to Section 3.4 but in the cooperative case. In particular, we point out a number of potential applications and also provide a theoretical basis of the framework. This includes approximation guarantees of the model, as well as policy gradient approximation properties. Furthermore, we perform an extensive comparison between our major-minor MFC learning framework and MARL methods. Our method appears to be able to outperform standard policy gradient MARL methods. The work is based on the conference publication

[2] K. Cui, S. Hauck, C. Fabian, and H. Koeppl, "Learning decentralized partially observable mean field control for artificial collective behavior", in *Proc. ICLR*, 2024, pp. 1–40.

Finally, in Section 4.3 we address the issue of all-knowing agents, which are not realistic especially in large decentralized systems. We formulate the first general MFC-based MARL system where each agent only observes limited information, correlated to the current MF and agent state. We provide a dynamic programming principle, a MARL algorithm, as well as approximation guarantees for both model and algorithm. The framework is verified on swarming models for engineering artificial collective behavior. The work is based on the conference publication

[1] K. Cui, C. Fabian, and H. Koeppl, "Major-minor mean field multi-agent reinforcement learning", *Proc. ICML*, 2024.

CHAPTER 5. After introducing general and tractable models for MARL based on MFG and MFC, we complete the thesis by discussing potential applications of such large-scale MARL. We also explicitly consider two possible applications of UAV swarms and edge computing. In Section 5.2 we apply MFC to formation control of UAV swarms. To accommodate real, embodied agents, we integrate collision avoidance into the framework and analyze the error resulting from it. On the other hand, in Section 5.3 we look at an edge computing and decentralized load balancing scenario where agents may choose to offload computation. We compare MFG and MFC points of view for this scenario, and demonstrate solution concepts to the problem. Further applications in the field of load balancing are found in the first part of Chapter 4 and external collaborations [10, 15, 21], which are only presented briefly in this thesis, see end of this chapter. Chapter 5 is based on the works

[5] K. Cui, M. Li, C. Fabian, and H. Koeppl, "Scalable task-driven robotic swarm control via collision avoidance and learning mean-field control", in *Proc. ICRA*, 2023, pp. 1192–1199.

[6] K. Cui, M. B. Yilmaz, A. Tahir, A. Klein, and H. Koeppl, "Optimal offloading strategies for edge-computing via mean-field games and control", in *Proc. GLOBECOM*, 2022, pp. 976–981.

[20] K. Cui, A. Tahir, G. Ekinci, A. Elshamanhory, Y. Eich, M. Li, and H. Koeppl, "A survey on large-population systems and scalable multi-agent reinforcement learning", *in preparation for AI Review, arXiv:2209.03859*, 2022.

CHAPTER 6. The last chapter gives a brief summary and discussion of the presented results, including their advantages and current limitations. As an outlook, future research directions and current limitations are established for theory, algorithms and applications.

APPENDICES A TO F. In the appendices, we give supplementary details on aforementioned contributions, including but not limited to full proofs of theoretical claims, briefly recapitulated material, and any additional supplementary experiments.

PUBLICATIONS NOT DISCUSSED IN THIS THESIS A number of publications and collaboration preprints generated during the work on this thesis are not discussed in detail. To understand their relation to the topic of this thesis, here we briefly give an overview of obtained results and their connection to material presented in this thesis. The following hence provides additional venues of application or generalization for MF models, which may be explored more in future works.

In [19], the GMFG framework for MFGs on graphs in Sections 3.2 and 3.3 is extended to directed and weighted graphs. In the following works [13, 14], we also move towards increasingly sparse graphs through the concept of L^p graphons and graphexes, which can handle sparsities up to degree distribution power law coefficients of 2. Algorithms and theoretical approximation guarantees are provided for the according systems, and the sparsity enables application on real graph datasets. In [18], we also developed continuous-time MF equations for higher-order dynamics using motifs (simple graph constellations) instead of hypergraphs as in Section 3.3. The above is based on the works

[13] C. Fabian, K. Cui, and H. Koeppl, "Learning mean field games on sparse graphs: A hybrid graphex approach", in *Proc. ICLR*, 2024, pp. 1–39.

[14] C. Fabian, K. Cui, and H. Koeppl, "Learning sparse graphon mean field games", in *Proc. AISTATS*, 2023, pp. 4486–4514.

[18] K. Cui, W. R. KhudaBukhsh, and H. Koeppl, "Motif-based mean-field approximation of interacting particles on clustered networks", *Phys. Rev. E*, vol. 105, no. 4, p. L042301, 2022.

[19] C. Fabian, K. Cui, and H. Koeppl, "Mean field games on weighted and directed graphs via colored digraphons", *IEEE Control Syst. Lett.*, vol. 7, pp. 877–882, 2022.

In relation to the UAV application in Section 5.2, we have also investigated various UAV swarm scenarios including coordination for collision avoidance in [16], networking in [4], low-level control of multi-drone structures in [12] and collaborative transportation of objects in [11]. While the above scenarios (except the first partially, see Section 5.2) are not considered through the lens of MF models here, they may be further investigated in future research. The above is based on the works

[4] K. Cui, L. Baumgärtner, M. B. Yilmaz, M. Li, C. Fabian, B. Becker, L. Xiang, M. Bauer, and H. Koeppl, "UAV swarms for joint data ferrying and dynamic cell coverage via optimal transport descent and quadratic assignment", in *Proc. LCN*, 2023, pp. 1–8.

[11] A. K. Sreedhara, D. Padala, S. Mahesh, K. Cui, M. Li, and H. Koeppl, "Optimal collaborative transportation for under-capacitated vehicle routing problems using aerial drone swarms", in *Proc. ICRA*, IEEE, 2024, pp. 8401–8407.

[12] M. Li, K. Cui, and H. Koeppl, "A modular aerial system based on homogeneous quadrotors with fault-tolerant control", pp. 8408–8414, 2024.

[16] R. Ourari, K. Cui, A. Elshamanhory, and H. Koeppl, "Nearest-neighbor-based collision avoidance for quadrotors via reinforcement learning", in *Proc. ICRA*, 2022, pp. 293–300.

Finally, related to the load balancing system in Section 4.1 and more distantly to the edge computing scenario in Section 5.3, in [15] we apply MFC-based MARL to scalable load balancing in the presence of many servers and schedulers. The approach is then extended to bounded-degree graph topologies in [21], and a related setting for optimization of age-of-information is considered in [10] using the partially-observable MFC framework presented in Section 4.3. The above is based on the works

[10] A. Tahir, K. Cui, A. Rizk, and H. Koeppl, "Collaborative optimization of the age of information under partial observability", *to appear in IFIP Networking 2024, arXiv:2312.12977*, 2023.

[15] A. Tahir, K. Cui, and H. Koeppl, "Learning mean-field control for delayed information load balancing in large queuing systems", in *Proc. ICPP*, 2022, pp. 1–11.

[21] A. Tahir, K. Cui, and H. Koeppl, "Sparse mean field load balancing in large localized queueing systems", *submitted to MobiHoc 2024, arXiv:2312.12973*, 2023.

2

BACKGROUND

In this chapter, we briefly introduce a background on concepts used in MARL and basic discrete-time MFGs / MFC, which are referred to throughout this work. While there are few works on continuous-time RL and many works on continuous-time MFGs / MFC, in discrete time on the other hand there are many works on RL but only few on MFGs / MFC. Since we focus on learning, we primarily consider discrete-time models in this thesis, and refer the reader to, e.g., [66, 67] for continuous-time literature. We reintroduce some concepts in this chapter in the respective sections to remind the reader, so one may skip to their sections of interest. Although we will overload some mathematical symbols in the following chapters according to the considered model setting, the semantics of introduced symbols will remain consistent and the same as in this exposition.

2.1 SINGLE-AGENT REINFORCEMENT LEARNING

The study of single-agent RL considers sequential decision-making in possibly stochastically evolving systems. The topic is closely related to the study of optimal control, which considers such sequential decision-making under known system models. In RL on the other hand, knowledge of the system model is waived. Instead, optimal sequential decision-making is "learned" through random interaction with the system, in a sample-based manner. In the following, we present the settings used in this thesis. We note that one can find a wealth of results in existing literature, and we present only a subset of classical results relevant to our work. We refer the reader interested in more details to [26, 68–70].

MARKOV DECISION PROCESS. The standard framework for RL is known as the Markov Decision Process (MDP), a tuple $(\mathcal{X}, \mathcal{U}, p, r, \gamma)$: It consists firstly of a system or agent state x_t at all times $t \in \mathcal{T}$, which is a random variable valued in some compact state space $\mathcal{X}$. The time index set is discrete and can be either finite, $\mathcal{T} = \{0, 1, \dots, T\}$ up to terminal time $T \in \mathbb{N}$, or infinite, $\mathcal{T} = \mathbb{N}_{\geq 0}$. The former is referred to as the finite-horizon case, whereas the latter is referred to as the infinite-horizon case. Furhter, the agent can influence the evolution of this state through the choice of actions u_t at all times $t \in \mathcal{T}$, which are similarly valued in some compact action space $\mathcal{U}$. As a result of actions, the state evolves over time according to some transition kernel p such that $x_{t+1} \sim p(x_{t+1} \mid x_t, u_t)$ at all times $t \in \mathcal{T}$. Commonly, the agent chooses actions according to a closed-loop feedback policy $\pi \in \Pi$, where Π is the set of all stochastic Markov policies such that the action depends only on the current time $t \in \mathcal{T}$ and state x_t, i.e. $u_t \sim \pi_t(u_t \mid x_t)$. In particular, due to the Markov property of MDPs, it is not necessary to consider policies that also depend on entire histories of past states or actions.

To obtain a controlled dynamical system, all that remains is to specify the initial state. We write $\mathcal{P}(\mathcal{X})$ for the space of all probability measures on $\mathcal{X}$, equipped with the 1-Wasserstein metric. Given an initial state distribution $\mu_0 \in \mathcal{P}(\mathcal{X})$, any choice of policy $\pi \in \Pi$ thus yields the controlled state process

$$x_0 \sim \mu_0, \quad u_t \sim \pi_t(u_t \mid x_t), \quad x_{t+1} \sim p(x_{t+1} \mid x_t, u_t) \quad \forall t \in \mathcal{T},$$

for which the goal of the agent is to maximize the sum of rewards over all times,

$$J(\pi) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} \gamma^t r(x_t, u_t)\right].$$

Here, the rewards are given by the reward function $r \colon \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ and discounted according to a discount factor $\gamma \in (0, 1)$ in the infinite-horizon case, or $\gamma = 1$ in the finite-horizon case. In general, the rewards may also be stochastic, which is included in the above model by taking the conditional expectation of rewards w.r.t. (x_t, u_t).

2.1.1 *Exact Dynamic Programming for Known Models*

In the infinite-horizon case, as long as the model is known, the exact solution to the above problem, i.e. an optimal policy $\pi^* \in \arg\max_{\pi \in \Pi} J(\pi)$, may be computed through classical dynamic programming by decomposing sequential optimal decision-making in the following way.

BELLMAN EQUATION. The well-known optimal state-action value function $Q^* \colon \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ specifies the best achievable expected sum of future rewards, whenever one is in a particular state and taking a particular action. It is defined as the solution to the Bellman equation

$$Q^*(x, u) = r(x, u) + \gamma \int_{\mathcal{X}} \max_{u' \in \mathcal{U}} Q^*(x', u') p(\mathrm{d}x' \mid x, u).$$

It essentially formalizes the idea that for optimality, it suffices to take an action u in state x such that the expected future rewards are maximal, given that we continue to act optimal in the future.

Under mild conditions [69, Theorem 4.2.3], it quantifies policies π^* putting full mass on actions u that maximize $Q_t^*(x, u)$ in any state x at time t as optimal solutions $\pi^* \in \arg\max_{\pi \in \Pi} J(\pi)$. Such a deterministic policy is guaranteed to exist. Note also the stationarity (time-independence) of value functions in the infinite-horizon case, such that it suffices to consider stationary policies.

VALUE ITERATION. To compute the solution to the Bellman equation by dynamic programming, its right-hand side may be repeatedly evaluated and assigned as a solution estimate, starting with any initial estimate $Q^{(0)}\colon \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. In other words, we apply fixed-point iteration to the Bellman equation as a fixed-point equation. The result is *value iteration*, where one repeatedly computes

$$Q^{(k+1)}(x,u) = r(x,u) + \gamma \int_{\mathcal{X}} \max_{u' \in \mathcal{U}} Q^{(k)}(x',u') p(\mathrm{d}x' \mid x,u)$$

in each iteration $k \in \mathbb{N}$, until the errors between left-hand and right-hand side (Bellman error) becomes sufficiently small. Indeed, value iteration is guaranteed to converge, since the map $Q^{(k)} \mapsto Q^{(k+1)}$ is a contraction (Lipschitz with constant $L < 1$), so $Q^{(k)} \to Q^*$ as $k \to \infty$.

POLICY ITERATION. The other classical dynamic programming approach to MDPs is to directly and iteratively calculate policies in *policy iteration*, which can be faster than value iteration. First, observe that one can evaluate any policy $\pi \in \Pi$ via the policy state-action value function $Q^\pi\colon \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, which specifies the expected future rewards in a particular state, taking a particular action, and following policy π thenceforth. Q^π solves the policy evaluation equation

$$Q^\pi(x,u) = r(x,u) + \gamma \int_{\mathcal{X}} \int_{\mathcal{U}} Q^*(x',u') \pi(\mathrm{d}u' \mid x) p(\mathrm{d}x' \mid x,u).$$

Similar to value iteration, one starts with some arbitrary initial policy $\pi^{(0)} \in \Pi$ and solves for $Q^{\pi^{(k)}}$ in each iteration $k \in \mathbb{N}$. The policies are then updated such that $\pi^{(k+1)} \in \Pi$ puts all mass on actions u in state x that maximize $Q^{\pi^{(k)}}(x,u)$ until convergence, giving the policy iteration algorithm.

ALTERNATIVE SETTINGS AND REFERENCES. In the above discounted infinite-horizon problem, it is sufficient for optimality to use time-independent values and policies. In the finite-horizon case, we can analogously compute the above, except by using *time-dependent* action-value functions $Q^*\colon \mathcal{T} \times \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ and policies. We note that there are also other settings and algorithms, such as optimizing average reward and using linear programming. For more details, see e.g. [26, 69].

2.1.2 *Reinforcement Learning for Unknown Models*

To learn an optimal solution from samples whenever the exact model is not known or explicitly used, a straightforward and basic RL approach is similar to value iteration but uses sample-based estimation. On the other hand, one may directly optimize policies by estimating gradients.

VALUE-BASED REINFORCEMENT LEARNING. The former falls into the area of so-called value-based techniques. They can be understood through stochastic approximation techniques and in its simplest form are given by Q-Learning [71]. Its idea is to not re-compute the entire action-value function in every iteration as in value iteration. Since the model is not known, one instead updates the current action-value estimate $\hat{Q}$ around new samples $(x_t, u_t, x_{t+1}) \in \mathcal{X} \times \mathcal{U} \times \mathbb{R} \times \mathcal{X}$ as

$$\hat{Q}(x_t,u_t) \leftarrow \hat{Q}(x_t,u_t) + \alpha \left(r(x_t,u_t) + \gamma \max_{u \in \mathcal{U}} \hat{Q}(x_{t+1},u) - \hat{Q}(x_t,u_t) \right)$$

with step size $\alpha > 0$. However, Q-Learning can be unstable when paired with (deep) function approximators for scaling to large state spaces. Recent modern RL has introduced stabilization techniques such as sampling from a replay buffer and using target networks, which leads to the

so-called Deep Q-Network (DQN) method [33] and its derivatives. Such modern RL and function approximation addresses the so-called "curse of dimensionality", where exact dynamic programming becomes difficult due to exponential scaling with the dimensionality of states and actions.

POLICY-BASED REINFORCEMENT LEARNING. The other common approach to RL is given by gradient-based methods, where one starts with an initial policy π^θ parametrized by some parameters θ. In contrast to value-based methods, policy-based gradient methods are not necessarily guaranteed to converge to globally optimal policies. However, they have the advantage of being immediately applicable to continuous action spaces, in contrast to using a value function which should output values for all possible action. Given a current policy π^θ, the idea is to estimate the gradient $\nabla_\theta J(\pi^\theta)$ from samples, since the model is assumed unknown. In particular, the celebrated policy gradient theorem implies that one can obtain the policy gradient as an expectation over seen states

$$\nabla_\theta J(\pi^\theta) = (1-\gamma)^{-1}\mathbb{E}_{x\sim d_{\pi^\theta}, u\sim\pi^\theta(\mu)}\left[Q^\theta(\mu,\xi)\nabla_\theta \log \pi^\theta(\xi \mid \mu)\right]$$

where $Q^\theta(\hat{\mu},\xi) = \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t r(\hat{\mu}_t) \mid \hat{\mu}_0 = \mu, \xi_0 = \xi]$ and $d_{\pi^\theta} = (1-\gamma)\sum_{t\in\mathcal{T}}\gamma^t \mathcal{L}_{\pi^\theta}(\hat{\mu}_t)$. Here, $\mathcal{L}_{\pi^\theta}$ denotes the probability law of π^θ. Using standard sample-based estimates of the above yields the most basic algorithm commonly known as REINFORCE [72].

RECENT TECHNIQUES. More recent techniques include many more algorithms such as Proximal Policy Optimization (PPO) [73], which is based on computationally simplified trust region optimization [74] and uses an actor-critic framework, where the actor is the policy and the critic is a value function estimate for variance reduction. It performs well empirically and finds application in modern LLMs [39] and complex games [75]. However, the basic principles are not different from the ones presented above. We refer readers interested in applying RL to comprehensive software frameworks such as [76] and references therein.

2.2 MULTI-AGENT REINFORCEMENT LEARNING

In the field of MARL, the preceding single-agent MDP is extended according to the considered scenario. In general, multiple agents are instantiated, whose states and actions may now interact with each other and must therefore be considered jointly.

MARL can be considered to subsume parts of algorithmic game theory, control theory and single-agent RL. As a result, there is also a great number of differing notations, models and scenarios in MARL. In this work, we focus on particular general instances of problems. We note however, that the multi-agent setting permits a great range of scenarios, and that we only present scenarios that are considered in this thesis. For more information and references on multi-agent settings, we refer the reader to various surveys [40, 43, 44].

2.2.1 *Competitive Setting*

In the typical competitive scenario, agents are assumed to care only about their own rewards. This gives rise to the setting of a Stochastic Game (SG), a tuple $(N, \mathcal{X}, (\mathcal{U}^i)_{i\in[N]}, p, (r^i)_{i\in[N]}, \gamma)$ (also known as Markov game): Consider some number $N \in \mathbb{N}$ of agents $i \in [N] \coloneqq \{1, 2, \dots, N\}$. As in the MDP, the system is endowed with a controlled state process x_t from some state space $\mathcal{X}$, but

in contrast to the MDP, each agent $i \in [N]$ may influence it through their associated control actions u_t^i from action spaces $\mathcal{U}^i$. As before, one can consider closed-loop feedback policies $\pi^i \in \Pi^i$ for each agent i, where Π^i is the set of all stochastic Markov policies such that the action depends only on the current time $t \in \mathcal{T}$ and state x_t, i.e. $u_t \sim \pi_t(u_t \mid x_t)$. Then, the state evolves over time according to some transition kernel p such that $x_{t+1} \sim p(x_{t+1} \mid x_t, u_t^1, \dots, u_t^N)$ at all times $t \in \mathcal{T}$. This gives us the model

$$x_0 \sim \mu_0(x_0), \quad u_t^i \sim \pi_t^i(u_t^i \mid x_t), \quad x_{t+1} \sim p(x_{t+1} \mid x_t, u_t^1, \dots, u_t^N) \quad \forall t \in \mathcal{T}, \forall i \in [N]. \tag{2.2.1}$$

Finally, the goal of each agent $i \in [N]$ is separate in general and given by some function

$$J^i(\pi^1, \dots, \pi^N) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} \gamma^t r^i(x_t, u_t^1, \dots, u_t^N)\right]. \tag{2.2.2}$$

SOLUTION CONCEPTS. A common goal or solution concept is then the so-called Nash equilibrium, which is the goal to be computed in equilibrium learning. It is defined as a tuple of policies $(\pi^1, \dots, \pi^N)$ such that no agent can single-handedly change its policy to improve its objective. By its definition, it is therefore a "stable" solution where no self-interested agent has incentive to deviate. While there are many other solution concepts, see e.g. [77] for some examples, in this work we primarily focus on the Nash equilibrium.

ALGORITHMS AND COMPLEXITY. In general, the complexity of computing Nash equilibria is difficult (PPAD-complete, [45]). For certain special cases, an algorithm can be found with guarantees of convergence. As one of many examples, classical Fictitious Play (FP) [78] is a method that repeatedly computes the best deviating policies against the tuple of average past policies in potential games or two-player zero-sum games. Indeed, one can consider the special case of two-player or two-team zero-sum games, which also enjoys improved tractability. As a result, a great amount of both classical and recent literature considers two-player zero-sum games [40]. In this work however, we focus on general-sum scenarios with many more agents than two. More generally, one can also consider partially-observable SGs or alternate frameworks such as extensive-form games, see e.g. [40, Section 2], but in this work we consider partial observations only in the cooperative setting. For partial observations in the MFG setting, see e.g. [79].

2.2.2 *Cooperative Setting*

In contrast to the competitive scenario, in the cooperative scenario all agents are assumed to share rewards. To be precise, one considers Multi-agent Markov Decision Processes (MMDPs) [43, 80] as a tuple $(N, \mathcal{X}, (\mathcal{U}^i)_{i \in [N]}, p, r, \gamma)$. Its definition then mirrors the definition of the competitive SG, with one important difference: The rewards r^i for each agent i are replaced by a shared reward function r. We obtain the overall model

$$x_0 \sim \mu_0(x_0), \quad u_t^i \sim \pi_t^i(u_t^i \mid x_t), \quad x_{t+1} \sim p(x_{t+1} \mid x_t, u_t^1, \dots, u_t^N) \quad \forall t \in \mathcal{T}, \forall i \in [N]. \tag{2.2.3}$$

as in Eq. (2.2.1), but with a centralized maximization objective for all agents,

$$J(\pi^1, \dots, \pi^N) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} \gamma^t r(x_t, u_t^1, \dots, u_t^N)\right]. \tag{2.2.4}$$

ALGORITHMS AND COMPLEXITY. It is then clear that, at least in the all-knowing fully-observable case, one can replace the multitude of agents by a single "super-agent", consisting of a joint state and a joint action space. As a result, the problem in theory reduces to an MDP, which can be solved by single-agent dynamic programming (PSPACE, [81]) and RL techniques discussed earlier. However, in general the joint action space scales exponentially with the number of agents, which can make the problem intractable. Furthermore, in many problems, each agent is also endowed with an associated state, the product of which should constitute part of the system state. The resulting exponentially large state and action spaces in the number of agents overall lead to tractability issues in the presence of many agents, which is referred to as the combinatorial nature of MARL [40].

PARTIAL INFORMATION. More importantly, if each agent is incapable of unrestricted communication with other agents while each agent only observes part of the system, the super-agent construction fails. This case is arguably more realistic in large-scale real systems. This leads us to the so-called Decentralized Partially-Observable Markov Decision Process (Dec-POMDP) model, a tuple $(N, \mathcal{X}, (\mathcal{U}^i)_{i\in[N]}, (\mathcal{Y}^i)_{i\in[N]}, p, (p^i_{\mathcal{Y}})_{i\in[N]}, r, \gamma)$: As before in the MMDP, we have a controlled state process x_t from some state space $\mathcal{X}$, and each agent $i \in [N]$ chooses control actions u^i_t from action spaces $\mathcal{U}^i$. In contrast to the prequel however, agents cannot see the true state x_t of the system. Instead, agents $i \in [N]$ are assumed to observe information $y^i_t \sim p^i_{\mathcal{Y}}(y^i_t \mid x^i_t)$ from observation space $\mathcal{Y}^i$ according to some observation probability kernel $p^i_{\mathcal{Y}}$. Agents have no access to the true state, and the information from the history of observation-actions cannot be reduced to a belief over the true state as in single-agent Partially-Observable Markov Decision Processs (POMDPs) [43, 82]. Hence, in general one usually considers closed-loop feedback policies $\pi^i \in \Pi^i$ for each agent i, where Π^i is the set of all stochastic history-dependent policies, such that the action depends on the current time $t \in \mathcal{T}$ and all past observations and actions $(y^i_\tau)_{\tau\le t}$, $(u^i_\tau)_{\tau<t}$, i.e. $u^i_t \sim \pi^i_t(u^i_t \mid y^i_0, u^i_0, \dots, y^i_{t-1}, u^i_{t-1}, y^i_t)$. This gives us the model

$$\begin{aligned} x_0 \sim \mu_0(x_0), \quad y^i_t \sim p^i_{\mathcal{Y}}(y^i_t \mid x^i_t), \quad u^i_t \sim \pi^i_t(u^i_t \mid y^i_0, u^i_0, \dots, y^i_{t-1}, u^i_{t-1}, y^i_t), \\ x_{t+1} \sim p(x_{t+1} \mid x_t, u^1_t, \dots, u^N_t) \quad \forall t \in \mathcal{T}, \forall i \in [N]. \end{aligned} \tag{2.2.5}$$

with centralized objective as before,

$$J(\pi^1, \dots, \pi^N) = \mathbb{E}\left[\sum_{t\in\mathcal{T}} \gamma^t r(x_t, u^1_t, \dots, u^N_t)\right]. \tag{2.2.6}$$

And again, it is known that the above decentralized control problem is highly intractable (NEXP-complete, [46]). Overall, the difficulties in scaling up to multi-agent problems are commonly known as the "curse of multiagents" or "combinatorial nature of MARL" [40]. Therefore, similar to how RL addresses the single-agent "curse of dimensions", one is motivated to (i) use approximate methods for MARL discussed in the following, or (ii) consider special case scenarios. We propose a combination of both of the above in Section 4.3.

2.2.3 *Common Recent Algorithms*

Various algorithms exist for solving MARL in special or in general. We only give a few of the most important examples in recent literature, and point to more comprehensive surveys such as [40].

VALUE DECOMPOSITION. A well-known value-based MARL algorithm is the Value Decomposition Network (VDN) [83], which focuses on the cooperative setting with a joint reward function to be maximized. The approach addresses the so-called lazy agent phenomenon, where one agent is active while the other policies remain inefficient, which is also related to the credit assignment problem of MARL [40]. Essentially, the credit assignment problem in MARL formalizes the fact that in the presence of many agents, rewards can be caused by any of the agent actions. The result is a higher variance of reward estimates, as other agent actions must be averaged out first. The above is solved by learning agent-wise value functions as an additive decomposition of the joint value function. Thereby, the learning of a joint value function over the exponential state and action space is avoided for scalability. Its many extensions such as QMIX [84, 85] further increase the range of representable decomposed value functions, by not assuming additive but instead monotonic decompositions.

INDEPENDENT LEARNING. Instead of performing the super-agent construction in the cooperative case, and in order to also handle the competitive case, a common idea is to perform independent learning [86]. In the simplest case, we can use Q-Learning on each agent separately, assuming a single-agent RL problem on each agent. The policy is then given as a maximizer of the learned action-value function. More popularly however, policy gradient methods such as PPO are used. This so-called Independent PPO (IPPO) method has repeatedly been demonstrated to give state-of-the-art performance on a diverse set of cooperative multi-agent benchmark tasks [42, 87–89]. Similarly, its extension using centralized critics (Multi-Agent PPO (MAPPO)) often performs well [88]. There, the idea is that it suffices for decentralized execution to have decentralized policies, while allowing for centralized information in the critics. This idea of learning with extra / full information is widely known as the Centralized Training Decentralized Execution (CTDE) paradigm.

PARAMETER SHARING Independent learning such as in the prequel is commonly coupled with the idea of parameter sharing [90], where all agents are assumed to have the same observation and action space, and learn a single set of policy parameters. The parameter sharing approach is one way of handling large numbers of agents tractably and even scale to unforeseen or dynamic numbers of agents in a system. However, general convergence guarantees of algorithms usually remain difficult, and common benchmarks often consider only limited numbers of agents [42, 89].

FURTHER TECHNIQUES. Again, as in the single-agent RL setting, many more algorithms have been proposed in literature that address particular issues with the above algorithms. Readers interested in more details on general MARL methods are referred to MARL frameworks such as [91] and references therein. Due to the difficulty of general MARL, often one considers particular settings that are more tractable, such as graph-based decompositions [92, 93] or also the MFG and MFC scenarios we are considering in the following.

2.3 INFINITE-AGENT MEAN FIELD REINFORCEMENT LEARNING

As discussed in the prequel, MARL can be hard to scale to many agents. In this work, recalling Figure 1.2, standard MFGs and MFC are obtained by a dynamical system where homogeneous agents are anonymously and weakly interacting with each other.

An increasingly popular and recent approach to the tractability issue has been to use the framework of learning in MFGs [9, 94–101] and their cooperative counterpart of MFC [8, 102–107]. It is

important to note that here, learning often also refers to classical equilibrium learning – i.e. iterative convergence to equilibria – in game theory, as opposed to e.g. RL, see also the discussion in [25]. Popularized by [23] and [22] in the context of differential games, MFGs and related approximations have since found application in a plethora of fields such as transportation and traffic control [54, 55, 108], large-scale batch processing and scheduling systems [56, 57, 109], peer-to-peer streaming systems [58], malware epidemics [59], crowd dynamics and evacuation of buildings [60, 110, 111], as well as many other applications in economics [64] and engineering [65].

Since the inception of MFGs, extensions have been manifold and include e.g. discrete-time models [24], partial observability [112], major-minor formulations [113] and many more. In the learning community, there has been immense recent interest in finding and analyzing solution methods for Mean Field Equilibrium (MFE) [9, 94–96, 114–118], solving the inverse RL problem [119] or applying related approximations directly to MARL [120]. Even more recently, focus increased also on the cooperative case of MFC [105, 107], for which dynamic programming holds on an enlarged state space, resulting in a high-dimensional MDP [8, 103, 104, 121]. Due to the extensive scale of prior conducted investigations, for a comprehensive overview of existing work in MFGs and learning thereof, we refer the interested reader to many extensive reviews on MFGs [25, 67, 122–125]. We also discuss works related to specific contributions of ours in the corresponding Chapters 3 to 5.

FINITE-AGENT MEAN FIELD MODELS. In essence, the typical finite-agent problem of interest is given by taking a multi-agent system as defined in the prequel with N agents $i \in [N] := \{1, 2, \dots, N\}$ and taking $N \to \infty$. To obtain a first mean field approximation result, it is assumed that the system state is simply a product of agent states, and that all agents are homogeneous (same state and action spaces $\mathcal{X}, \mathcal{U}$, same initial distribution μ_0, same dynamics p and reward functions r). Furthermore, one assumes that all agent interdependence happens through the MF $\mu_t^N = \sum_{i \in [N]} \delta_{x_t^i}$ ("histogram of agent states", here δ is the Dirac measure)

$$x_0^i \sim \mu_0(x_0^i), \quad u_t^i \sim \pi_t^i(u_t^i \mid x_t^i), \quad x_{t+1} \sim p(x_{t+1} \mid x_t, u_t, \mu_t^N) \quad \forall t \in \mathcal{T}, \forall i \in [N]. \tag{2.3.7}$$

For MFGs, each agent shares policies $\pi^1 = \ldots = \pi^N = \pi \in \Pi$. In this section, Π is the set of all policies such that the action depends only on the current time $t \in \mathcal{T}$ and local agent state x_t, which suffices over local history-dependent policies [24, Proposition 3.2]. Finally, in MFGs each agent is equipped with its own objective

$$J_i^N(\pi) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} \gamma^t r(x_t^i, u_t^i, \mu_t^N)\right]. \tag{2.3.8}$$

We can hence view the finite-agent MFG as a special case of SGs in Eqs. (2.2.1) and (2.2.2).

In contrast, the cooperative finite-agent MFC problem is equipped with a single global objective

$$J^N(\pi) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} \gamma^t r(\mu_t^N)\right]. \tag{2.3.9}$$

We can understand the above model as a special case of MMDPs, or Dec-POMDPs in Eqs. (2.2.5) and (2.2.6) for the more general partially-observable case introduced in Section 4.3.

LIMITING INFINITE-AGENT PROBLEMS. Taking the limit of $N \to \infty$ gives us the classical MFG and MFC problems, where the empirical MF μ_t^N is instead replaced by a deterministic MF μ_t via LLN. Any single agent can be replaced by a single representative that acts according to arbitrary $\hat{\pi} \in \Pi$, under the presence of all other agents acting according to $\pi \in \Pi$, i.e.

$$x_0 \sim \mu_0(x_0), \quad u_t \sim \hat{\pi}_t(u_t \mid x_t), \quad x_{t+1} \sim p(x_{t+1} \mid x_t, u_t, \mu_t) \quad \forall t \in \mathcal{T}, \tag{2.3.10}$$

$$\mu_{t+1} = \int_{\mathcal{X}} \int_{\mathcal{U}} p(\cdot \mid x, u) \pi_t(\mathrm{d}u \mid x) \mu_t(\mathrm{d}x) \quad \forall t \in \mathcal{T}. \tag{2.3.11}$$

The objective in the MFG case then becomes

$$J^{\mu}(\hat{\pi}) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} \gamma^t r(x_t, u_t, \mu_t)\right], \tag{2.3.12}$$

and for the MFC case, irrespective of any single agent deviating from all other agents' policy π,

$$J(\pi) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} \gamma^t r(\mu_t)\right]. \tag{2.3.13}$$

The above MF $\mu \in \mathcal{M}$ is generated by the policy $\pi \in \Pi$, where $\mathcal{M} \coloneqq \mathcal{P}(\mathcal{X})^{\mathcal{T}}$ is the space of MFs. As a result, the concept of a Nash equilibrium becomes that of a policy π that generates a MF μ, under which the representative agent's optimal policy is π again. We usually write $\Phi(\mu)$ for the map from MF μ to all optimal (best response) policies maximizing Eq. (2.3.12), and we write $\mu = \Psi(\pi)$ for the MF generated by π. Therefore, a Nash MFE π fulfills $\pi \in \Phi(\Psi(\pi))$.

On the other hand, in MFC one can see that any single representative agent alone does not matter. Performing a "super-agent" construction, at each time step t one essentially controls the MF μ_t (all agents) via choosing π_t, or equivalently the joint $h_t \coloneqq \mu_t \otimes \pi_t(\mu_t)$, which gives rise to the MFC MDP (e.g., [107]). The MFC MDP is then simply given by an MDP with dynamics

$$h_t \sim \hat{\pi}_t(h_t \mid \mu_t), \quad \mu_{t+1} = T(\mu_t, h_t) \coloneqq \iint p(\cdot \mid x, u, \mu_t) h_t(\mathrm{d}x, \mathrm{d}u) \tag{2.3.14}$$

and objective $J(\hat{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(\mu_t)\right]$, where the system states are μ_t and actions are h_t, controlled by a "higher-level" super-agent policy $\hat{\pi}$. Here, $\mu_t \otimes \pi_t(\mu_t)$ is the *desired joint state-action distribution* under some "lower-level" policy $\pi_t(\mu_t) \in \Pi$. See also Chapter 4.

THEORETICAL GUARANTEES. The infinite-agent limit is a proxy for the finite-agent problem of interest, which is a good approximation for large finite problems with many agents under certain conditions: Propagation of chaos [126] – the convergence of the empirical MF μ_t^N to its deterministic limit μ_t as $N \to \infty$ by a LLN – is typically obtained to guarantee the approximate optimality of MF solutions in the finite problems of interest. For example, under Lipschitz continuity assumptions on transitions and rewards, a classical rate of approximation is $\mathcal{O}(1/\sqrt{N})$ [23], e.g., as $\sup_{f \in C_b(\mathcal{P}(\mathcal{X}))} \mathbb{E}\left[f(\mu_t^N) - f(\mu_t)\right] = \mathcal{O}(1/\sqrt{N})$ over bounded and continuous functions f.

In MFGs, this ideally means obtaining approximate Nash equilibria where each finite agent can only gain at most $\mathcal{O}(1/\sqrt{N})$ by deviating from its MFE policy. On the other hand, in MFC an approximate $\mathcal{O}(1/\sqrt{N})$-optimality of MFC solutions can be shown. Even with only simple model continuity, asymptotic results can sometimes be obtained (see, e.g., Theorem 3.2.3), but a quantified rate becomes more difficult. Furthermore, it is possible to obtain the existence of Nash equilibria in the MFG case, making the model useful. If needed, under additional assumptions such as monotonicity (i.e. agents disliking crowded states) [127], one can often obtain uniqueness of equilibria and algorithmic guarantees of convergence to the desired equilibrium.

ASSUMPTIONS OF MEAN FIELD MODELS We note that the assumptions are worth discussing: Firstly, we would like to briefly discuss the assumption of identical behavior between agents. In particular, for the derivation of the above limits, it is assumed that all agents use the same policy. One can however argue that symmetry suffices: In the competitive case, we are interested in finding Nash equilibria, which are guaranteed to exist in the MF limit, as approximate symmetric Nash equilibria. Meanwhile, in MFC (or two-team MF problems), recent literature [128] has shown that using identical policies between all agents is associated with only negligible suboptimality.

Secondly, note that the assumed continuity is strictly required in MF models, in order to obtain the propagation of chaos as a theoretical grounding of MF models to the real finite-agent system. In contrast, consider an example of voting, where the majority between two choices wins. If agents use a uniformly random policy, it is not possible to find a deterministic MF model that adequately describes the random winning behavior in the finite agent systems.

Finally, the closely related weak interaction between agents is therefore a founding principle of the MF models introduced above. Otherwise, if an agent has no negligible effect on the whole, it is not possible to reduce all other agents to a MF that remains unaffected by it. Weak interaction and the other issues of all-knowing agents and homogeneity in RQ II are addressed in this thesis, by extending the basic models in Chapter 3 and Chapter 4.

ALGORITHMIC SOLUTIONS AND MARL. As for single-agent and multi-agent RL in infinite-agent systems, various algorithms have been proposed in the last few years to solve MFGs and MFC. There has also been work on special cases such as stationary problems (where the MF does not change) or one-shot problems without time. A recent survey of learning algorithms can be found in [25].

In MFGs, learning typically refers to equilibrium learning as the computation of equilibria, instead of using RL. A classical approach is through classical Fixed Point Iteration (FPI) [23], where one simply iterates optimal policies $\pi^{(k+1)} \in \Phi(\Psi(\pi^{(k)}))$ over iterations $k \in \mathbb{N}$. Another approach [94, 99, 114] focuses on potential MFGs by adapting the classical game-theoretical FP algorithm [77]. Together with deep techniques including normalizing flows and deep RL ([129], deep RL as in Chapter 3), as well as online mirror descent [117, 130], recent algorithms are also scaled to larger state spaces. Finally, optimization and linear program formulations of MDPs may be used [131]. For a quick start, see also software frameworks such as MFGLib [132] or OpenSpiel [133].

However, unfortunately FPI tends to fail in discrete-time and it is usually difficult to verify theoretical convergence guarantees, as we show in our first contribution in Chapter 3. There, we analyze regularized games as a trade-off between convergence and optimality, and later extend algorithms to more general settings on graphs and with strong interactions. In particular, existing algorithms mostly fall into two categories of assumptions: Methods such as FPI assume the contractivity of the best response map, i.e. the map from policy to its optimal policy under the generated mean field. On the other hand, methods such as FP and online mirror descent assume monotonicity conditions. These assumptions limit the applicability of MFGs to practical problems that do not fulfill the assumptions. In general, these methods can then be applied as oracles in a model-based RL manner for unknown system models [134].

Meanwhile, in MFC learning, existing works focus on solving the MFC MDP through policy-based RL methods [105] and discretization [107]. This means directly using the methods that we have presented in Section 2.1.2, but on a high-dimensional or possibly infinite-dimensional MDP with infinitely many agents, as its state is a distribution over agent states, and similarly its action. Furthermore, MFC algorithms were so far analyzed mostly as algorithms on the infinite-agent system. Instead, as part of our contributions in Chapter 4, we look at kernel-based parametrizations instead

of discretization for scalability, learn on the finite system instead of the infinite MFC MDP in a true MARL manner, and analyze the resulting error of policy gradients. Apart from purely algorithmic contributions, we also extend methods and theories to more general settings with either partial information or strong interactions.

2.4 CONCLUSION OF CHAPTER 2

We have introduced basic models and problem scenarios studied in this thesis. The frameworks were discussed in discrete time, which is less standard for MFGs and MFC literature, but more standard for MARL and MF-learning literature.

First, we briefly introduced basic concepts of single-agent RL. In the initial exposition, we discussed the single-agent MDP as a model of sequential decision-making or control, that can be solved exactly through dynamic programming methods such as value iteration or policy iteration. Similarly, RL was introduced with value-based and policy-based approaches, as a sample-based and tractable way of dealing with high-dimensional MDPs and their "curse of dimensionality".

We then explained SGs and Dec-POMDPs as two intractable but general models for MARL as the generalization of RL to multiple agents problems, in the competitive and cooperative case respectively. The models generalize the single-agent MDP and add significant additional complexity, known as the "curse of multiagents" or "combinatorial nature of MARL". Some of the most common techniques for MARL were presented, together with some of their ideas on how to address scaling to multiple agents via additional assumptions such as parameter sharing or decomposable value functions.

Finally, we moved on to introduce basic discrete-time MFGs and MFC as another set of tractable but specialized models for MARL problems. Some basic algorithmic approaches to MARL using MFGs and MFC were described. However, firstly MFGs and MFC restrict the space of problems that may be solved, and secondly the solution thereof still remains to be explored more. For example, MFG learning algorithms are yet limited to assumptions such as contractivity or monotonicity, while MFC-based MARL should be analyzed on finite-agent systems.

In the following chapters and sections, we introduce generalizations of the above models, as well as novel algorithms that either address issues in existing algorithms or generalize to more advanced MF models. As discussed in Chapter 1, we hope our work improves applicability of MFGs and MFC, making them more applicable and solvable, by providing both novel theoretical frameworks with guarantees and novel algorithms.

3

COMPETITIVE MEAN FIELD GAMES

In this chapter, we study the competitive case of MFGs for tractable equilibrium learning in large-scale games. In general, one may focus on various simplifying settings of MFGs such as the static case where agents have no dynamics and there is no time evolution, or the stationary case where the MF is stationary over time. Here, we focus on the general setting of evolutive MFGs, where the MF evolves over time. We first analyze general MFGs with respect to their solvability through FPI, and give some algorithms. We then move on to extensions of MFGs to graph-based interaction and major agents in order to increase the generality and flexibility of MFGs.

3.1 REGULARIZATION FOR APPROXIMATE LEARNING OF MEAN FIELD GAMES

The recent MFG formalism promises otherwise intractable computation of approximate Nash equilibria in many-agent settings. In this section, we consider discrete-time finite MFGs subject to finite-horizon objectives. We show that all discrete-time finite MFGs with non-constant fixed point operators fail to be contractive as typically assumed in existing MFG literature, barring convergence via FPI. Instead, we incorporate entropy-regularization and Boltzmann policies into the FPI. As a result, we obtain provable convergence to approximate fixed points where existing methods fail, and reach the original goal of approximate Nash equilibria. All proposed methods are evaluated with respect to their exploitability, on both instructive examples with tractable exact solutions and high-dimensional problems where exact methods become intractable. In high-dimensional scenarios, we apply established deep RL methods and empirically combine FP with our approximations. The material presented in this section is based upon our work [9].

Computing an MFE remains difficult in the general case. Standard assumptions in existing literature are MFE uniqueness and operator contractivity [23, 95, 135] to obtain convergence via simple FPI. While these assumptions hold true for some games, we address the case where such restrictive assumptions fail. Applications for such MF models are manifold and include e.g. finance [123], power control [136], wireless communication [137] or public health models [138].

A MOTIVATING EXAMPLE. Consider the following trivial situation informally: Let a large number of agents choose simultaneously between going left (L) or right (R). Afterwards, each agent shall be punished proportional to the number of agents that chose the same action. If we had infinitely many independent, identically acting agents, the only stable solution would be to have all agents pick uniformly at random.

The MFG formalism models this problem by picking one representative agent and abstracting all other agents into their state distribution. Unfortunately, analytically obtaining fixed points in general proves difficult and existing computational methods can fail.

OUR CONTRIBUTION. We begin by formulating the MF analogue to finite games in game theory. In this setting we give simplified proofs for both existence and the approximate Nash equilibrium property of MFE. Moreover, we show that in finite MFGs, all non-constant fixed point operators are non-contractive, necessitating a different approach than naive FPI.

Consequently, we approximate the fixed point operator by introducing relative entropy regularization and Boltzmann policies. We prove guaranteed convergence for sufficiently high temperatures, while remaining arbitrarily exact for sufficiently low temperatures. Furthermore, repeatedly iterating on the prior policy allows us to perform an iterative descent on exploitability, successively improving the equilibrium approximation.

Finally, our methods are extensively evaluated and compared to other methods such as FP [127], which in general fail to converge to a fixed point. We outperform existing methods in terms of exploitability in our problems, allowing us to find approximate MFE in the general case and paving the way to practical application of MFGs. In otherwise intractable problems, we apply deep RL techniques together with particle-based simulations.

3.1.1 *Simple Finite Mean Field Games*

Consider a discrete-time N-agent stochastic game with finite agent state space $\mathcal{X}$ and finite agent action space $\mathcal{U}$, equipped with the discrete metric. Let $\mathcal{T} = \{0, 1, \dots, T-1\}$ denote the time index set. Denote by $\mathcal{P}(\mathcal{X})$ the set of all Borel probability measures on a metric space $\mathcal{X}$. Since we work with finite spaces, we abuse notation and denote both a measure ν and its probability mass function by $\nu(\cdot)$. For each agent, the dynamical behavior is described by the state transition function $p : \mathcal{X} \times \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \to [0, 1]$ and the initial state distribution $\mu_0 : \mathcal{X} \to [0, 1]$. For agents $i = 1, \dots, N$ at times $t \in \mathcal{T}$, their states x_t^i and actions u_t^i are random variables with values in $\mathcal{X}$ and $\mathcal{U}$ respectively. Let $\mu_t^N \equiv \frac{1}{N} \sum_{i=1}^N \delta_{x_t^i}$ denote the empirical MF, where δ is the Dirac measure. Further, let $\mu[\mathbf{x}] \equiv \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ denote the empirical measure of agent states $\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{X}^N$. We also define $\mathbf{x}_t = (x_t^1, \dots, x_t^N)$. Consider for each agent i a Markov policy $\pi^i = (\pi_t^i)_{t \in \mathcal{T}} \in \Pi$, where $\pi_t^i : \mathcal{U} \times \mathcal{X} \to [0, 1]$ and Π is the space of all Markov policies. The state evolution of agent i begins with $x_0^i \sim \mu_0$, and subsequently for all applicable times t follows

$$u_t^i \sim \pi_t^i(u_t^i \mid x_t^i), \quad x_{t+1}^i \sim p(x_{t+1}^i \mid x_t^i, u_t^i, \mu[\mathbf{x}_t]) \quad \forall i \in [N].$$

Finally, define agent i's finite horizon objective function

$$J_i^N(\pi^1, \dots, \pi^N) \equiv \mathbb{E}\left[\sum_{t=0}^{T-1} r(x_t^i, u_t^i, \mu_t^N)\right]$$

to be maximized, where $r : \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is the agent reward function, and note $\mu_t^N = \mu[\mathbf{x}_t]$. With this, we can give the notion of optimality used by [24].

Definition 3.1.1. *A Markov-Nash equilibrium is a* 0*-Markov-Nash equilibrium. For* $\varepsilon \geq 0$*, an* ε*-Markov-Nash equilibrium (approximate Markov-Nash equilibrium) is defined as a tuple of policies* $(\pi^1, \dots, \pi^N) \in \Pi^N$ *such that for any* $i = 1, \dots, N$*, we have*

$$J_i^N(\pi^1, \dots, \pi^N) \geq \max_{\pi \in \Pi} J_i^N(\pi^1, \dots, \pi^{i-1}, \pi, \pi^{i+1}, \dots, \pi^N) - \varepsilon\,.$$

Since analyzing policies acting on joint state information or the state history is difficult, optimality has been restricted to the set of Markov policies Π acting on the agent's own state. Although this may seem like a significant restriction, in the $N \to \infty$ limit, the evolution of all other agents – the MF – becomes deterministic and therefore non-informative.

INFINITE AGENT LIMIT. The $N \to \infty$ limit of the N-agent game constitutes its corresponding finite MFG (i.e. with a finite state and action space). It consists of the same elements $\mathcal{T}, \mathcal{X}, \mathcal{U}, p, r, \mu_0$. However, instead of modeling N separate agents, it models a single representative agent and collapses all other agents into their common state distribution, i.e. the MF $\mu = (\mu_t)_{t \in \mathcal{T}} \in \mathcal{M}$ with

$\mu_t : \mathcal{X} \to [0,1]$, where $\mathcal{M}$ is the space of all MFs and μ_0 is given. The deterministic mean field μ replaces the empirical measure of the finite game. Consider a Markov policy $\pi \in \Pi$ as before. For some fixed mean field μ, the evolution of random states x_t and actions u_t begins with $x_0 \sim \mu_0$ and subsequently for all applicable times t follows

$$u_t \sim \pi_t(u_t \mid x_t), \quad x_{t+1} \sim p(x_{t+1} \mid x_t, u_t, \mu_t)\,,$$

and the objective analogously becomes

$$J^\mu(\pi) \equiv \mathbb{E}\left[\sum_{t=0}^{T-1} r(x_t, u_t, \mu_t)\right].$$

The mean field μ induced by some fixed policy π begins with the given μ_0 and is defined recursively by

$$\mu_{t+1}(x') \equiv \sum_{x\in\mathcal{X}} \mu_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) p(x' \mid x, u, \mu_t)\,.$$

By fixing a mean field $\mu \in \mathcal{M}$, we obtain an induced MDP with time-dependent transition function $p(x' \mid x, u, \mu_t)$ and reward function $r(x, u, \mu_t)$. Denote the set-valued map from mean field to optimal policies π of the induced MDP as $\hat{\Phi} : \mathcal{M} \to 2^\Pi$ (i.e. such that π is optimal at any time and state). Analogously, define the map from a policy to its induced mean field as $\Psi : \Pi \to \mathcal{M}$. Finally, we can define the $N \to \infty$ analogue to Markov-Nash equilibria.

Definition 3.1.2. *A MFE is a pair $(\pi, \mu) \in \Pi \times \mathcal{M}$ such that $\pi \in \hat{\Phi}(\mu)$ and $\mu = \Psi(\pi)$ holds.*

By defining any single-valued map $\Phi : \mathcal{M} \to \Pi$ to an optimal policy, we obtain a composition $\Gamma = \Psi \circ \Phi : \mathcal{M} \to \mathcal{M}$, henceforth MFE operator. Shown by [24] for general Polish $\mathcal{X}$ and $\mathcal{U}$, the MFE exists and constitutes an approximate Markov-Nash equilibrium for sufficiently many agents under technical conditions. In Appendix A, we give simplified proofs for finite MFGs under the following standard assumption.

Assumption 3.1.1. *The functions r and p are continuous, hence bounded.*

Note that we metrize probability measure spaces $\mathcal{P}(\mathcal{X})$ with the total variation distance d_{TV}. For probability measures ν, ν' on finite spaces $\mathcal{X}$, d_{TV} simplifies to

$$d_{TV}(\nu, \nu') = \frac{1}{2}\sum_{x\in\mathcal{X}} |\nu(x) - \nu'(x)|\,.$$

Accordingly, we equip $\Pi, \mathcal{M}$ with sup metrics, i.e. for policies $\pi, \pi' \in \Pi$ and MFs $\mu, \mu' \in \mathcal{M}$ we define the metric spaces (Π, d_Π) and $(\mathcal{M}, d_\mathcal{M})$ with

$$d_\Pi(\pi, \pi') \equiv \max_{t\in\mathcal{T}} \max_{x\in\mathcal{X}} d_{TV}(\pi_t(\cdot \mid x), \pi'_t(\cdot \mid x))\,,$$
$$d_\mathcal{M}(\mu, \mu') \equiv \max_{t\in\mathcal{T}} d_{TV}(\mu_t, \mu'_t)\,.$$

Proposition 3.1.1. *Under Assumption 3.1.1, there exists at least one MFE $(\pi^*, \mu^*) \in \Pi \times \mathcal{M}$.*

Theorem 3.1.1. *Under Assumption 3.1.1, if (π^*, μ^*) is an MFE, then for any $\varepsilon > 0$ there exists $N' \in \mathbb{N}$ such that for all $N > N'$, the policy $(\pi^*, \ldots, \pi^*)$ is an ε-Markov-Nash equilibrium in the N-agent game.*

For proofs, see Appendix A. Importantly, finding Nash equilibria in large-N games is hard [45], whereas an MFE can be significantly more tractable to compute. Accordingly, solving the limiting MFG approximately solves the finite-N game for large N in a tractable manner.

3.1.2 *Fixed Point Iteration Fails*

Repeated application of the MFE operator constitutes the exact FPI approach to finding MFE. The standard assumption for convergence in the literature is contractivity and thereby MFE uniqueness (e.g. [95, 139]).

Proposition 3.1.2. *Let Φ, Ψ be Lipschitz with constants c_1, c_2, fulfilling $c_1 c_2 < 1$. Then, the FPI $\mu^{n+1} = \Psi(\Phi(\mu^n))$ converges to the MF of the unique MFE for any initial $\mu^0 \in \mathcal{M}$.*

Proof. Let $\mu, \mu' \in \mathcal{M}$ arbitrary, then

$$\begin{aligned} d_{\mathcal{M}}(\Gamma(\mu), \Gamma(\mu')) &= d_{\mathcal{M}}(\Psi(\Phi(\mu)), \Psi(\Phi(\mu'))) \\ &\leq c_2 \cdot d_{\Pi}(\Phi(\mu), \Phi(\mu')) \\ &\leq c_2 \cdot c_1 \cdot d_{\mathcal{M}}(\mu, \mu') \,. \end{aligned}$$

Since μ, μ' are arbitrary, Γ is Lipschitz with constant $c_1 \cdot c_2 < 1$. (Π, d_{Π}) and $(\mathcal{M}, d_{\mathcal{M}})$ are complete metric spaces (see Appendix A). Therefore, Banach's fixed point theorem implies convergence to the unique fixed point for any starting $\mu^0 \in \mathcal{M}$. □

Unfortunately, it remains unclear how to proceed if multiple optimal policies of an induced MDP exist, or if contractivity fails, e.g. when multiple MFE exist. In the following, consider again the illuminating example from the introduction.

3.1.2.1 *Toy Example*

Consider $\mathcal{X} = \{C, L, R\}, \mathcal{U} = \mathcal{X} \setminus \{C\}$, $\mu_0(C) = 1$, $r(x, u, \mu_t) = -\mathbf{1}_{\{L\}}(x) \cdot \mu_t(L) - \mathbf{1}_{\{R\}}(x) \cdot \mu_t(R)$ and $\mathcal{T} = \{0, 1\}$. The transition function allows picking the next state directly, i.e. for all $x, x' \in \mathcal{X}, u \in \mathcal{U}$,

$$\mathbb{P}(x_{t+1} = x' \mid x_t = x, u_t = u) = p(x' \mid x, u) = \mathbf{1}_{\{x'\}}(u) \,.$$

Clearly, any MFE (π^*, μ^*) must fulfill $\pi_0^*(L \mid C) = \pi_0^*(R \mid C) = 1/2$, while π_1^* can be arbitrary. Even if the operator Φ chooses suitable optimal policies, the fixed point operator Γ remains non-contractive, as the MF will necessarily alternate between left and right for any non-uniform starting $\mu^0 \in \mathcal{M}$.

We observe that the example has infinitely many MFE, but no deterministic MFE, i.e. an MFE such that for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ either $\pi_t(u \mid x) = 0$ or $\pi_t(u \mid x) = 1$ holds, similar to the classical game-theoretical insight of mixed Nash equilibrium existence (cf. [77]). Therefore, choosing optimal, deterministic policies will typically fail.

3.1.2.2 *General Non-Contractivity*

Most existing work assumes contractivity, which is too restrictive. In many scenarios, agents need to "coordinate" with each other. For example, a herd of hunting animals may collectively choose one of multiple hunting grounds, allowing for multiple MFEs. Hence, it can be difficult to apply existing MFG methodologies in practice, as many problems automatically fail contractivity.

From the previous example, we may be led to believe that non-contractivity is a general property of finite MFGs. And indeed, regardless of number of MFEs, it turns out that in any finite MFG with non-constant MFE operator, a policy selection operator Φ with finite image Π_Φ will lead to non-contractivity. Note that this includes both the conventional $\arg\max$ and the $\arg\max$-e (cf. [95]) choice of actions.

Theorem 3.1.2. *Let the image of Φ be a finite set $\Pi_\Phi \subseteq \Pi$. Then, either it holds that $\Gamma = \Psi \circ \Phi$ is a constant, or Γ is not Lipschitz continuous and thus not a contraction.*

Therefore, typical discrete-time finite MFGs have non-contractive fixed point operators and we must change our approach. Note that although non-contractivity does not imply non-convergence, the trivial example from before strongly suggests that non-convergence is the case for many finite MFGs.

3.1.3 *Approximating Mean Field Equilibria Can Help*

Exact FPI fails to solve most finite MFGs. Therefore, a different solution approach is necessary. In the following, we present two related approaches that guarantee convergence while plausibly remaining approximate Nash equilibria in the finite-N case. For our results, we require a stronger Lipschitz assumption that implies Assumption 3.1.1.

Assumption 3.1.2. *The functions r and p are Lipschitz continuous, hence bounded.*

3.1.3.1 *Relative Entropy Mean Field Games*

A straightforward idea is regularization by replacing the objective by the well-known (see e.g. [140]) relative entropy objective

$$\tilde{J}^\mu(\pi) \equiv \mathbb{E}\left[\sum_{t=0}^{T-1} r(x_t, u_t, \mu_t) - \eta \log \frac{\pi_t(u_t \mid x_t)}{q_t(u_t \mid x_t)}\right]$$

with temperature $\eta > 0$ and positive prior policy $q \in \Pi$, i.e. $q_t(u \mid x) > 0$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$. Shown in Appendix A, the unique optimal policy $\tilde{\pi}_t^{\mu,\eta}$ fulfills

$$\tilde{\pi}_t^{\mu,\eta}(u \mid x) = \frac{q_t(u \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,x,u')}{\eta}\right)}$$

for the MDP induced by fixed $\mu \in \mathcal{M}$, with the soft action-value function $\tilde{Q}_\eta(\mu, t, x, u)$ given by the smooth-maximum Bellman recursion

$$\tilde{Q}_\eta(\mu,t,x,u) = r(x,u,\mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \cdot \eta \log\left(\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp \frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right)$$

of the MDP induced by fixed $\mu \in \mathcal{M}$, with terminal condition $\tilde{Q}_\eta(\mu, T-1, x, u) \equiv r(x, u, \mu_{T-1})$. Note that we recover optimality as $\eta \to 0$, see Theorem 3.1.4. Define the relative entropy MFE operator $\tilde{\Gamma}_\eta \equiv \Psi \circ \tilde{\Phi}_\eta$ with policy selection $\tilde{\Phi}_\eta(\mu) \equiv \tilde{\pi}^{\mu,\eta}$ for all $\mu \in \mathcal{M}$.

Definition 3.1.3. *An η-relative entropy MFE (η-RelEnt MFE) for some positive prior policy $q \in \Pi$ is a pair $(\pi^E, \mu^E) \in \Pi \times \mathcal{M}$ such that $\pi^E = \tilde{\Phi}_\eta(\mu^E)$ and $\mu^E = \Psi(\pi^E)$ hold. An η-maximum entropy MFE (η-MaxEnt MFE) is an η-RelEnt MFE with uniform prior policy q.*

3.1.3.2 *Boltzmann Iteration*

Since only deterministic policies fail, a derivative approach is to use softmax policies directly with the unregularized action-value function, also called Boltzmann policies. Assume that the action-value function Q^* fulfilling the Bellman equation

$$Q^*(\mu, t, x, u) = r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \cdot \max_{u' \in \mathcal{U}} Q^*(\mu, t+1, x', u') \,.$$

of the MDP induced by fixed $\mu \in \mathcal{M}$ with terminal condition $Q^*(\mu, T-1, x, u) \equiv r(x, u, \mu_{T-1})$ is known. Define the map $\Phi_\eta(\mu) \equiv \pi^{\mu,\eta}$ for any $\mu \in \mathcal{M}$, where

$$\pi_t^{\mu,\eta}(u \mid x) \equiv \frac{q_t(u \mid x) \exp\left(\frac{Q^*(\mu,t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{Q^*(\mu,t,x,u')}{\eta}\right)}$$

for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ and temperature $\eta > 0$.

Definition 3.1.4. *An η-Boltzmann MFE (η-Boltzmann MFE) for some positive prior policy $q \in \Pi$ is a pair $(\pi^B, \mu^B) \in \Pi \times \mathcal{M}$ such that $\pi^B = \Phi_\eta(\mu^B)$ and $\mu^B = \Psi(\pi^B)$ hold.*

3.1.3.3 *Theoretical Properties*

Both η-RelEnt MFE and η-Boltzmann MFE are guaranteed to exist for any temperature $\eta > 0$.

Proposition 3.1.3. *Under Assumption 3.1.1, η-Boltzmann and η-RelEnt MFE exist for any temperature $\eta > 0$.*

Contractivity of both η-Boltzmann MFE operator $\Gamma_\eta \equiv \Psi \circ \Phi_\eta$ and η-RelEnt MFE operator $\tilde{\Gamma}_\eta \equiv \Psi \circ \tilde{\Phi}_\eta$ is guaranteed for sufficiently high temperatures, even if all possible original Φ are not Lipschitz continuous.

Theorem 3.1.3. *Under Assumption 3.1.2, $\mu \mapsto Q^*(\mu, t, x, u)$, $\mu \mapsto \tilde{Q}_\eta(\mu, t, x, u)$ and $\Psi(\pi)$ are Lipschitz continuous with constants K_{Q^*}, $K_{\tilde{Q}}$ and K_Ψ for arbitrary $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}, \eta > \eta', \eta' > 0$. Furthermore, Γ_η and $\tilde{\Gamma}_\eta$ are a contraction for*

$$\eta > \max\left(\eta', \frac{|\mathcal{U}|\,(|\mathcal{U}|-1) K_Q K_\Psi q_{\max}^2}{2 q_{\min}^2}\right)$$

where $K_Q = K_{Q^*}$ for Γ_η, $K_Q = K_{\tilde{Q}}$ for $\tilde{\Gamma}_\eta$, $q_{\max} \equiv \max_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}} q_t(u \mid x) > 0$ and $q_{\min} \equiv \min_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}} q_t(u \mid x) > 0$.

Sufficiently large η hence implies convergence via FPI. On the other hand, for sufficiently low temperatures η, both η-Boltzmann and η-RelEnt MFE will also constitute an approximate Markov-Nash equilibrium of the finite-N game.

Theorem 3.1.4. *Under Assumption 3.1.2, if $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ is a sequence of η_n-Boltzmann or η_n-RelEnt MFE with $\eta_n \to 0$, then for any $\varepsilon > 0$ there exist $n', N' \in \mathbb{N}$ such that for all $n > n', N > N'$, the policy $(\pi_n^*, \dots, \pi_n^*) \in \Pi^N$ is an ε-Markov-Nash equilibrium of the N-agent game, i.e.*

$$J_i^N(\pi_n^*, \dots, \pi_n^*) \geq \max_{\pi_i \in \Pi} J_i^N(\pi_n^*, \dots, \pi_n^*, \pi_i, \pi_n^*, \dots, \pi_n^*) - \varepsilon\,.$$

If we can obtain contractivity for sufficiently low η, we can find good approximate Markov-Nash equilibria. As it is impossible to have both $\eta \to 0$ and $\eta \to \infty$, it depends on the problem and prior whether we can converge to a good solution. Nonetheless, we find that it is often possible to empirically find low η that provide convergence as well as a good approximate MFE.

3.1.3.4 *Prior Descent*

In principle, we can insert arbitrary prior policies $q \in \Pi$. Under Assumption 3.1.1, by boundedness of both $\tilde{Q}_\eta$ and Q^* (see Appendix A), both η-RelEnt and η-Boltzmann MFE policies converge to the prior policy as $\eta \to \infty$. Therefore, in principle we can show that for any $\varepsilon > 0$, for sufficiently large η and N, the η-RelEnt and η-Boltzmann MFE under q will be at most an ε-worse approximate Nash equilibrium than the prior policy. Furthermore, we obtain guaranteed contractivity by Theorem 3.1.3. Thus, any prior policy gives a worst-case bound on the performance achievable over all $\eta > 0$. On the other hand, if we obtain better results for sufficiently low η, we may iteratively improve our policy and thus our equilibrium quality.

3.1.4 *Relation to Prior Work*

The original work of [23] introduces contractivity and uniqueness assumptions into the continuous MFG setting. Analogously, [95] and [139] assume contractivity for discrete-time MFGs and dense graph limit MFGs respectively. Further existing work on discrete-time MFGs similarly assumes uniqueness of the MFE, which includes [24] and [141] for approximate optimality and existence results, and [135] for an analysis on contractivity requirements. [114] solve discrete-time continuous state MFG problems under the classical uniqueness conditions of [22]. Further extensions of the MFG formula include partial observability [112] or major agents [113].

The work of [142] is related and studies theoretical properties of finite-N regularized games and their limiting MFG. In their work, the existence and approximate Nash property of MFE in stationary regularized games is shown, and Q-Learning error propagation is investigated. In comparison, we consider the original, unregularized finite-N game in a transient setting and perform extensive empirical evaluations. [95] and [120] previously proposed to apply Boltzmann policies. The former applies the approximation without analyzing the resulting contractivity, while the latter focuses on directly solving finite-N games.

An orthogonal approach to computing MFE is FP. Rooted in game-theory and classical economic works [78], it has since been adapted to MFGs. In FP, all past MFs [94] and policies [127] are averaged to produce a new MF or policy. Importantly, convergence is guaranteed in certain special cases only (cf. [96]). Although introduced in a continuous-time setting, we evaluate FP empirically in our setting and find that both our regularization and FP may be combined successfully.

3.1.5 *Experiments*

In practice, we find that our approaches are capable of generating solutions of lower exploitability than otherwise obtained. Unless stated otherwise, we compute everything exactly, use the maximum entropy objective (MaxEnt) with the uniform prior policy q where $q_t(u \mid x) = 1/|\mathcal{U}|$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$, and initialize with $\mu^0 = \Psi(q)$ generated by q. As the main evaluation metric, we define the exploitability of a policy $\pi \in \Pi$ with induced MF $\mu \equiv \Psi(\pi)$ as

$$\Delta J(\pi) \equiv \max_{\pi^*} J^{\mu}(\pi^*) - J^{\mu}(\pi) \, .$$

Clearly, the exploitability of π is zero if and only if (π, μ) is an MFE. Indeed, for any $\varepsilon > 0$, any policy $\pi \in \Pi$ is a $(\Delta J(\pi) + \varepsilon)$-Markov Nash equilibrium if N sufficiently large, i.e. the exploitability translates directly to the limiting equilibrium quality in the finite-N game, see also Theorem 3.1.4 and its proof.

We evaluate the algorithms on the LR, RPS, SIS and Taxi problems, ordered in increasing complexity. Details of the algorithms, hyperparameters, problems and experiment configurations as well as further experimental results can be found in Appendix A.

3.1.5.1 *Exploitability*

In Figure 3.1, we plot the minimum, maximum and mean exploitability for varying temperatures η during the last 10 fixed point iterations, i.e. a single value when the exploitability (and usually MF) converges. Observe that the lowest convergent temperature outperforms not only the exact FPI (drawn at temperature zero), but also the uniform prior policy.

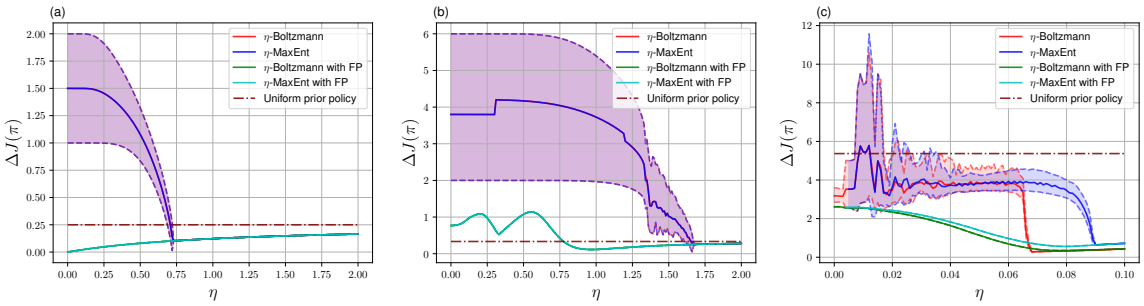


FIGURE 3.1: Convergence in exploitability of regularized MFG algorithms. Mean exploitability over the final 10 iterations. Dashed lines represent maximum and minimum over the final 10 iterations. (a) LR, 10000 iterations; (b) RPS, 10000 iterations; (c) SIS, 10000 iterations. Maximum entropy (MaxEnt) results begin at higher temperatures due to limited floating point accuracy. Temperature zero depicts the exact FPI for both η-MaxEnt and η-Boltzmann MFE. In LR and RPS, η-MaxEnt and η-Boltzmann MFE coincide both with and without FP, here averaging both policy and MF over all past iterations. The exploitability of the prior policy is indicated by the dashed horizontal line.

Although developed for a different setting, we also show results of FP similar to the version from [127], i.e. both policies and MFs are averaged over all past iterations. It can be seen that FP only converges to the optimal solution in the LR problem. In the other examples, supplementing FP with entropy regularization is effective at producing better results. A non-existent FP variant averaging only the policies finds the exact MFE in RPS, but nevertheless fails in SIS. See Appendix A for further results.

Evaluating and solving finite-N games is highly intractable by the curse of dimensionality, as the local state is no longer sufficient to perform dynamic programming in the presence of the random empirical state measure. Since it has already been proven that the exploitability for $N \to \infty$ will converge to the exploitability of the corresponding MFG, we refrain from evaluating on finite-N games.

Note that the plots are entirely deterministic and not stochastic as it would seem at first glance, since the depicted shaded area visualizes the non-convergence of exploitability and is a result of the fixed point updates running into a limit cycle (cf. Figure 3.2).

3.1.5.2 *Convergence*

In Figure 3.2, the difference between the exploitability of the current policy and the minimal exploitability reached during the final 10 iterations is shown for η-Boltzmann MFE. As the temperature η decreases, time to convergence increases until non-convergence is reached in form of a limit cycle. Analogous results for η-RelEnt MFE can be found in Appendix A.

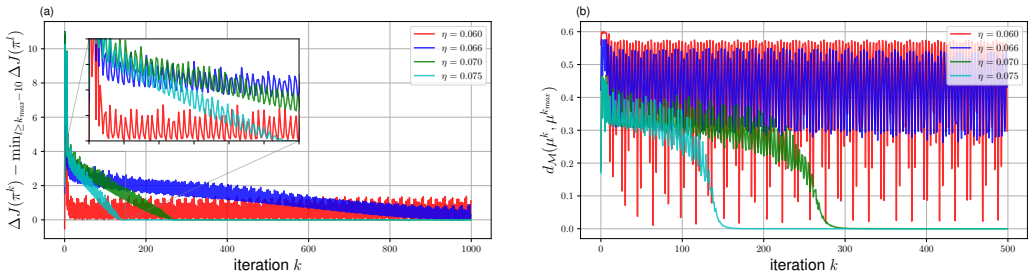


FIGURE 3.2: Change in exploitability and MF over iterations. (a) Difference between current and final minimum exploitability over the last 10 iterations; (b) Distance between current and final MF. Plotted for the η-Boltzmann MFE iterations in SIS for different indicated temperature settings. Note the periodicity of the lowest temperature setting, indicating a limit cycle.

Note also that in LR, we can analytically find $K_Q = 1$ and $K_\Psi = 1$. Thus, we obtain guaranteed convergence via η-Boltzmann MFE iteration if $\eta > 1$. In Figure 3.1, we see convergence already for $\eta \geq 0.7$. Note further that the non-converged regime can allow for lower exploitability. However, it is unclear a priori when to stop, and for approximate solutions where DQN is used for evaluation, the evaluation of exploitability may become inaccurate.

3.1.5.3 *Deep Reinforcement Learning*

For problems with intractably large state spaces, we adopt the DQN algorithm [33], using the implementation of [143] as a base. Particle-based simulations are used for the MF, and stochastic performance evaluation on the induced MDP is performed (see Appendix A). Note that the approximation introduces three sources of stochasticity into the otherwise deterministic algorithms, i.e. stochastic evaluation, MF simulation and DQN. To counteract the randomness, we average our results over multiple runs. The hyperparameters and architectures used are standard and can be found in Appendix A.

Fitting the soft action-value function directly using a network is numerically problematic, as the log-exponential transformation of approximated action-values quickly fails due to limited floating point accuracy. Thus, we limit ourselves to the classical Bellman equation with Boltzmann policies only.

In Figure 3.3, we evaluate the exploitability of Boltzmann DQN iteration, evaluated exactly in SIS and RPS, and stochastically in Taxi over 2000 realizations. Minimum, maximum and mean exploitability are taken over the final 5 iterations and averaged over 5 seeds. Note that it is very time-consuming to solve a full RL problem using DQN repeatedly in every iteration. Nonetheless, we observe that a temperature larger than zero appears to improve exploitability and convergence in the SIS example. Both due to the noisy nature of approximate solutions and the lower number of iterations, it can be seen that a higher temperature is required to converge than in the exact case.

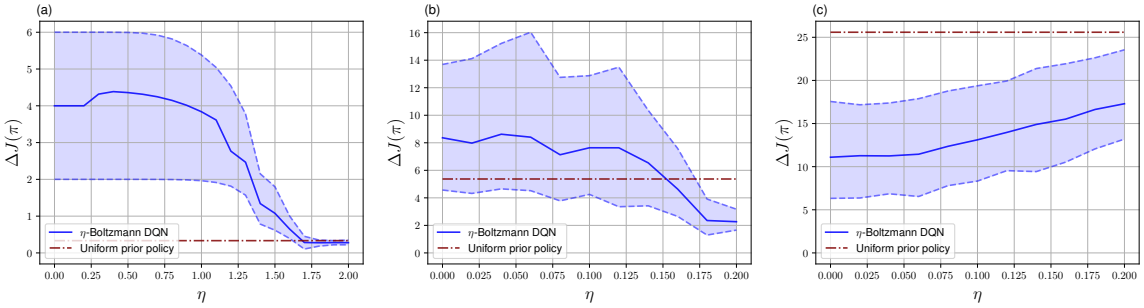


FIGURE 3.3: Convergence in exploitability of deep RL-based MFG algorithms. Mean exploitability over the final 5 iterations using DQN, averaged over 5 seeds. Dashed lines represent the averaged maximum and minimum exploitability over the last 5 iterations. (a) RPS, 1000 iterations; (b) SIS, 50 iterations; (c) Taxi, 15 iterations. Evaluation of exploitability is exact except in Taxi, which uses DQN and averages over 1000 episodes. The point of zero temperature depicts FPI using exact DQN policies.

In the intractable Taxi environment, the policy oscillates between two modes as in exact LR, and regularization fails to obtain better results, see also Appendix A. An important reason is that the prior policy performs extremely bad (exploitability of ~ 35) as most states require specific actions for optimality. Hence we cannot find an $\eta > 0$ for which the algorithm both converges and performs well. Using prior descent and iteratively refining a better prior policy would likely increase performance, but is deferred to future investigations as the required computations grow very large.

Fictitious play is expensive in combination with approximate Q-Learning and particle simulations, as policies and particles of past iterations must be kept to perform exact FP. For this reason, we do not attempt approximate FP with approximate solution methods. In theory, supervised learning for fitting summarizing policies and randomly sampling particles may help, but is out of scope of this work.

3.1.5.4 *Prior Descent*

In Figure 3.4, we repeatedly perform outer iterations consisting of 100 η-RelEnt MFE iterations each with the indicated fixed temperature parameters in SIS. After each outer iteration, the prior policy is updated to the newest resulting policy. Note again that the results are entirely deterministic.

Searching for a suitable η dynamically every iteration would keep the exploitability from increasing, as for $\eta \to \infty$ we obtain the original prior policy. Since it is expensive to scan over all temperatures in each outer iteration, we use a heuristic. Intuitively, since the prior will become increasingly good, it will be increasingly difficult to obtain a better policy. Thus, increasing the temperature will help sticking close to the prior and converge. Consequently, we use the simple heuristic

$$\eta_{i+1} = \eta_i \cdot c$$

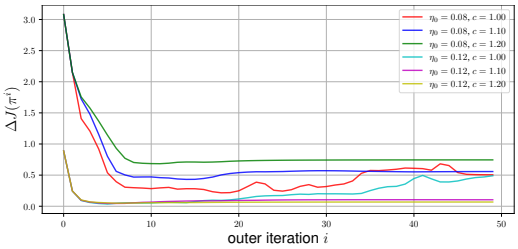


FIGURE 3.4: Convergence in exploitability of prior iteration algorithm. Exploitability over outer iterations in SIS, using 100 η-RelEnt MFE iterations per outer iteration. Note that the results are deterministic. Not shown: Running the fixed temperature settings $c = 1$ for longer does not converge for at least 1000 iterations.

for each outer iteration i, where $c \geq 1$ adjusts the temperature after each outer iteration.

Importantly, even for our simple heuristic, prior descent already achieves an exploitability of ~ 0.068, whereas the best results for the fixed uniform policy from Figure 3.1 show an optimal mean exploitability of ~ 0.281. Furthermore, repeated prior policy updates succeed in computing the exact MFE in RPS and LR under a fixed temperature (see Appendix A).

Note that prior descent creates a double loop around solving the optimal control problem, becoming highly expensive under deep RL. Hence, we refrain from prior descent with DQN. Automatically adjusting temperatures to monotonically improve exploitability is left for potential future work.

3.1.6 *Summary*

In this work, we have investigated the necessity and feasibility of approximate MFG solution approaches – entropy regularization, Boltzmann policies and prior descent – in the context of finite MFGs. We have shown that the finite MFG case typically cannot be solved by exact FPI or FP alone. Entropy regularization and Boltzmann policies in combination with deep RL may enable feasible computation of approximate MFE. We believe that lifting the restriction of inherent contractivity is an important step in ensuring applicability of MFG models in practical problems.

For future work, an efficient, automatic temperature adjustment for prior descent could be fruitful. Furthermore, it would be interesting to generalize relative entropy MFGs to infinite horizon discounted problems, continuous time, and continuous state and action spaces. Moreover, it could be of interest to investigate theoretical properties of FP in finite MFGs in combination with entropy regularization. For non-Lipschitz mappings from policy to induced MF, the proposed approach does not provide a solution. It could nonetheless be important to consider problems with threshold-type dynamics and rewards, e.g. majority vote problems. Most notably, the current formalism precludes common noise entirely, i.e. any games with common states. In practice, many problems will allow for some type of common states between agents, leading to non-independent agent distributions and stochastic as opposed to deterministic MFs, see Section 3.4.

3.2 LEARNING MEAN FIELD GAMES ON GRAPHS

Recent advances at the intersection of dense large graph limits and MFGs have begun to enable the scalable analysis of a broad class of dynamical sequential games with large numbers of agents. Results had been largely limited to graphon MF systems with continuous-time diffusive or jump dynamics, typically without control and with little focus on computational methods. We propose a novel discrete-time formulation for GMFGs as the limit of non-linear dense graph Markov games with weak interaction. On the theoretical side, we give extensive and rigorous existence and approximation properties of the graphon MF solution in sufficiently large systems. On the practical side, we provide general learning schemes for graphon MFE by either introducing agent equivalence classes or reformulating the graphon MF system as a classical MF system. By repeatedly finding a regularized optimal control solution and its generated MF, we successfully obtain plausible approximate Nash equilibria in otherwise infeasible large dense graph games with many agents. Empirically, we are able to demonstrate on a number of examples that the finite-agent behavior comes increasingly close to the MF behavior for our computed equilibria as the graph or system size grows, verifying our theory. More generally, we successfully apply policy gradient RL in conjunction with sequential Monte Carlo methods. The material presented in this section is based upon our work [7]. Further extensions to more sparse and weighted or directed graphs in our collaborations [13, 14, 19] are not presented in this thesis.

MEAN FIELD SYSTEMS ON GRAPHS. For MF systems on dense graphs, prior work mostly considers MF systems without control [144] or time-dynamics, i.e. the static case [145, 146]. In contrast to our work, [119] consider states instead of agents on a graph, while [120] requires restrictive assumptions and considers average actions instead of distributions of the neighbors. To the best of our knowledge, [147] and [139] are the first to consider general continuous-time diffusion-type graphon MF systems with control, the latter proposing many clusters of agents as well as proving an approximate Nash property as the number of clusters and agents grows. There have since been efforts to control cooperative graphon MF systems with diffusive linear dynamics using spectral methods [148, 149]. On the other hand, [150, 151] consider large non-clustered systems in a continuous-time diffusion-type setting without control, while [152] and [153] consider continuous-time linear-quadratic systems and continuous-time jump processes respectively. To the best of our knowledge, only [154] have considered solving and formulating a GMFGs in discrete time, though requiring analytic computation of an infinite-dimensional value function defined over all MFs and thus being inapplicable to arbitrary problems in a black-box, learning manner. In contrast, we give a general learning scheme and also provide extensive theoretical analysis of our algorithms and (slightly different) model. Finally, for sparse graphs there exist preliminary results [155, 156] including also our collaborations [13, 14, 19], though the setting remains to be developed.

OUR CONTRIBUTION. In this work, we propose a dense graph limit extension of MFGs in discrete time, combining graphon MF systems with MFGs. More specifically, we consider limits of many-agent systems with discrete-time graph-based dynamics and weak neighbor interactions. In contrast to prior works, we consider one of the first general discrete-time formulations as well as its controlled case, which is a natural setting for many problems that are inherently discrete in time or to be controlled digitally at discrete decision times. Our contribution can be summarized as: (i) formulating one of the first general discrete-time GMFGs frameworks for approximating otherwise intractable large dense graph games; (ii) providing an extensive theoretical analysis of existence and approximation properties in such systems; (iii) providing general learning schemes for finding graphon MFE, and (iv) empirically evaluating our proposed approach with verification of

theoretical results in the finite N-agent graph system, finding plausible approximate Nash equilibria for otherwise infeasible large dense graph games with many agents.

3.2.1 *Mean Field Games on Dense Graphs*

In the following, we will give a dense graph N-agent model as well as its corresponding MF system, where agents are affected only by the overall state distribution of all neighbors, as visualized in Figure 3.5. As a result of the LLN, this distribution will become deterministic – the MF – as $N \to \infty$. We begin with graph-theoretical preliminaries, see also [157] for a review. The study of dense large graph limits deals with the limiting representation of adjacency matrices called graphons. Define $\mathcal{I} := [0,1]$ and $\mathcal{W}_0$ as the space of all bounded, symmetric and measurable functions (graphons) $W \in \mathcal{W}_0$, $W : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$ bounded by $0 \leq W \leq 1$. For any simple graph $G = (\{1,\dots,N\}, \mathcal{E})$, we define its step-graphon a.e. uniquely by

$$W_G(x,y) = \sum_{i,j \in \{1,\dots,N\}} \mathbf{1}_{(i,j)\in\mathcal{E}} \cdot \mathbf{1}_{x \in (\frac{i-1}{N}, \frac{i}{N}]} \cdot \mathbf{1}_{y \in (\frac{j-1}{N}, \frac{j}{N}]}, \tag{3.2.1}$$

see e.g. Figure 3.5. We equip $\mathcal{W}_0$ with the cut (semi-)norm $\|\cdot\|_\square$ and cut (pseudo-)metric $\delta_\square$

$$\|W\|_\square := \sup_{S,T} \left| \int_{S\times T} W(x,y)\,\mathrm{d}x\,\mathrm{d}y \right|, \quad \delta_\square(W,W') := \inf_\varphi \|W - W'_\varphi\|_\square, \tag{3.2.2}$$

for graphons $W, W' \in \mathcal{W}_0$ and $W'_\varphi(x,y) := W'(\varphi(x),\varphi(y))$, where the supremum is over all measurable subsets $S,T \subseteq \mathcal{I}$ and the infimum is over measure-preserving bijections $\varphi : \mathcal{I} \to \mathcal{I}$.

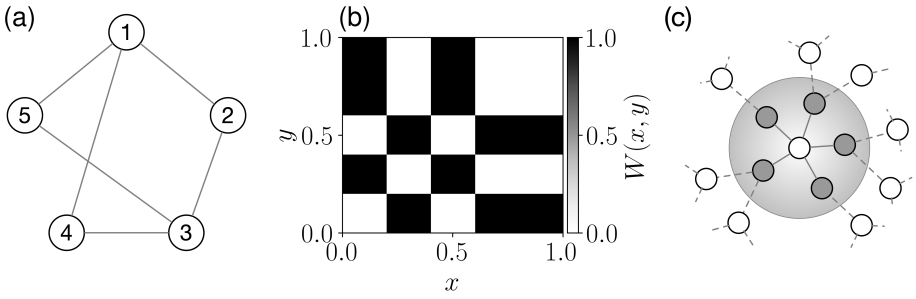


FIGURE 3.5: Visualization of graphical interactions. **(a)**: A graph with 5 nodes; **(b)**: The associated step graphon of the graph in (a) as a continuous domain version of its adjacency matrix; **(c)**: A visualization of the dynamics, i.e. the center agent is affected only by its neighbors (grey).

To provide motivation, note that convergence in $\delta_\square$ is equivalent to e.g. convergence of probabilities of locally encountering any fixed subgraph by randomly sampling a subset of nodes. Many such properties of graph sequences $(G_N)_{N\in\mathbb{N}}$ converging to some graphon $W \in \mathcal{W}_0$ can then be described by W, and we point to [157] for details. In this work, we will primarily use the analytical fact that for converging graphon sequences $\|W_{G_N} - W\|_\square \to 0$, we equivalently have

$$\|W_{G_N} - W\|_{L_\infty \to L_1} = \sup_{\|g\|_\infty \leq 1} \int_{\mathcal{I}} \left| \int_{\mathcal{I}} (W_{G_N}(\alpha,\beta) - W(\alpha,\beta)) g(\beta)\,\mathrm{d}\beta \right| \mathrm{d}\alpha \to 0 \tag{3.2.3}$$

under the operator norm of operators $L_\infty \to L_1$, see e.g. [157], Lemma 8.11.

By [157], Theorem 11.59, the above is equivalent to convergence in the cut metric $\delta_\square(W_{G_N}, W) \to 0$ up to relabeling. In the following, we will therefore assume sequences of simple graphs $G_N = (\mathcal{V}_N, \mathcal{E}_N)$ with vertices $\mathcal{V}_N = \{1,\dots,N\}$, edge sets $\mathcal{E}_N$, edge indicator variables $\xi^N_{i,j} := \mathbf{1}_{(i,j)\in\mathcal{E}_N}$ for all nodes $i,j \in \mathcal{V}_N$, and associated step graphons W_N converging in cut norm.

Assumption 3.2.1. *The sequence of step-graphons $(W_N)_{N\in\mathbb{N}}$ converges in cut norm $\|\cdot\|_\square$ or equivalently in operator norm $\|\cdot\|_{L_\infty\to L_1}$ as $N\to\infty$ to some graphon $W\in\mathcal{W}_0$, i.e.*

$$\|W_N - W\|_\square \to 0, \quad \|W_N - W\|_{L_\infty\to L_1} \to 0\,. \tag{3.2.4}$$

Next, we define W-random graphs to consist of vertices $\mathcal{V}_N \coloneqq \{1,\dots,N\}$ with adjacency matrices $\boldsymbol{\xi}^N$ generated by sampling graphon indices α_i uniformly from $\mathcal{I}$ and edges $\xi^N_{i,j} \sim \text{Bernoulli}(W(\alpha_i,\alpha_j))$ for all vertices $i,j\in\mathcal{V}_N$. For experiments, by [157], Lemma 10.16, we can thereby generate a.s. converging graph sequences by sampling W-random graphs for any fixed graphon $W\in\mathcal{W}_0$. In principle, one could also consider arbitrary graph generating processes whenever a valid relabeling function φ is known.

In our work, the usage of graphons enables us to find MF systems on dense graphs and to extend the expressiveness of classical MFGs. As examples, we will use the limiting graphons of uniform attachment, ranked attachment and Erdős–Rényi (ER) random graphs given by $W_{\text{unif}}(x,y) = 1-\max(x,y)$, $W_{\text{rank}}(x,y) = 1-xy$ and $W_{\text{er}}(x,y) = p$ respectively [157, 158], each of which exhibits different node connectivities as shown in Figure 3.6.

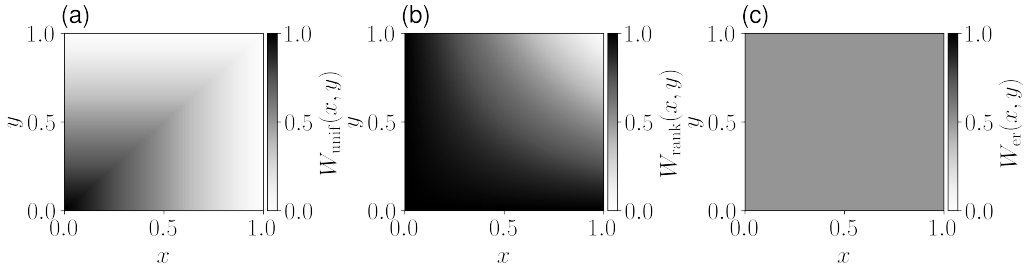


FIGURE 3.6: Three example graphons used in our experiments. **(a)**: Uniform attachment graphon; **(b)**: Ranked attachment graphon; **(c)**: ER graphon with edge probability 0.5.

FINITE GRAPH GAME. For simplicity of analysis, we consider finite state and action spaces $\mathcal{X},\mathcal{U}$ as well as times $\mathcal{T}\coloneqq\{0,1,\dots,T-1\}$. On a metric space $\mathcal{A}$, define the spaces of all Borel probability measures $\mathcal{P}(\mathcal{A})$ and all Borel measures $\mathcal{B}_1(\mathcal{A})$ bounded by 1, equipped with the L_1 norm. For simplified notation, we denote both a measure ν and its probability mass function by ν. Define the space of policies $\Pi\coloneqq\mathcal{P}(\mathcal{U})^{\mathcal{T}\times\mathcal{X}}$, i.e. agents apply Markovian feedback policies $\pi^i=(\pi^i_t)_{t\in\mathcal{T}}\in\Pi$ that act on local state information. This allows for the definition of weakly interacting agent state and action random variables

$$x^i_0\sim\mu_0, \quad u^i_t\sim\pi^i_t(u^i_t\mid x^i_t), \quad x^i_{t+1}\sim p(x^i_{t+1}\mid x^i_t,u^i_t,\mathbb{G}^i_t), \quad \forall t\in\mathcal{T},\forall i\in\mathcal{V}_N \tag{3.2.5}$$

under some transition kernel $p\colon\mathcal{X}\times\mathcal{U}\times\mathcal{B}_1(\mathcal{X})\to\mathcal{P}(\mathcal{X})$, where the empirical neighborhood MF $\mathbb{G}^i_t$ of agent i is defined as the $\mathcal{B}_1(\mathcal{X})$-valued (unnormalized) neighborhood state distribution

$$\mathbb{G}^i_t \coloneqq \frac{1}{N}\sum_{j\in\mathcal{V}_N}\xi^N_{i,j}\delta_{x^j_t}, \tag{3.2.6}$$

where δ is the Dirac measure, i.e. each agent affects each other at most negligibly with factor $1/N$. Finally, for each agent i we define separate, competitive objectives

$$J^N_i(\pi^1,\dots,\pi^N)\coloneqq\mathbb{E}\left[\sum_{t=0}^{T-1} r(x^i_t,u^i_t,\mathbb{G}^i_t)\right] \tag{3.2.7}$$

to be maximized over π^i, where $r:\mathcal{X}\times\mathcal{U}\times\mathcal{B}_1(\mathcal{X})\to\mathbb{R}$ is an arbitrary reward function.

Remark 3.2.1. *We can also consider infinite horizons, under the alternative objective* $\tilde{J}_i^N(\pi^1,\dots,\pi^N) \equiv \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(x_t^i,u_t^i,\mathbb{G}_t^i)\right]$ *with all results but Theorem 3.2.4 holding. One may also extend to state-action distributions, heterogeneous starting conditions and time-dependent* r, P, *though we avoid this for expositional simplicity.*

Remark 3.2.2. *Note that it is straightforward to extend to heterogeneous agents by modelling agent types as part of the agent state, see also e.g. [105]. It is only required to model agent states in a unified manner, which does not imply that there can be no heterogeneity.*

With this, we can give a typical notion of Nash equilibria as found e.g. in [24]. However, under graph convergence in Assumption 3.2.1, it is always possible for a finite number of nodes to have an arbitrary neighborhood differing from the graphon as $N \to \infty$. Thus, it is impossible to show approximate optimality for all nodes and only possible to show for an increasingly large fraction $1-\delta \approx 1$ of nodes. For this reason, we slightly weaken the notion of Nash equilibria by restricting to a fraction $1-\delta$ of agents, as e.g. considered in [96, 159].

Definition 3.2.1. *An* (ε,δ)*-Markov-Nash equilibrium (almost Markov-Nash equilibrium) for* $\varepsilon,\delta>0$ *is defined as a tuple of policies* $(\pi^1,\dots,\pi^N)\in\Pi^N$ *such that for any* $i\in\mathcal{J}_N$*, we have*

$$J_i^N(\pi^1,\dots,\pi^N) \geq \sup_{\pi\in\Pi} J_i^N(\pi^1,\dots,\pi^{i-1},\pi,\pi^{i+1},\dots,\pi^N) - \varepsilon, \tag{3.2.8}$$

for some set $\mathcal{J}_N\subseteq\mathcal{V}_N$ *containing at least* $\lfloor(1-\delta)N\rfloor$ *agents, i.e.* $|\mathcal{J}_N|\geq\lfloor(1-\delta)N\rfloor$.

The minimal such $\varepsilon>0$ for any fixed policy tuple (and typically $\delta=0$) is also called its exploitability. Whilst we ordain ε-optimality only for a fraction $1-\delta$ of agents, if the fraction δ is negligible, it will have negligible impact on other agents as a result of the weak interaction property. Thus, the solution will remain approximately optimal for almost all agents for sufficiently small δ regardless of the behavior of that fraction δ of agents. In the following, we will give a limiting system that shall provide (ε,δ)-Markov-Nash equilibria with $\varepsilon,\delta\to0$ as $N\to\infty$.

GRAPHON MEAN FIELD GAME. The formal $N\to\infty$ limit of the N-agent game constitutes its GMFG, which shall be rigorously justified in Section 3.2.2. We define the space of measurable state marginal ensembles $\mathcal{M}_t \coloneqq \mathcal{P}(\mathcal{X})^{\mathcal{I}}$ and measurable MF ensembles $\boldsymbol{\mathcal{M}} \coloneqq \mathcal{P}(\mathcal{X})^{\mathcal{T}\times\mathcal{I}}$, in the sense that $\alpha\mapsto\mu_t^\alpha(x)$ is measurable for any $\boldsymbol{\mu}\in\boldsymbol{\mathcal{M}}$, $t\in\mathcal{T}$, $x\in\mathcal{X}$. Similarly, we define the space of measurable policy ensembles $\boldsymbol{\Pi}\subseteq\Pi^{\mathcal{I}}$, i.e. with measurable $\alpha\mapsto\pi_t^\alpha(u\mid x)$ for any $\boldsymbol{\pi}\in\boldsymbol{\Pi}, t\in\mathcal{T}, x\in\mathcal{X}, u\in\mathcal{U}$.

In the GMFG, we will consider infinitely many agents $\alpha\in\mathcal{I}$ instead of the finitely many $i\in\mathcal{V}_N$. As a result, we will have infinitely many policies $\pi^\alpha\in\Pi$ – one for each agent α – through some measurable policy ensemble $\boldsymbol{\pi}\in\boldsymbol{\Pi}$. We again define state and action random variables

$$x_0^\alpha\sim\mu_0,\quad u_t^\alpha\sim\pi_t^\alpha(u_t^\alpha\mid x_t^\alpha),\quad x_{t+1}^\alpha\sim p(x_{t+1}^\alpha\mid x_t^\alpha,u_t^\alpha,\mathbb{G}_t^\alpha),\quad \forall(\alpha,t)\in\mathcal{I}\times\mathcal{T} \tag{3.2.9}$$

where we introduce the (now deterministic) $\mathcal{B}_1(\mathcal{X})$-valued neighborhood MF of agents α as

$$\mathbb{G}_t^\alpha \coloneqq \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^\beta\,\mathrm{d}\beta \tag{3.2.10}$$

for some deterministic $\boldsymbol{\mu} \in \boldsymbol{\mathcal{M}}$. Under fixed $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, μ_t^α should be understood as the law of x_t^α, $\mu_t^\alpha \equiv \mathcal{L}(x_t^\alpha)$. Finally, define the maximization objective of agent α over π^α for fixed $\boldsymbol{\mu} \in \boldsymbol{\mathcal{M}}$ as

$$J_\alpha^{\boldsymbol{\mu}}(\pi^\alpha) \equiv \mathbb{E}\left[\sum_{t=0}^{T-1} r(x_t^\alpha, u_t^\alpha, \mathbb{G}_t^\alpha)\right]. \tag{3.2.11}$$

To formulate the limiting version of Nash equilibria, we define a map $\Psi\colon \boldsymbol{\Pi} \to \boldsymbol{\mathcal{M}}$ mapping from a policy ensemble $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ to the corresponding generated MF ensemble $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi}) \in \boldsymbol{\mathcal{M}}$ by

$$\mu_0^\alpha \equiv \mu_0, \quad \mu_{t+1}^\alpha(x') \equiv \sum_{x\in\mathcal{X}} \mu_t^\alpha(x) \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) p(x' \mid x, u, \mathbb{G}_t^\alpha), \quad \forall \alpha \in \mathcal{I} \tag{3.2.12}$$

where integrability in Eq. (3.2.10) holds by induction, and note how then $\mu_t^\alpha = \mathcal{L}(x_t^\alpha)$.

Similarly, let $\Phi\colon \boldsymbol{\mathcal{M}} \to 2^{\boldsymbol{\Pi}}$ map from a MF ensemble $\boldsymbol{\mu}$ to the set of optimal policy ensembles $\boldsymbol{\pi}$ characterized by $\pi^\alpha \in \arg\max_{\pi\in\boldsymbol{\Pi}} J_\alpha^{\boldsymbol{\mu}}(\pi)$ for all $\alpha \in \mathcal{I}$, which is particularly fulfilled if $\pi_t^\alpha(u \mid x) > 0 \implies u \in \arg\max_{u'\in\mathcal{U}} Q_\alpha^{\boldsymbol{\mu}}(t, x, u')$ for all $\alpha \in \mathcal{I}$, $t \in \mathcal{T}$, $x \in \mathcal{X}$, $u \in \mathcal{U}$, where $Q_\alpha^{\boldsymbol{\mu}}$ is the optimal action value function under fixed $\boldsymbol{\mu} \in \boldsymbol{\mathcal{M}}$ following the Bellman equation

$$Q_\alpha^{\boldsymbol{\mu}}(t, x, u) = r(x, u, \mathbb{G}_t^\alpha) + \sum_{x'\in\mathcal{X}} p(x' \mid x, u, \mathbb{G}_t^\alpha) \arg\max_{u'\in\mathcal{U}} Q_\alpha^{\boldsymbol{\mu}}(t+1, x', u') \tag{3.2.13}$$

with $Q_\alpha^{\boldsymbol{\mu}}(T, x, u) \equiv 0$ and generally time-dependent, see [68] for a review.

We can now define the GMFG version of Nash equilibria as policy ensembles $\boldsymbol{\pi}$ generating MF ensembles $\boldsymbol{\mu}$ under which they are optimal, as $\mu_t^\alpha = \mathcal{L}(x_t^\alpha)$ if all agents $\alpha \in \mathcal{I}$ follow π^α.

Definition 3.2.2. *A Graphon Mean Field Equilibrium (GMFE) is a pair $(\boldsymbol{\pi}, \boldsymbol{\mu}) \in \boldsymbol{\Pi} \times \boldsymbol{\mathcal{M}}$ such that $\boldsymbol{\pi} \in \Phi(\boldsymbol{\mu})$ and $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$.*

3.2.2 *Theoretical Foundations*

To obtain meaningful optimality results beyond empirical MF convergence, we will need a Lipschitz assumption as in the uncontrolled, continuous-time case (cf. [150], Condition 2.3) and typical in MF theory [23].

Assumption 3.2.2. *Let r, p, W be Lipschitz continuous with Lipschitz constants $L_r, L_p, L_W > 0$.*

Note that all proofs but Theorem 3.2.1 also hold for only block-wise Lipschitz continuous W, see Appendix B.1. Since $\mathcal{X} \times \mathcal{U} \times B_1(\mathcal{X})$ is compact, r is bounded by the extreme value theorem.

Proposition 3.2.1. *Under Assumption 3.2.2, r will be bounded by $|r| \leq M_r$ for some constant $M_r > 0$.*

We then obtain existence of a GMFE by reformulating the GMFG as a classical MFG and applying existing results from [24]. More precisely, we consider the equivalent MFG with extended state space $\mathcal{X} \times \mathcal{I}$, action space $\mathcal{U}$, policy $\tilde{\pi} \in \mathcal{P}(\mathcal{U})^{\mathcal{T}\times\mathcal{X}\times\mathcal{I}}$, MF $\tilde{\mu} \in \mathcal{P}(\mathcal{X} \times \mathcal{I})^{\mathcal{T}}$, reward function $\tilde{r}((x, \alpha), u, \tilde{\mu}) := r(x, u, \int_{\mathcal{I}} W(\alpha_t, \beta)\tilde{\mu}_t(\cdot, \beta)\,\mathrm{d}\beta)$ and transition dynamics such that the states $(\tilde{x}_t, \alpha_t)$ follow $(\tilde{x}_0, \alpha_0) \sim \tilde{\mu}_0 := \mu_0 \otimes \mathrm{Unif}([0, 1])$ and

$$\tilde{u}_t \sim \tilde{\pi}_t(\tilde{u}_t \mid \tilde{x}_t, \alpha_t), \quad \tilde{x}_{t+1} \sim p(\tilde{x}_{t+1} \mid \tilde{x}_t, \tilde{u}_t, \int_{\mathcal{I}} W(\alpha_t, \beta)\tilde{\mu}_t(\cdot, \beta)\,\mathrm{d}\beta), \quad \alpha_{t+1} = \alpha_t\,. \tag{3.2.14}$$

Theorem 3.2.1. *Under Assumption 3.2.2, there exists a GMFE* $(\boldsymbol{\pi}, \boldsymbol{\mu}) \in \boldsymbol{\Pi} \times \boldsymbol{\mathcal{M}}$.

Meanwhile, in finite games, even showing the existence of Nash equilibria in local feedback policies is problematic [24]. Note however, that while this reformulation will be useful for learning and existence, it does not allow us to conclude that the **finite graph** game is well approximated, as classical MFG approximation theorems e.g. in [24] do not consider the graph structure and directly use the limiting graphon W in the dynamics Eq. (3.2.14).

As our next main result, we shall therefore show rigorously that the GMFE can provide increasingly good approximations of the N-agent finite graph game as $N \to \infty$. As mentioned, the following also holds for only block-wise Lipschitz continuous W instead of fully Lipschitz continuous W. Complete mathematical proofs together with additional theoretical supplements can be found in Appendix B. To obtain joint N-agent policies as approximate Nash equilibria from a GMFE $(\boldsymbol{\pi}, \boldsymbol{\mu})$, we define the map $\Gamma_N(\boldsymbol{\pi}) \coloneqq (\pi^1, \pi^2, \dots, \pi^N) \in \Pi^N$, where

$$\pi_t^i(u \mid x) \coloneqq \pi_t^{\alpha_i}(u \mid x), \quad \forall (\alpha, t, x, u) \in \mathcal{I} \times \mathcal{T} \times \mathcal{X} \times \mathcal{U} \tag{3.2.15}$$

with $\alpha_i = \frac{i}{N}$, as by Assumption 3.2.1 the agents are correctly labeled such that they match up with their limiting graphon indices $\alpha_i \in \mathcal{I}$. In our experiments, we use the α_i generated during the generation process of the W-random graphs, though for arbitrary finite systems one would have to first identify the graphon as well as an appropriate assignment of agents to graphon indices $\alpha_i \in \mathcal{I}$, which is a separate, non-trivial problem requiring at least graphon estimation, e.g. [160].

For theoretical analysis, we propose to lift the empirical distributions and policy tuples to the continuous domain $\mathcal{I}$, i.e. under an N-agent policy tuple $(\pi^1, \dots, \pi^N) \in \Pi^N$, we define the step policy ensemble $\boldsymbol{\pi}^N \in \boldsymbol{\Pi}$ and the random empirical step measure ensemble $\boldsymbol{\mu}^N \in \boldsymbol{\mathcal{M}}$ by

$$\pi_t^{N,\alpha} \coloneqq \sum_{i \in \mathcal{V}_N} \mathbf{1}_{\alpha \in (\frac{i-1}{N}, \frac{i}{N}]} \cdot \pi_t^i, \quad \mu_t^{N,\alpha} \coloneqq \sum_{i \in \mathcal{V}_N} \mathbf{1}_{\alpha \in (\frac{i-1}{N}, \frac{i}{N}]} \cdot \delta_{x_t^j}, \quad \forall (\alpha, t) \in \mathcal{I} \times \mathcal{T}\,. \tag{3.2.16}$$

In the following, we consider deviations of the i-th agent from $(\pi^1, \pi^2, \dots, \pi^N) = \Gamma_N(\boldsymbol{\pi}) \in \Pi^N$ to $(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) \in \Pi^N$, i.e. the i-th agent deviates by instead applying $\hat{\pi} \in \Pi$. Note that this includes the special case of no agent deviations. For any $f \colon \mathcal{X} \times \mathcal{I} \to \mathbb{R}$ and state marginal ensemble $\boldsymbol{\mu}_t \in \mathcal{M}_t$, define $\boldsymbol{\mu}_t(f) \coloneqq \int_{\mathcal{I}} \sum_{x \in \mathcal{X}} f(x, \alpha) \mu_t^\alpha(x) \, \mathrm{d}\alpha$. We are now ready to state our first result of convergence of empirical state distributions to the MF, potentially at the classical rate $\mathcal{O}(1/\sqrt{N})$ and consistent with results in uncontrolled, continuous-time diffusive graphon MF systems (cf. [150], Theorem 3.2).

Theorem 3.2.2. *Consider Lipschitz continuous* $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ *up to a finite number of discontinuities* D_π*, with associated MF ensemble* $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$*. Under Assumption 3.2.1 and the* N*-agent policy* $(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) \in \Pi^N$ *with* $(\pi^1, \pi^2, \dots, \pi^N) = \Gamma_N(\boldsymbol{\pi}) \in \Pi^N$*,* $\hat{\pi} \in \Pi$*,* $t \in \mathcal{T}$*, we have for all measurable functions* $f \colon \mathcal{X} \times \mathcal{I} \to \mathbb{R}$ *uniformly bounded by some* $M_f > 0$*, that*

$$\mathbb{E}\left[\left|\boldsymbol{\mu}_t^N(f) - \boldsymbol{\mu}_t(f)\right|\right] \to 0 \tag{3.2.17}$$

uniformly over all possible deviations $\hat{\pi} \in \Pi, i \in \mathcal{V}_N$*. Furthermore, if the graphon convergence in Assumption 3.2.1 is at rate* $\mathcal{O}(1/\sqrt{N})$*, then this rate of convergence is also* $\mathcal{O}(1/\sqrt{N})$.

In particular, the technical Lipschitz requirement of $\boldsymbol{\pi}$ typically holds for neural-network-based policies [105, 116] and includes also the case of finitely many optimality regimes over all graphon indices $\alpha \in \mathcal{I}$, which is sufficient to achieve arbitrarily good approximate Nash equilibria through

our algorithms as shown in Section 3.2.3. We would like to remark that the above result generalizes convergence of state histograms to the MF solution, since the state marginals of agents are additionally close to each of their graphon MF equivalents. The above will be necessary to show convergence of the dynamics of a deviating agent to

$$\hat{x}_0^{\overset{i}{N}} \sim \mu_0, \quad \hat{u}_t^{\overset{i}{N}} \sim \hat{\pi}_t(\hat{u}_t^{\overset{i}{N}} \mid \hat{x}_t^{\overset{i}{N}}), \quad \hat{x}_{t+1}^{\overset{i}{N}} \sim p(\hat{x}_{t+1}^{\overset{i}{N}} \mid \hat{x}_t^{\overset{i}{N}}, \hat{u}_t^{\overset{i}{N}}, \mathbb{G}_t^{\overset{i}{N}}), \quad \forall t \in \mathcal{T} \tag{3.2.18}$$

for almost all agents i, i.e. the dynamics are approximated by using the limiting deterministic neighborhood MF $\mathbb{G}^{\overset{i}{N}}$, see Appendix B.1. This will imply the approximate Nash property:

Theorem 3.2.3. *Consider a GMFE $(\boldsymbol{\pi}, \boldsymbol{\mu})$ with Lipschitz continuous $\boldsymbol{\pi}$ up to a finite number of discontinuities D_π. Under Assumptions 3.2.1 and 3.2.2, for any $\varepsilon, \delta > 0$ there exists N' such that for all $N > N'$, the policy $(\pi^1, \dots, \pi^N) = \Gamma_N(\boldsymbol{\pi}) \in \Pi^N$ is an (ε, δ)-Markov Nash equilibrium, i.e.*

$$J_i^N(\pi^1, \dots, \pi^N) \geq \max_{\pi \in \Pi} J_i^N(\pi^1, \dots, \pi^{i-1}, \pi, \pi^{i+1}, \dots, \pi^N) - \varepsilon \tag{3.2.19}$$

for all $i \in \mathcal{J}_N$ and some $\mathcal{J}_N \subseteq \mathcal{V}_N \colon |\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$.

In general, Nash equilibria are highly intractable [45]. Therefore, solving the GMFG allows obtaining approximate Nash equilibria in the N-agent system for sufficiently large N, since $\varepsilon, p \to 0$ as $N \to \infty$. As a side result, we also obtain first results for the uncontrolled discrete-time case by considering trivial action spaces with $|\mathcal{U}| = 1$, see Corollary B.1.2 in the Appendix.

3.2.3 *Learning Graphon Mean Field Equilibria*

By learning GMFE, one may potentially solve otherwise intractable large N-agent games. For learning, we can apply any existing techniques for classical MFGs (e.g. [95, 114, 115]), since by Eq. (3.2.14) we have reformulated the GMFG as a classical MFG with extended state space. Nonetheless, it may make sense to treat the graphon index $\alpha \in \mathcal{I}$ separately, e.g. when treating special cases such as block graphons, or by grouping graphically similar agents. We repeatedly apply two functions $\hat{\Phi}, \hat{\Psi}$ by beginning with the MF $\boldsymbol{\mu}^0 = \hat{\Psi}(\boldsymbol{\pi}^0)$ generated by the uniformly random policy $\boldsymbol{\pi}^0$, and computing $\boldsymbol{\pi}^{n+1} = \hat{\Phi}(\boldsymbol{\mu}^n)$, $\boldsymbol{\mu}^{n+1} = \hat{\Psi}(\boldsymbol{\pi}^{n+1})$ for $n = 0, 1, \dots$ until convergence using one of the following two approaches:

- **Equivalence classes method.** We introduce agent equivalence classes, or discretization, of $\mathcal{I}$ for the otherwise uncountably many agents $\alpha \in \mathcal{I}$ by partitioning $\mathcal{I}$ into M subsets. For example, in the special case of block graphons (block-wise constant W), one can solve separately for each block equivalence class (type) of agents, since all agents in the class share the same dynamics. Note that in contrast to multi-class MFGs [23], GMFGs are rigorously connected to finite graph games and can handle an uncountable number of classes α. To deal with general graphons, we choose equidistant representatives $\alpha_i \in \mathcal{I}$, $i = 1, \dots, M$ covering the whole interval $\mathcal{I}$, and approximate each agent $\alpha \in \tilde{\mathcal{I}}_i$ by the nearest α_i for the intervals $\tilde{\mathcal{I}}_i \subseteq \mathcal{I}$ of points closest to that α_i to obtain M approximate equivalence classes. Formally, we approximate MFs $\hat{\Psi}(\boldsymbol{\pi}) = \sum_{i=1}^M \mathbf{1}_{\alpha \in \tilde{\mathcal{I}}_i} \hat{\mu}^{\alpha_i}$ recursively computed over all times for any fixed policy ensemble $\boldsymbol{\pi}$, and similarly policies $\hat{\Phi}(\boldsymbol{\mu}) = \sum_{i=1}^M \mathbf{1}_{\alpha \in \tilde{\mathcal{I}}_i} \pi^{\alpha_i}$ where π^{α_i} is the optimal policy of α_i for fixed $\boldsymbol{\mu}$. We solve the optimal control problem for each equivalence class using backwards induction (alternatively, one may use RL), and solve the evolution

equation for the representatives α_i of the equivalence classes recursively. The details are found in Appendix B.11. Note that this does not mean that we consider the N-agent problem with $N = M$, but instead we approximate the limiting problem with the limiting graphon W, and the solution will be near-optimal for all sufficiently large finite systems at once.

- **Direct RL.** We directly apply RL as $\hat{\Phi}$. The central idea is to consider the GMFG as a classical MFG with extended state space $\mathcal{X} \times \mathcal{I}$, i.e. for fixed MFs, we solve the MDP defined by Eq. (3.2.14). Agents condition their policy not only on their own state, but also their node index $\alpha \in \mathcal{I}$ and the current time $t \in \mathcal{T}$, since the MFs are non-stationary in general and require time-dependent policies for optimality. Here, we assume that we can sample from a simulator of Eq. (3.2.9) for a given fixed MF as commonly assumed in MFG learning literature [95, 115]. For application to arbitrary finite systems, one could apply a model-based RL approach coupled with graphon estimation, though this remains outside the scope of this work. For solving the MF evolution equation Eq. (3.2.12), we can again use any applicable numerical method and choose a conventional sequential Monte Carlo method for $\hat{\Psi}$. While it is possible to exactly solve optimal control problems for each agent equivalence class with finite state-action spaces, this is generally not the case for e.g. continuous state-action spaces. Here, a general RL solution can solve otherwise intractable problems in an elegant manner, since the graphon index α simply becomes part of a continuous state space.

For convergence, we begin by stating the classical feedback regularity condition [23, 95] after equipping $\boldsymbol{\Pi}$, $\boldsymbol{\mathcal{M}}$ e.g. with the supremum metric.

Proposition 3.2.2. *Assume that the maps $\hat{\Psi}, \hat{\Phi}$ are Lipschitz with constants c_1, c_2 and $c_1 \cdot c_2 < 1$. Then the FPI $\boldsymbol{\mu}^{n+1} = \hat{\Psi}(\hat{\Phi}(\boldsymbol{\mu}^n))$ converges.*

Feedback regularity is not assured, and thus there is no general convergence guarantee. Whilst one could attempt to apply FP [114], additional assumptions will be needed for convergence. Instead, whenever necessary for convergence, we regularize by introducing Boltzmann policies $\pi_t^\alpha(u \mid x) \propto \exp(\frac{1}{\eta} Q_\alpha^{\boldsymbol{\mu}}(t, x, u))$ with temperature η, provably converging to an approximation for sufficiently high temperatures [9].

Theorem 3.2.4. *Under Assumptions 3.2.1 and 3.2.2, the equivalence classes algorithm with Boltzmann policies $\hat{\Phi}(\boldsymbol{\mu})_t^\alpha(u \mid x) \propto \exp(\frac{1}{\eta} Q_\alpha^{\boldsymbol{\mu}}(t, x, u))$ converges for sufficiently high temperatures $\eta > 0$.*

Importantly, even an exact solution of the GMFG only constitutes an approximate Nash equilibrium in the finite-graph system. Furthermore, even the existence of exact finite-system Nash equilibria in local feedback policies is not guaranteed, see the discussion in [24] and references therein. Therefore, little is lost by introducing slight additional approximations for the sake of a tractable solution, if at all needed (e.g. the Investment-Graphon problem in the following converges without introducing Boltzmann policies), since near optimality holds for small temperatures [9]. Indeed, we find that we can show optimality of the equivalence classes approach for sufficiently fine partitions of $\mathcal{I}$, giving us a theoretical foundation for our proposed algorithms.

Theorem 3.2.5. *Under Assumptions 3.2.1 and 3.2.2, for a solution $(\boldsymbol{\pi}, \boldsymbol{\mu}) \in \boldsymbol{\Pi} \times \boldsymbol{\mathcal{M}}$, $\boldsymbol{\pi} \in \hat{\Phi}(\boldsymbol{\mu})$, $\boldsymbol{\mu} = \hat{\Psi}(\boldsymbol{\pi})$ of the M equivalence classes method and for any $\varepsilon, \delta > 0$ there exists $N', M \in \mathbb{N}$ such that for all $N > N'$, the policy $(\pi^1, \dots, \pi^N) = \Gamma_N(\boldsymbol{\pi}) \in \Pi^N$ is an (ε, δ)-Markov Nash equilibrium.*

A theoretically rigorous analysis of the elegant direct RL approach is beyond our scope and deferred to future works, though we empirically find that both methods agree.

3.2.4 *Experiments*

In this section, we will give an empirical verification of our theoretical results. As we are unaware of any prior discrete-time GMFGs (except for the example in [154], which is similar to the first problem in the following), we propose two problems adapted from existing non-graph-based works on the three graphons in Figure 3.6. We defer detailed descriptions of problems and algorithms, plots as well as further analysis, including exploitability and a verification of stability of our solution with respect to the number of equivalence classes – to Appendix B.11.

The **SIS-Graphon** problem was considered in [9] as a classical discrete-time MFG. We impose an epidemics scenario where people (agents) are infected with probability proportional to the number of infected neighbors and recover with fixed probability. People may choose to take precautions (e.g. social distancing), avoiding potential costly infection periods at a fixed cost.

In the **Investment-Graphon** problem – an adaptation of a problem studied by [98], where it was in turn adapted from [161] – we consider many firms maximizing profits, where profits are proportional to product quality and decrease with total neighborhood product quality, i.e. the graph models overlap in e.g. product audience or functionality. Firms can invest to improve quality, though it becomes more unlikely to improve quality as their quality rises.

LEARNED EQUILIBRIUM BEHAVIOR. For the SIS-Graphon problem, we apply softmax policies for each approximate equivalence class to achieve convergence, see Appendix B.11 for details on temperature choice and influence. In Figure 3.7, the learned behavior can be observed for various α. As expected, in the ER graphon case, behavior is identical over all α. Otherwise, we find that agents take more precautions with many connections (low α) than with few connections (high α). For the uniform attachment graphon, we observe no precautions in case of negligible connectivity ($\alpha \to 1$), while for the ranked attachment graphon there is no such $\alpha \in \mathcal{I}$ (cf. Figure 3.6). Further, the fraction of infected agents at stationarity rises as α falls. A similar analysis holds for Investment-Graphon without need for regularization, see Appendix B.11.

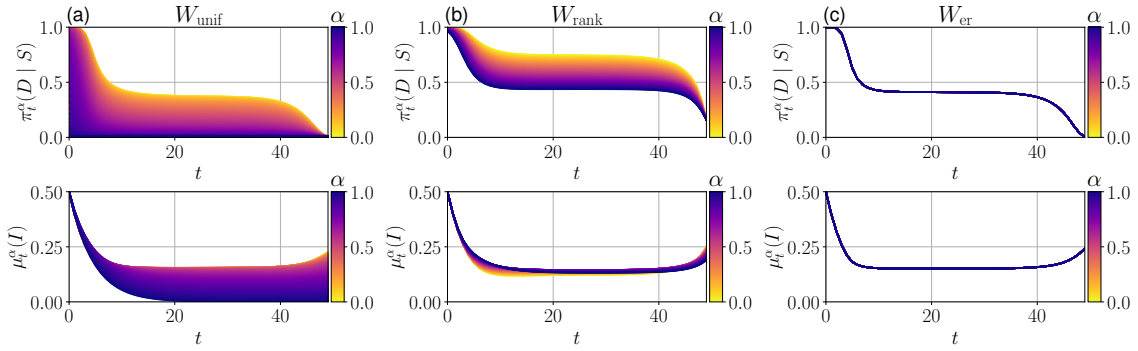


FIGURE 3.7: Achieved equilibrium via $M = 100$ approximate equivalence classes in SIS-Graphon, plotted for each agent $\alpha \in \mathcal{I}$. Top: Probability of taking precautions when healthy. Bottom: Probability of being infected. It can be observed that agents with less connections (higher α) will take less precautions. **(a)**: Uniform attachment graphon; **(b)**: Ranked attachment graphon; **(c)**: ER graphon.

Note that the specific method of solution is not of central importance here, as in general any RL and filtering method can be substituted to handle 1. otherwise intractable or 2. inherently sample-based settings. Indeed, we achieve similar results using PPO [73] in Investment-Graphon, enabling a general RL-based methodology for GMFGs. In Appendix B.11, we find that PPO achieves qualitatively and quantitatively similar behavior to the equivalence classes method, with slight deviations due to the approximations from PPO and Monte Carlo. In particular, the PPO exploitability $\varepsilon \approx 2$ remains low compared to $\varepsilon > 30$ for the uniform random policy, see Appendix B.11. In Appendix B.11, we also show how, due to the non-stationarity of the environment, a naive application of MARL [88] fails to converge, while existing MF MARL techniques [120] remain incomparable as agents must observe the average actions of all neighbors. On SIS-Graphon, we require softmax policies to achieve convergence, which is not possible with PPO as no Q-function is learned. In general, one could use entropy regularized policies, e.g. SAC [162], or alternatively use any value-based RL method, though an investigation of the best approach is outside of our scope.

QUANTITATIVE VERIFICATION OF THE MF APPROXIMATION. To verify the rigorously established accuracy of our MF system empirically, we will generate W-random graphs. Note that there are considerable difficulties associated with an empirical verification of Eq. (3.2.19), since 1. for any N one must check the Nash property for (almost) all N agents, 2. finding optimal $\hat{\pi}$ is intractable, as no Dynamic Programming Principle (DPP) holds on the non-Markovian local agent state, while acting on the full state fails by the curse of dimensionality, and 3. the inaccuracy from estimating all J_i^N, $i = 1, \dots, N$ at once increases with N due to variance, i.e. cost scales fast with N for fixed variance. Instead, we verify Eq. (B.1.7) in Appendix B.1 using the GMFE policy on systems of up to $N = 100$ agents, i.e. $\hat{\pi} = \pi^{\alpha_i}$ for the closest α_i and comparing for all agents at once ($p = 0$). Shown in Figure 3.8, for W-random graph sequences, at each N we performed 10000 runs to estimate $\max_i |J_i^N - J_{\alpha_i}|$. We find that the maximum deviation between achieved returns and MF return decreases as $N \to \infty$, verifying that we obtain an increasingly good approximation of the finite N-agent graph system. The oscillations in Figure 3.8 stem from the randomly sampled graphs.

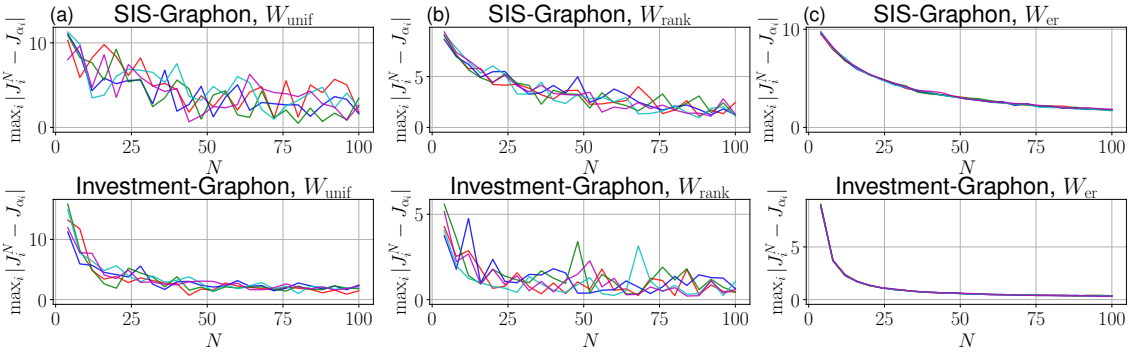


FIGURE 3.8: Decreasing maximum deviation between average N-agent objective and MF objective over all agents for the GMFE policy and 5 W-random graph sequences. **(a)**: Uniform attachment graphon; **(b)**: Ranked attachment graphon; **(c)**: ER graphon.

3.2.5 *Summary*

In this work, we have formulated a new framework for dense graph-based dynamical games with the weak interaction property. On the theoretical side, we have given one of the first general

discrete-time GMFG formulations with existence conditions and approximate Nash property of the finite graph system, thus extending classical MFGs and allowing for a tractable, theoretically well-founded solution of competitive large-scale graph-based games on large dense graphs. On the practical side, we have proposed a number of computational methods to tractably compute GMFE and experimentally verified the plausibility of our methodology on a number of examples. Venues for further extensions are manifold and include extensions of theory to e.g. continuous spaces, partial observability or common noise. So far, graphons assume dense graphs and cannot properly describe sparse graphs ($W = 0$), which remain an active frontier of research. Finally, real-world application scenarios may be of interest, where estimation of agent graphon indices becomes important for model-based MARL. We hope that our work inspires further applications and research into scalable MARL using graphical dynamical systems based on graph limit theory and MF theory, such as our extensions to more sparse and weighted or directed graphs [13, 14, 19].

3.3 MEAN FIELD GAMES ON HYPERGRAPHS

We propose an approach to modelling large-scale multi-agent dynamical systems allowing interactions among more than just pairs of agents using the theory of MFGs and the notion of hypergraphons, which are obtained as limits of large hypergraphs. To the best of our knowledge, ours is the first work on MFGs on hypergraphs. Together with an extension to a multi-layer setup, we obtain limiting descriptions for large systems of non-linear, weakly-interacting dynamical agents. On the theoretical side, we prove the well-foundedness of the resulting hypergraphon MFG, showing both existence and approximate Nash properties. On the applied side, we extend numerical and learning algorithms to compute the hypergraphon MFE. To verify our approach empirically, we consider a social rumor spreading model, where we give agents intrinsic motivation to spread rumors to unaware agents, and an epidemics control problem. The material presented in this section is based on [17].

Tractably finding competitive equilibria and decentralized, cooperative optimal control solutions has been the focus of many recent works [7, 115, 117, 118, 130, 147, 154]. MF systems have also been extended to dynamical systems on graphs, typically using the theory of large graph limits called graphons [157, 163], as seen in Section 3.2. The graphon MF systems can be considered either as the limit of systems with weakly-interacting node state processes [7, 150], or alternatively as the result of a double limit procedure where each node constitutes a large population, or 'cluster' of agents, each of which interacts with each other via inter- and intra-cluster coupling. First, infinitely many nodes are considered according to the graphon, and then infinitely many agents are considered per node, see e.g. [139, 164].

In this section, we will consider the former. The goal of our work is the synthesis of dynamical systems on hypergraphs with competitive or selfish agents. Existing analysis of hypergraph MF systems typically remains restricted to special dynamics such as epidemiological equations [165–167] or opinion dynamics [168] on sparse graphs. In contrast, our work deals with general, agent-controlled non-linear dynamics and equilibrium solutions. We build upon prior results for discrete-time, graph-based MF systems [7, 24, 150] and extend them to incorporate higher-order hypergraphs as well as multiple layers.

OUR CONTRIBUTION. Our contribution can be summarized as follows: (i) To the best of our knowledge, ours is the first general MFG-theoretical framework for non-linear dynamics on multi-layer hypergraphs. Multi-layer networks [169] have proven extremely useful in many application areas including infectious disease epidemiology, where different layers could be used to describe community, household and hospital settings [170]. (ii) We prove the existence and the approximation properties of the proposed MFE. (iii) We propose and empirically verify algorithms for solving such hypergraphon MF systems, and thereby obtain a tractable approach to solving and analyzing otherwise intractable Nash equilibria on multi-layer hypergraph games. The proposed framework is of great generality, extending the recently established graphon MFGs and thereby also standard MFGs (via fully-connected graphs).

After introducing some graph-theoretical preliminaries, in Section 3.3.1 we will begin by formulating the motivating mathematical dynamical model and game on hypergraphs, as well as its more tractable MF analogue. Then, in Section 3.3.2 we will show the existence of solutions for the MF problem as well as quantify its approximation qualities of the finite hypergraph game, building a mathematical foundation for hypergraphon MFGs. Lastly, in Section 3.3.3 we will evaluate our model numerically for an illustrative rumor spreading game, verifying our theoretical approximation results and the obtained equilibrium behavior. All of the proofs can be found in the Appendix.

Notation. On a discrete space A, define the spaces of all (Borel) probability measures $\mathcal{P}(A)$ and all sub-probability measures $\mathcal{B}_1(A)$, equipped with the L_1 norm. Define the unit interval $\mathcal{I} \coloneqq [0,1]$ and its N equal-length subintervals $I_1^N, \dots, I_N^N$ such that $\bigsqcup_{i=1}^N I_i^N = \mathcal{I}$ for any integer N, where $\bigsqcup$ denotes disjoint union and each I_i^N includes its rightmost point i/N. Denote the expectation and variance of random variables X by $\mathbb{E}[X]$, $\mathbb{V}[X]$. Define the indicator function $\mathbf{1}_A(x)$ mapping to 1 whenever $x \in A$ and 0 otherwise. For any integer k, define $[k] \coloneqq \{1, \dots, k\}$. Let $r(A,m)$ denote the set of all distinct non-empty subsets of any set A with at most m elements, and denote the set of all distinct non-empty, proper subsets by $r_<(A) \coloneqq r(A, |A|-1)$ as well as the set of all distinct non-empty subsets by $r(A) \coloneqq r(A,|A|)$. To keep notation simple, in the following we write $r_<[k] \coloneqq r_<([k])$, $r[k] \coloneqq r([k])$ and identify e.g. $r_<[k]$ with $[|r_<[k]|] \coloneqq \{1, \dots, |r_<[k]|\}$ whenever helpful. Denote the set of permutations of a set A as $\mathrm{Sym}(A)$. Define the space of bounded, $r_<[k]$-dimensional, symmetric functions $\mathrm{Sym}^{\mathrm{ind}}_<[k]$ induced by permutations of the underlying set $[k]$, i.e. any bounded function $f \colon \mathcal{I}^{r_<[k]} \to \mathbb{R}$ is in $\mathrm{Sym}^{\mathrm{ind}}_<[k]$ whenever f is invariant to all permutations $\sigma \in \mathrm{Sym}([k])$, $f(x_1, \dots, x_k, x_{11}, x_{12}, \dots) = f(x_{\sigma(1)}, \dots, x_{\sigma(k)}, x_{\sigma(1)\sigma(1)}, x_{\sigma(1)\sigma(2)}, \dots)$. Analogously, we define spaces of such functions $\mathrm{Sym}^{\mathrm{ind}}_\leq[k]$ and $\mathrm{Sym}^{\mathrm{ind}}[k]$ over $r[k]$ and $[k]$, respectively.

3.3.1 *Mean Field Games on Dense Hypergraphs*

Before we formulate the stochastic dynamic hypergraph game and its limiting analogue in the following subsections, we discuss some graph-theoretical preliminaries. A (undirected) hypergraph is defined as a pair $H = (V, E)$ of a set of vertices V and a set of hyperedges $E \subseteq 2^V \setminus \{\varnothing\}$. In contrast to edges in graphs, here hyperedges may connect an arbitrary number of vertices instead of only two. If there is no scope of confusion, we will call hyperedges of a hypergraph just edges. Denote by $V[H]$ and $E[H]$ the vertex set and edge set of a hypergraph H. The maximum cardinality of all edges of a hypergraph H is called its rank. A k-uniform hypergraph is defined as a hypergraph where all edges have cardinality k. A multi-layer hypergraph $H = (V, E^1, \dots, E^D)$ with D layers is obtained by allowing for multiple edge sets $E^1, \dots, E^D \subseteq 2^V \setminus \{\varnothing\}$, and we analogously write $E^d[H]$ for the d-th set of edges of a multi-layer hypergraph H. We define the d-th sub-hypergraph H^d of a multi-layer hypergraph H as the hypergraph with vertex set $V[H]$ and edge set $E[H^d] = E^d[H]$.

Consider any (non-uniform) hypergraph H with bounded rank $k_{\max}$. Observe the isomorphism between multi-layer uniform hypergraphs and such H by splitting hyperedges of each cardinality $k \leq k_{\max}$ into their own layer. Since this procedure can be repeated for each layer of a multi-layer hypergraph, any multi-layer hypergraph is therefore equivalent to a correspondingly defined multi-layer uniform hypergraph. Hence, from here on it suffices to define and consider $[k_1, \dots, k_D]$-uniform hypergraphs H as D-layer hypergraphs, where each layer $d = 1, \dots, D$ is given by a k_d-uniform hypergraph with $k_d \leq k_{\max}$, see also Figure 3.9 for a visualization. For instance, in social networks each layer could model e.g. the k-cliques of acquaintances formed at work, friendship at school or family relations.

To formulate the infinitely-large MF system, we define the limiting description of sufficiently dense multilayer hypergraphs as the graphs intuitively become infinite in size, called hypergraphons [171]. Here, dense means a number of edges on the order of $O(N^2)$, where N is the number of vertices, to which existing hypergraphon theory remains limited to. However, we note that an extension to more sparse models by fusing the theory of hypergraphons with L^p graphons [14, 172, 173] could be part of future work. The space of k-uniform hypergraphons $\mathcal{W}_k$ is now defined as the space of

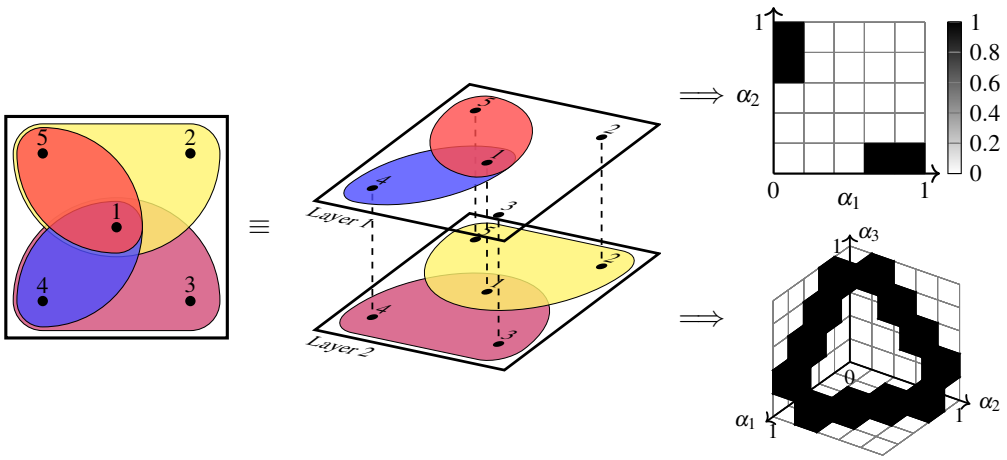


FIGURE 3.9: An example hypergraph H is transformed into a multi-layer uniform hypergraph. On the left, a hypergraph H with nodes $V[H] = \{1, \dots, 5\}$ and hyperedges $E[H] = \{\{1,4\},\{1,5\},\{1,3,4\},\{1,2,5\}\}$ is depicted. An equivalent representation of H as a $[2,3]$-uniform hypergraph H_{unif} as well as its associated hypergraphons are given, where the first and second layers each consist of edges $\{\{1,4\},\{1,5\}\}$ and 3-hyperedges $\{\{1,3,4\},\{1,2,5\}\}$ respectively. The associated (step-)hypergraphons $W[H^1_{\text{unif}}]$ and $W[H^2_{\text{unif}}]$ are given as continuous versions of the (multi-dimensional) $\{0,1\}$-valued adjacency matrices. Here, we depict only the first three coordinates for the second layer step-hypergraphon $W[H^2_{\text{unif}}]$, given by the constant 1 (black) or 0 (white). Note that while each edge corresponds to two entries in the adjacency matrix of the 2-uniform case, for the 3-uniform case each hyperedge corresponds to six entries, resulting in the step graphon shown (bottom right).

all bounded and symmetric functions $W \in \mathrm{Sym}^{\mathrm{ind}}_{<}[k], W\colon \mathcal{I}^{r<[k]} \to \mathcal{I}$ that are measurable. We equip $\mathrm{Sym}^{\mathrm{ind}}_{<}[k]$ with the cut (semi-)norm $\|\cdot\|_{\square^{k-1}}$ proposed by [174], defined by

$$\|W\|_{\square^{k-1}} := \sup_{\substack{u_i\colon \mathcal{I}^{r[k-1]} \to \mathcal{I}, \\ u_i \in \mathrm{Sym}^{\mathrm{ind}}_{\leq}[k-1]}} \left| \int_{\mathcal{I}^{r<[k]}} W(\alpha) \prod_{i=1}^{k} u_i(\alpha_{r([k]\setminus\{i\})}) \,\mathrm{d}\alpha \right|, \tag{3.3.20}$$

which (see e.g. [157, Lemma 8.10]) coincides with the standard graphon case for $k = 2$,

$$\|W\|_{\square} = \sup_{f,g\colon \mathcal{I}\to\mathcal{I}} \left| \int_{\mathcal{I}^2} W(\alpha,\beta) f(\alpha) g(\beta) \,\mathrm{d}(\alpha,\beta) \right|. \tag{3.3.21}$$

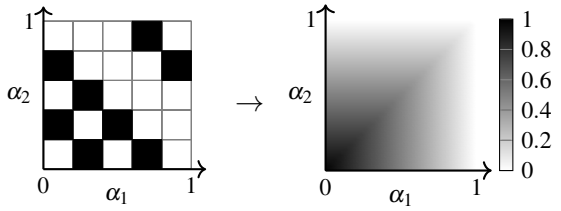


FIGURE 3.10: Visualization of the convergence of 2-dimensional step-graphons to the uniform attachment graphon $W_{\text{unif}}(\alpha_1, \alpha_2) = 1 - \max(\alpha_1, \alpha_2)$.

To analytically connect k-uniform hypergraphs to hypergraphons, we define the step-hypergraphons of any k-uniform hypergraph H as

$$W[H](\boldsymbol{\alpha}) = \sum_{\mathbf{m}\in[N]^k} \mathbf{1}_{E[H]}(\mathbf{m}) \cdot \prod_{i\in[k]} \mathbf{1}_{I^N_{m_i}}(\alpha_i). \tag{3.3.22}$$

For motivation, note that for any sequence of graphs with converging homomorphism densities, equivalently the step graphons converge in the cut norm to the limiting graphon, and their limiting homomorphism densities can be described by the limiting graphon [157]. Similarly, cut-norm convergence for the more general uniform hypergraphs at least implies the convergence of hypergraph homomorphism densities [174]. Accordingly, we assume hypergraph convergence in each layer of a given sequence of $[k_1, \dots, k_D]$-uniform hypergraphs $(H_N)_{N\in\mathbb{N}}$ via convergence of their step-hypergraphons $W_d^N := W[H_N^d]$ to a limiting hypergraphon $W_d \in \mathcal{W}_{k_d}$ in the cut norm as visualized in Figure 3.10, similar as in standard graphon MF systems [7, 150].

Assumption 3.3.1. *The sequence of step-hypergraphons $W^N := (W_d^N)_{d\in[D]}$ converges on each layer in cut norm $\|\cdot\|_\square$ to some hypergraphons $W := (W_d)_{d\in[D]} \in \bigtimes_{d\in[D]} \mathcal{W}_{k_d}$, i.e.*

$$\left\|W_d^N - W_d\right\|_\square \to 0, \quad \forall d \in [D]. \tag{3.3.23}$$

3.3.1.1 *Finite Hypergraph Game*

In this subsection, we will formulate a dynamical model on hypergraphs where each node is understood as an agent that is influenced by the state distribution of all of its neighbors, according to some time-varying dynamics. Furthermore, each agent is expected to selfishly optimize its own objective, which gives rise to Nash equilibria as the solution of interest.

Consider a $[k_1, \dots, k_D]$-uniform hypergraph and let $\mathcal{T}$ be the time index set, either $\mathcal{T} = \{0, 1, \dots, T-1\}$ or $\mathcal{T} = \mathbb{N}_0 := \{0, 1, 2, \dots\}$. We define N agents $i \in [N]$ each endowed with local states x_t^i and actions u_t^i from a finite state space $\mathcal{X}$ and finite action space $\mathcal{U}$, respectively. Here, $\mathcal{X}$ and $\mathcal{U}$ are assumed finite for technical reasons, though we believe that results could be extended to more general spaces in the future. States have an initial distribution $x_0^i \sim \mu_0 \in \mathcal{P}(\mathcal{X})$. For all times $t \in \mathcal{T}$ and agents $i \in [N]$, their actions are random variables following the law

$$u_t^i \sim \pi_t^i(u_t^i \mid x_t^i), \tag{3.3.24}$$

with policy (i.e. probability distribution over actions) $\pi^i \in \Pi := \mathcal{P}(\mathcal{U})^{\mathcal{T}\times\mathcal{X}}$, that, for each node i, depends on the i-th state at time t. Then, the states are random variables following the law

$$x_{t+1}^i \sim p(x_{t+1}^i \mid x_t^i, u_t^i, \nu_t^{N,i}), \tag{3.3.25}$$

with transition kernels $p\colon \mathcal{X} \times \mathcal{U} \times \mathcal{B}_1(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ that, for each node i, depends on the i-th state and action at time t, and $\nu_t^{N,i}$. Here, the $\bigtimes_{d=1}^D \mathcal{P}(\mathcal{X}^{k_d-1})$-valued multi-layer empirical neighborhood MF $\nu_t^{N,i}$ is defined as

$$\nu_{t,d}^{N,i} := \frac{1}{N^{k_d-1}} \sum_{\mathbf{m}\in[N]^{k_d-1}} \mathbf{1}_{E^d[H_N]}(\mathbf{m} \cup i)\delta_{\bigtimes_{j\neq i} x_t^{m_j}}, \tag{3.3.26}$$

in its d-th layer, consisting of the unnormalized state distributions of an agent i's neighbors on each layer. In other words, the state dynamics of an agent depend only on the states of nodes in their immediate neighborhood and can be influenced by the agent via its actions u_t^i.

For example, in an epidemics spread scenario, the states of each agent could model their infection status, while the actions of an agent could be to take protective measures. As a result, each agent will randomly become infected with probability depending on how many neighboring agents are infected and whether the agent is taking protective measures.

The cost functions $r\colon \mathcal{X}\times\mathcal{U}\times\mathcal{B}_1(\mathcal{X})\to\mathbb{R}$ with discount factor $\gamma\in(0,1)$ or in the finite horizon case $\gamma\in(0,1]$ define the objective function for the i-th agent

$$J_i^N(\pi^1,\dots,\pi^N) \coloneqq \mathbb{E}\left[\sum_{t\in\mathcal{T}}\gamma^t r(x_t^i,u_t^i,\nu_t^{N,i})\right], \tag{3.3.27}$$

which can describe also e.g. random rewards R_t^i that are conditionally independent given $x_t^i,u_t^i,\nu_t^{N,i}$ by the law of total expectation and taking the conditional expectation, $r(x_t^i,u_t^i,\nu_t^{N,i})\equiv \mathbb{E}\left[r_t^i \,\middle|\, x_t^i,u_t^i,\nu_t^{N,i}\right]$.

Our goal is now to find Nash equilibria, i.e. stable policies where no agent can singlehandedly deviate and improve their own objective. Note that finding Nash equilibria in games such as the above is difficult, since a) even existence of Nash equilibria under the above, decentralized information structure of policies is hard to show, and b) computation of the Nash equilibria fails due to both curse of dimensionality under full observability and general complexity of computing Nash equilibria [45], see also [24] and the discussion therein.

Thus, in the finite game, we are interested in finding the following weaker notion of approximate equilibria [7, 96], where a negligible fraction of agents that remains insignificant to all other agents may remain suboptimal.

Definition 3.3.1. *An (ε,δ)-Nash equilibrium for $\varepsilon,\delta>0$ is defined as a tuple of policies $(\pi^1,\dots,\pi^N)\in\Pi^N$ such that for any $i\in\mathcal{J}^N$, we have*

$$J_i^N(\pi^1,\dots,\pi^N)\geq \sup_{\pi\in\Pi} J_i^N(\pi^1,\dots,\pi^{i-1},\pi,\pi^{i+1},\dots,\pi^N)-\varepsilon \tag{3.3.28}$$

for some set $\mathcal{J}^N\subseteq[N]$ of at least $\lfloor(1-\delta)N\rfloor$ agents.

While it may seem excessive to reduce to approximate optimality limited to a fraction of the agents, it is always possible under Assumption 3.3.1 for a finite number of agents to deviate arbitrarily from the limiting system description. Therefore, under our assumptions it is only possible to obtain an approximate equilibrium solution for almost all agents via the MF formulation. Although we could make stronger assumptions on the mode of convergence for hypergraphons, such a concept of convergence would be difficult to motivate from a graph theoretical perspective. Therefore, we restrict ourselves to the cut-norm convergence [174] and the above solution concept.

3.3.1.2 *Hypergraphon Mean Field Game*

Next, we will formally let $N\to\infty$ and obtain a more tractable, reduced model consisting of any single representative agent and the distribution of agent states, the so-called MF.

To analyze the case $N\to\infty$ however, we first introduce some preliminary definitions. We define the space of MFs $\mathcal{M}\subseteq\mathcal{P}(\mathcal{X})^{\mathcal{T}\times\mathcal{I}}$ such that $\boldsymbol{\mu}\in\mathcal{M}$ whenever $\alpha\mapsto\mu_t^\alpha(x)$ is measurable for all $t\in\mathcal{T},x\in\mathcal{X}$. Intuitively, a MF is the distribution of states each of the infinitely many agents in $\mathcal{I}$ is in. Analogously, the space of policies $\boldsymbol{\Pi}\subseteq\Pi^{\mathcal{I}}$ is given by policies $\pi\in\Pi^{\mathcal{I}}$ where $\alpha\mapsto\pi_t^\alpha(u\mid x)$ is measurable for any $t\in\mathcal{T},x\in\mathcal{X},u\in\mathcal{U}$. Intuitively, $\pi\in\Pi^{\mathcal{I}}$ defines the behavior for each agent $\alpha\in\mathcal{I}$. For any $f\colon\mathcal{X}\times\mathcal{I}\to\mathbb{R}$ and state marginal ensemble $\boldsymbol{\mu}\in\mathcal{P}(\mathcal{X})^{\mathcal{I}}$, define

$$\boldsymbol{\mu}(f)\coloneqq\int_{\mathcal{I}}\sum_{x\in\mathcal{X}} f(x,\alpha)\mu^\alpha(x)\,\mathrm{d}\alpha. \tag{3.3.29}$$

In the limit of $N \to \infty$, assuming that all agents follow a policy $\mathbf{\Pi} \subseteq \Pi^{\mathcal{I}}$, we obtain infinitely many agents $\alpha \in \mathcal{I}$, for each of whom we define the limiting hypergraphon MF dynamics analogously to the finite hypergraph game.

The agent states have the initial distribution $x_0^\alpha \sim \mu_0 \in \mathcal{P}(\mathcal{X})$. For all times $t \in \mathcal{T}$ and agents $\alpha \in \mathcal{I}$, their actions will be random variables following the law

$$u_t^\alpha \sim \pi_t^\alpha(u_t^\alpha \mid x_t^\alpha), \tag{3.3.30}$$

under the policy $\pi^\alpha \in \Pi$, while their states follow the law

$$x_{t+1}^\alpha \sim p(x_{t+1}^\alpha \mid x_t^\alpha, u_t^\alpha, \nu_t^\alpha), \tag{3.3.31}$$

with the limiting, now deterministic neighborhood MF $\nu_t^\alpha \in \bigtimes_{d=1}^D \mathcal{P}(\mathcal{X}^{k_d-1})$. Informally, by a LLN, we have replaced the distribution of finitely many neighbor states by the limiting MF distribution ν_t^α. The d-th component of this MF is given by

$$\nu_{t,d}^\alpha(x) \coloneqq \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} W_d(\alpha, \boldsymbol{\beta}) \prod_{j=1}^{k_d-1} \mu_t^{\beta_j}(x_j)\,\mathrm{d}\boldsymbol{\beta}, \tag{3.3.32}$$

where x_j denotes separate coordinates of the input x (the order does not matter due to symmetry). In other words, the d-layer neighborhood MF distributions give the probability of random neighbors of a shared hyperedge on layer d to be in states $(x_1, \ldots, x_{k_d-1}) \in \mathcal{X}^{k_d-1}$

Note that the same, shared $\alpha \in \mathcal{I}$ is used for all D layers, i.e. all layer neighborhood distributions of agents jointly converge to the limiting descriptions ν_t^α. This makes sense, since by Assumption 3.3.1, we assume that the agents are already ordered such that the corresponding step-hypergraphons converge to the limiting hypergraphon in cut norm on all layers jointly.

Finally, the objective will be given by

$$J_\alpha^{\boldsymbol{\mu}}(\pi^\alpha) \coloneqq \mathbb{E}\left[\sum_{t\in\mathcal{T}} \gamma^t r(x_t^\alpha, u_t^\alpha, \nu_t^\alpha)\right], \tag{3.3.33}$$

which leads to the MF counterpart of Nash equilibria. Informally, a MFE is given by a 'consistent' tuple of policy and MF, such that the policy is optimal under the MF and the MF is generated by the policy. As a result, if all agents follow the policy, they will be optimal under the generated MF, leading to a Nash equilibrium.

More formally, we define the maps $\Phi\colon \mathcal{M} \to 2^{\mathbf{\Pi}}$ mapping from fixed MF $\boldsymbol{\mu} \in \mathcal{M}$ to all optimal policies $\boldsymbol{\pi} \in \mathbf{\Pi}\colon \forall \alpha \in \mathcal{I}\colon \pi^\alpha \in \arg\max_{\tilde{\pi}} J_\alpha^{\boldsymbol{\mu}}(\tilde{\pi})$ and similarly $\Psi\colon \mathbf{\Pi} \to \mathcal{M}$ mapping from policy $\boldsymbol{\pi} \in \mathbf{\Pi}$ to its induced MF $\boldsymbol{\mu} \in \mathcal{M}$ such that for all $\alpha \in \mathcal{I}$, $t \in \mathcal{T}$ we have the initial distribution $\mu_0^\alpha = \mu_0$ and MF evolution

$$\mu_{t+1}^\alpha = \int_{\mathcal{X}} \int_{\mathcal{U}} p(x, u, \nu_t^\alpha) \pi_t^\alpha(\mathrm{d}u \mid x) \mu_t^\alpha(\mathrm{d}x). \tag{3.3.34}$$

Definition 3.3.2. *A Hypergraphon Mean Field Equilibrium (HMFE) is a pair $(\boldsymbol{\pi}, \boldsymbol{\mu}) \in \mathbf{\Pi} \times \mathcal{M}$ such that $\boldsymbol{\pi} \in \Phi(\boldsymbol{\mu})$ and $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$.*

Importantly, the MFG will be motivated rigorously in the following, and its computational complexity is independent of the number of agents. Instead, the complexity of the problem will scale with the size of agent state and action spaces $\mathcal{X}, \mathcal{U}$ and the considered time horizon in case of a finite horizon cost function, since we will solve for equilibria by repeatedly (i) computing optimal policies for discrete MDPs [68] $\pi^\alpha \in \arg\max_{\tilde{\pi}} J_\alpha^{\boldsymbol{\mu}}(\tilde{\pi})$, and (ii) solving the MF evolution equations Eq. (3.3.34). In particular, MFE are guaranteed to exist, and the corresponding equilibrium policy will provide an equilibrium for large finite systems.

To obtain meaningful results, we need a standard continuity assumption (e.g. [150]), since otherwise weak interaction is not guaranteed: Without continuity, a change of behavior in only one of many agents could otherwise cause arbitrarily large changes in the dynamics or rewards.

Assumption 3.3.2. *Let r, p, W be Lipschitz continuous with Lipschitz constants $L_r, L_p, L_W > 0$.*

Remark 3.3.1. *For all but Theorem 3.3.1, we may alternatively let W be Lipschitz on finitely many disjoint hyperrectangles, i.e. let there be disjoint intervals $\{I_1, \dots, I_Q\}$, $\cup_i I_i = \mathcal{I}$ such that $\forall i \in \{1, \dots, Q\}$, $\forall \alpha, \tilde{\alpha} \in I_i$, $\forall d \in [D]$, $\forall \beta \in \mathcal{I}^{r<[k_d]\setminus\{1\}}$ we have*

$$|W_d(\alpha, \beta) - W_d(\tilde{\alpha}, \beta)| \leq L_W |\alpha - \tilde{\alpha}| . \tag{3.3.35}$$

Remark 3.3.2. *Note that our model is quite general: In particular, it is also possible to model dynamics and rewards dependent on the state-action distributions instead of only state distributions, replacing $\delta_{\times_{j\neq i} x_t^{m_j}}$ by $\delta_{\times_{j\neq i}(x_t^{m_j}, u_t^{m_j})}$ in Eq. (3.3.26). This can be done by reformulating any problem as follows. Assume a problem with state and action spaces $\mathcal{X}$, $\mathcal{U}$ and dependence of rewards and transitions on joint state-action distributions. We can rewrite the problem as a new problem with new state space $\mathcal{X} \cup (\mathcal{X} \times \mathcal{U})$, where in the new problem, each two decision epochs t, $t+1$ correspond to a single original decision epoch, where in the first step t we transition deterministically from $x_t^{m_j}$ to $(x_t^{m_j}, u_t^{m_j})$ for the taken action $u_t^{m_j}$, while in the second step $t+1$ we transition and compute rewards according to the original system, ignoring any second actions taken. Choosing the square root of the discount factor and normalizing rewards will give a problem in our form that is equivalent to the original problem.*

3.3.2 *Theoretical Foundations*

In this section, we rigorously motivate the MF formulation by providing existence and approximation results of an HMFE. Essentially, HMFE are guaranteed to exist and will give approximate Nash equilibria in finite hypergraph games with many agents. The reader interested primarily in applications may skip this section.

We lift the empirical distributions and policies to the continuous domain $\mathcal{I}$, i.e. for any $(\pi^1, \dots, \pi^N) \in \Pi^N$ we define the step policy $\boldsymbol{\pi}^N \in \boldsymbol{\Pi}$ and step empirical measures $\boldsymbol{\mu}^N \in \mathcal{M}$ by

$$\pi_t^{N,\alpha} \coloneqq \sum_{i\in[N]} \mathbf{1}_{I_i^N}(\alpha) \cdot \pi_t^i, \quad \forall (\alpha, t) \in \mathcal{I} \times \mathcal{T}, \tag{3.3.36}$$

$$\mu_t^{N,\alpha} \coloneqq \sum_{i\in[N]} \mathbf{1}_{I_i^N}(\alpha) \cdot \delta_{x_t^j}, \quad \forall (\alpha, t) \in \mathcal{I} \times \mathcal{T}. \tag{3.3.37}$$

Proofs for the results to follow can be found in the Appendix and are at least structurally similar to proofs in [7], though they contain a number of additional considerations we highlight in Appendix C.

3.3.2.1 *Existence of Equilibria*

First, we show that there exists an HMFE. We do this by rewriting the problem in a more convenient form as done in [7]. Consider an equivalent, more standard MFG with states $(\alpha_t, \tilde{x}_t)$, i.e. we integrate the graphon indices α into the state. The newfound states follow the initial distribution $\tilde{x}_0 \sim \mu_0$, $\alpha_0 \sim \mathrm{Unif}(\mathcal{I})$. Then, the actions and original state transitions follow as before, while the α_t part of the state remains fixed at all times, i.e.

$$\begin{aligned} \tilde{u}_t &\sim \tilde{\pi}_t(\tilde{u}_t \mid \tilde{x}_t, \alpha_t), \\ \tilde{x}_{t+1} &\sim p(\tilde{x}_{t+1} \mid \tilde{x}_t, \tilde{u}_t, \tilde{\nu}_t), \quad \alpha_{t+1} = \alpha_t \end{aligned} \tag{3.3.38}$$

where we used the standard (non-graphical) MF $\tilde{\mu}_t \in \mathcal{P}(\mathcal{X} \times \mathcal{I})$ (cf. [24]) and let

$$\tilde{\nu}_{t,d}(x) = \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} W_d(\alpha_t, \boldsymbol{\beta}) \prod_{j=1}^{k_d-1} \tilde{\mu}_t(x_j, \beta_j) \, \mathrm{d}\boldsymbol{\beta}, \tag{3.3.39}$$

Using existing results for MFGs [24], we obtain existence of a potentially non-unique HMFE.

Theorem 3.3.1. *Under Assumption 3.3.2, there exists a HMFE* $(\boldsymbol{\pi}, \boldsymbol{\mu}) \in \boldsymbol{\Pi} \times \mathcal{M}$.

For uniqueness results, we refer to existing results such as the classical monotonicity condition [22, 117]. However, using existing theory will not analyze the finite hypergraph structure and instead directly uses the limiting hypergraphons. In the following, we thus show also that the finite hypergraph games are indeed approximated well.

3.3.2.2 *Approximation Properties*

Next, we will show that the finite hypergraph game and its dynamics are well-approximated by the hypergraphon MFG, which implies that the HMFE solution of the hypergraphon MFG will give us the desired (ε, δ)-Nash equilibrium in large finite hypergraph games.

To begin, we define and obtain finite N-agent system equilibria from an HMFE via the policy sharing map $\mathrm{Id}_N(\boldsymbol{\pi}) \coloneqq (\pi^1, \dots, \pi^N) \in \Pi^N$, i.e. Id_N is defined such that each agent will act according to its position α on the hypergraphon,

$$\pi_t^i(u \mid x) \coloneqq \pi_t^{\frac{i}{N}}(u \mid x), \quad \forall (i, t, x, u) \in [N] \times \mathcal{T} \times \mathcal{X} \times \mathcal{U}. \tag{3.3.40}$$

Now consider $(i, \hat{\pi})$-deviated policy tuples where the i-th agent deviates from an equilibrium policy tuple to its own policy $\hat{\pi}$, i.e. policy tuples $(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N)$. Note that this includes the deviation-free case as a special case. In order to obtain a (ε, δ)-Nash equilibrium, we must show that for almost all i and policies $\hat{\pi}$, the $(i, \hat{\pi})$-deviated policy tuple will be approximately described by the interaction with the limiting hypergraphon MF. For this purpose, the first step is to show the convergence of agent state distributions to the MF.

Define for any $n \in \mathbb{N}$ the evaluation of measurable functions $f \colon \mathcal{X}^n \times \mathcal{I}^n \to \mathbb{R}$ under any n-dimensional product measures $\bigotimes^n \boldsymbol{\mu} \in \mathcal{P}(\mathcal{X}^n)^{\mathcal{I}^n}$ as

$$\boldsymbol{\mu}(f) \coloneqq \int_{\mathcal{I}^n} \sum_{x \in \mathcal{X}^n} f(x, \boldsymbol{\beta}) \prod_{i \in [n]} \mu^{\beta_i}(x_i) \, \mathrm{d}\boldsymbol{\beta}, \tag{3.3.41}$$

where $\bigotimes^n \boldsymbol{\mu}$ denotes the n-fold product of the measure $\boldsymbol{\mu}$, i.e. the n-dimensional distribution over agent states.

Then, our first main result is the convergence of the finite-dimensional agent state marginals to the limiting deterministic MF, given sufficient regularity of the applied policy. For this purpose, we introduce and optimize over a class $\boldsymbol{\Pi}_{\text{Lip}}$ of Lipschitz-continuous policies up to at most D_π discontinuities, i.e. $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\text{Lip}}$ whenever $\alpha \mapsto \pi_t^\alpha$ at any time t has at most D_π discontinuities. Note however, that we could in principle approximate non-Lipschitz policies by classes of Lipschitz-continuous policies.

Theorem 3.3.2. *Consider a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\text{Lip}}$ with associated MF $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$. Let $(\pi^1, \dots, \pi^N) = \text{Id}_N(\boldsymbol{\pi})$, $\hat{\pi} \in \Pi$, $t \in \mathcal{T}$. Under the policy tuple $(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) \in \Pi^N$ and Assumption 3.3.1, we have for all finite dimensionalities $n \in \mathbb{N}$ and all measurable functions $f \colon \mathcal{X}^n \times \mathcal{I}^n \to \mathbb{R}$ uniformly bounded by fixed $M_f > 0$, that*

$$\mathbb{E}\left[\left|\bigotimes^n \boldsymbol{\mu}_t^N(f) - \bigotimes^n \boldsymbol{\mu}_t(f)\right|\right] \to 0, \tag{3.3.42}$$

uniformly over all possible deviations. Furthermore, the rate of convergence follows the hypergraphon convergence rate in Assumption 3.3.1 up to $O(1/\sqrt{N})$.

As a special case, by considering $n = 1$ and $f = \mathbf{1}_{\{x\}}$ for any $x \in \mathcal{X}$, we find convergence in L_1 of the empirical distribution of agent states $\frac{1}{N}\sum_{i\in[N]} \delta_{x_t^i}$ to the limiting MF $\int_{\mathcal{I}} \mu_t^\alpha \, \mathrm{d}\alpha$.

Our second main result is the (uniform) convergence of the system for almost any agent $i \in [N]$ with deviating policy $\hat{\pi} \in \Pi$ to the system where the interaction with other agents is replaced by the interaction with the limiting deterministic MF. Hence, we introduce new random variables for the single deviating agent, beginning with initial distribution $\hat{x}_0^{\frac{i}{N}} \sim \mu_0$. The action variables follow the deviating policy

$$\hat{u}_t^{\frac{i}{N}} \sim \hat{\pi}_t(\hat{u}_t^{\frac{i}{N}} \mid \hat{x}_t^{\frac{i}{N}}), \tag{3.3.43}$$

with the state transition laws

$$\hat{x}_{t+1}^{\frac{i}{N}} \sim p(\hat{x}_{t+1}^{\frac{i}{N}} \mid \hat{x}_t^{\frac{i}{N}}, \hat{u}_t^{\frac{i}{N}}, \nu_t^{\frac{i}{N}}), \tag{3.3.44}$$

i.e. we assume that all other agents act according to their corresponding equilibrium policy $\text{Id}_N(\boldsymbol{\pi})$, such that the neighborhood state distributions of most agents can be replaced by the limiting term $\nu_t^{\frac{i}{N}}$ with little error in large hypergraphs.

Theorem 3.3.3. *Consider a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\text{Lip}}$ with associated MF $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$. Let $(\pi^1, \dots, \pi^N) = \text{Id}_N(\boldsymbol{\pi})$, $\hat{\pi} \in \Pi$, $t \in \mathcal{T}$. Under the policy tuple $(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) \in \Pi^N$ and Assumptions 3.3.1 and 3.3.2, for any uniformly bounded family of functions $\mathcal{G}$ from $\mathcal{X}$ to $\mathbb{R}$ and any $\varepsilon, p > 0$, $t \in \mathcal{T}$, there exists $N' \in \mathbb{N}$ such that for all $N > N'$*

$$\sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[g(x_t^i)\right] - \mathbb{E}\left[g(\hat{x}_t^{\frac{i}{N}})\right] \right| < \varepsilon \tag{3.3.45}$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}^N$ for some $\mathcal{J}^N \subseteq [N]$, $|\mathcal{J}^N| \geq \lfloor (1-p)N \rfloor$.

Further, for any uniformly Lipschitz, uniformly bounded family of measurable functions $\mathcal{H}$ from $\mathcal{X} \times \mathcal{B}_1(\mathcal{X})$ to $\mathbb{R}$ and any $\varepsilon, p > 0$, $t \in \mathcal{T}$, there exists $N' \in \mathbb{N}$ such that for all $N > N'$

$$\sup_{h \in \mathcal{H}} \left| \mathbb{E}\left[h(x_t^i, \nu_t^{N,i}) \right] - \mathbb{E}\left[h(\hat{x}_t^{\frac{i}{N}}, \nu_t^{\frac{i}{N}}) \right] \right| < \varepsilon \tag{3.3.46}$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}^N$ for some $\mathcal{J}^N \subseteq [N]$ with $|\mathcal{J}^N| \geq \lfloor (1-p)N \rfloor$.

As a corollary, we will have good approximation of the finite hypergraph game objective through the hypergraphon MF objective, and correspondingly the approximate Nash property of hypergraphon MFE, motivating the hypergraphon MFG framework.

Corollary 3.3.1. *Consider a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\text{Lip}}$ with associated MF $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$. Let $(\pi^1, \dots, \pi^N) = \mathrm{Id}_N(\boldsymbol{\pi})$, $\hat{\pi} \in \Pi$, $t \in \mathcal{T}$. Under the policy tuple $(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) \in \Pi^N$ and Assumptions 3.3.1 and 3.3.2, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have*

$$\left| J_i^N(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) - J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\hat{\pi}) \right| < \varepsilon \tag{3.3.47}$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}^N$ for some $\mathcal{J}^N \subseteq [N]$ with $|\mathcal{J}^N| \geq \lfloor (1-p)N \rfloor$.

Corollary 3.3.2. *Consider an HMFE $(\boldsymbol{\pi}, \boldsymbol{\mu}) \in \boldsymbol{\Pi}_{\text{Lip}} \times \mathcal{M}$. Under Assumptions 3.3.1 and 3.3.2, for any $\varepsilon, \delta > 0$ there exists N' such that for all $N > N'$, the policy $(\pi^1, \dots, \pi^N) = \mathrm{Id}_N(\boldsymbol{\pi})$ is an (ε, δ)-Nash equilibrium.*

Therefore, we find that a solution of the MF system is a good equilibrium solution of sufficiently large finite hypergraph games.

The assumption of a class $\boldsymbol{\Pi}_{\text{Lip}}$ of Lipschitz continuous policies up to finitely many discontinuities may seem restrictive. However – similar to [7, Theorem 5] – we may discretize and partition $\mathcal{I}$ in order to solve hypergraphon MFGs to an arbitrary degree of exactness, preserving the good approximation properties on large hypergraph games.

3.3.3 *Experiments*

In this section, we shall introduce an exemplary numerical problem of rumor spreading, and show associated numerical solutions to demonstrate the hypergraphon MF framework, verifying the theoretical results.

In order to learn an HMFE in our model, we shall adopt the well-founded discretization method proposed in [7] analogous to the technique used in the proof of Theorem 3.3.1 to convert the GMFG into a classical MFG, and thereby allow application of any existing MFG algorithms such as FPI to solve for an equilibrium. In other words, we will split $\mathcal{I}$ into subintervals $I_1^N, \dots, I_N^N$, for each of which we will pick a representing $\alpha \in I_i^N$. This α together with an agent's original state in $\mathcal{X}$ will constitute the new state. In Appendix C.6, we perform additional experiments for another numerical problem of epidemics control, where existing algorithms fail, pointing out potential future work.

3.3.3.1 *Hypergraphons*

In our experiments, we shall sample finite hypergraphs directly from given limiting hypergraphons, which should ensure that we obtain hypergraph sequences that fulfill Assumption 3.3.1 analogous to the standard graphon case at rate $O(\frac{1}{\sqrt{\log N}})$, see [157, Lemma 10.16]. To sample a k-uniform hypergraph with N nodes from a k-uniform hypergraphon W, we sample $|r_<[k]|$ uniformly distributed values from the unit interval $\{\alpha_{\mathbf{j}} \colon \alpha_{\mathbf{j}} \sim \mathrm{Unif}([0,1])\}_{\mathbf{j} \in r_<[k]}$. Then, we add any hyperedge $B \subseteq [N]$ with probability $W(\alpha_{r_<(B)})$.

For the sake of illustration, unless otherwise noted, we will in the following consider two-layer hypergraphons, where the first layer is a 2-uniform hypergraph (standard graph), while the second layer shall be a 3-uniform hypergraph. For the first layer, we consider the uniform attachment graphon

$$W_{\mathrm{unif}}(\alpha_1, \alpha_2) = 1 - \max(\alpha_1, \alpha_2),$$

the ranked attachment graphon

$$W_{\mathrm{rank}}(\alpha_1, \alpha_2) = 1 - \alpha_1 \alpha_2$$

and the flat (or p-ER) random graphon

$$W_{\mathrm{flat}} \coloneqq p = 0.5.$$

In particular, the uniform attachment graphon is the limit of a random graph sequence where we iteratively add a new node N and then connect all unconnected nodes with probability $\frac{1}{N}$. Similarly, for the ranked attachment graphon, at each iteration n we first add a new (n-th) node. Before adding the node, the nodes $1, \ldots, n-1$ exist from prior iterations. The new node n is connected to all previous nodes $i = 1, \ldots, n-1$ with probability $1 - \frac{i}{n}$. Then, all other nodes that are not yet connected with each other will connect with probability $\frac{2}{n}$. See also [157, Chapter 11] and Figure 3.11.

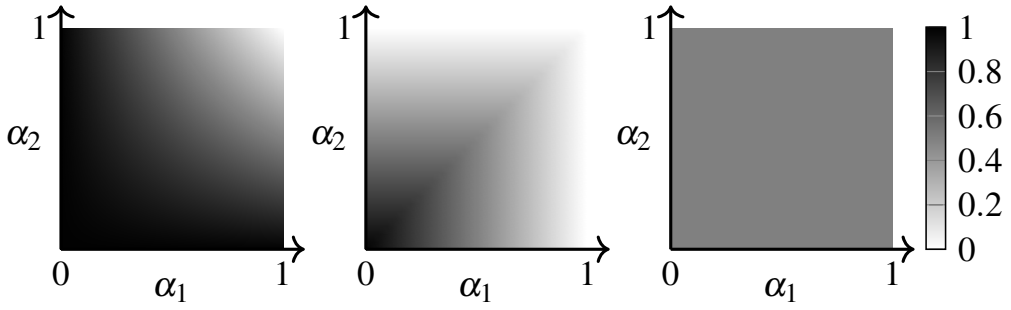


FIGURE 3.11: Visualization of example graphons in the 2-dimensional case. Left: Uniform attachment graphon; Middle: Ranked attachment graphon; Right: 0.5-ER graphon.

For the second, 3-uniform layer, we similarly consider the hypergraphon resulting from converting all triangles in a standard p-ER graph into hyperedges [174]

$$\hat{W}_{\mathrm{ind}}(\boldsymbol{\alpha}) \coloneqq \mathbf{1}_{\mathcal{I}^3 \times [0,p]^3}(\alpha),$$

as well as the uniform attachment hypergraphon

$$\hat{W}_{\mathrm{unif}}(\boldsymbol{\alpha}) = 1 - \max(\alpha_1, \alpha_2, \alpha_3)$$

and its inverted version

$$\hat{W}_{\text{inv}-\text{unif}}(\boldsymbol{\alpha}) = 1 - \max(1-\alpha_1, 1-\alpha_2, 1-\alpha_3)$$

resulting from a similar construction as in the standard case.

3.3.3.2 *Rumor Spreading Dynamics*

In this section, we will describe some simple social dynamics and epidemics problems to illustrate potential applications of hypergraphon MFGs. Here, each layer could model different types of interpersonal relationships. In our particular example of 2-uniform and 3-uniform layers, the latter can model small cliques of friends, while the former could model general acquaintanceship. We do note that social networks are typically more sparse, possessing significantly less edges than on the order of $O(N^2)$. However, our model is a first step towards rigorous limiting hypergraph models and in the future could be extended by using other graph limit theories such as L^p graphons [14, 172, 173] by extending their theory towards hypergraphons. We further imagine that similar approaches could be used e.g. in economics [64] or engineering applications [65].

In the classical Maki-Thompson model [175, 176], spread of rumors is modelled via three node states: ignorant, spreader and stifler. Ignorants are unaware of the rumor, while spreaders attempt to spread the rumor. When spreaders attempt to spread to nodes that are already aware of the rumor too often, they stop spreading and become a stifler. In this work, instead of a priori assuming the above behavior, we will give agents an intrinsic motivation to spread or stifle rumors, giving rise to the Rumor problem. We shall consider ignorant (I) and aware (A) nodes. The behavior of aware nodes is then motivated by the gain and loss of social standing resulting from spreading rumors to ignorant and aware nodes respectively. The possible actions $\mathcal{U} \coloneqq \{\bar{S}, S\}$ of nodes are to actively spread the rumor (S) or to refrain from doing so ($\bar{S}$). The probability of an ignorant node becoming aware of the rumor at any decision epoch is then simply given by a linear combination of all layer neighborhood densities of aware, spreading nodes.

Since transition dynamics will depend on the spreading actions of neighbors, following Remark 3.3.2 we define instead the extended state space $\mathcal{X} = \{I, A\} \cup (\{I, A\} \times \mathcal{U})$. We then assume the dynamics are given at all times t by

$$\begin{aligned} p((x,u) \mid x, u, \boldsymbol{\nu}) &= 1, \quad p(A \mid (A,u'), u, \boldsymbol{\nu}) = 1, \\ p(A \mid (I,u'), u, \boldsymbol{\nu}) &= 1 - p(I \mid (I,u'), u, \boldsymbol{\nu}) \\ &= \min\left(1, \sum_{d\in[D]} \tau_d \boldsymbol{\nu}_d \left(\sum_{i\in[k_d]} \mathbf{1}_{\{(A,S)\}}(x \mapsto x_i)\right)\right) \end{aligned}$$

for all $x \in \{I, A\}$, $u, u' \in \mathcal{U}$, and similarly the rewards are given

$$R((A,S), u, \boldsymbol{\nu}) = \sum_{d\in[D]} \boldsymbol{\nu}_d \left(\sum_{i\in[k_d]} r_d \mathbf{1}_{\{I\}\times\mathcal{U}}(x \mapsto x_i) - c_d \mathbf{1}_{\{A\}\times\mathcal{U}}(x \mapsto x_i)\right)$$

with $R \equiv 0$ otherwise. In other words, any aware and spreading agent obtains a reward in each layer that is proportional to the probability of a neighbor of any hyperedge sampled uniformly-at-random out of all connected hyperedges to be ignorant. In our experiments, we use $\tau_1 = 0.3$, $\tau_2 = 0.5$, $r_d = 0.5$, $c_d = 0.8$, $\mu_0(A) = 0.01$ and $\mathcal{T} = \{0, 1, \ldots, 49\}$.

3.3.3.3 *Numerical Results*

In our experiments, we restrict ourselves to finite time horizons with $\gamma = 1$, 50 discretization points, and use backwards induction with exact forward propagation to compute exact solutions. Note that simple FPI by repeatedly computing an arbitrary optimal deterministic policy and its corresponding MF converges to an equilibrium in the Rumor problem. In general however, FPI (as well as more advanced state-of-the-art techniques) may fail to converge, see e.g. the SIS problem in Appendix C.6.

In Figure 3.12, we can observe that the behavior for the Rumor problem is as expected. At the equilibrium, agents will continue to spread rumors until the number of aware agents reaches a critical point at which the penalty for spreading to aware agents is larger than the reward for spreading to ignorant agents. The agents with higher connectivity are more likely to be aware of the rumor. Particularly in the uniform attachment hypergraphon case, the threshold is reached at different times, since the neighborhoods of different α reach awareness at different rates depending on their connectivity. Here, a number of nodes with very low degrees will continue spreading the rumors. In Appendix C.6, we show additional results for inverted 3-uniform hypergraphons, which give similar results to the ones seen here.

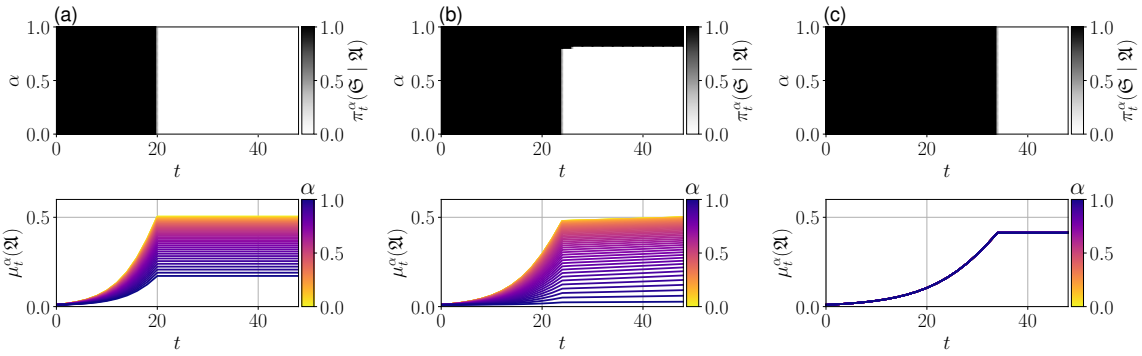


FIGURE 3.12: Equilibrium behavior for the Rumor problem. Top: The threshold policy allows spreading of rumors. It can be seen that agents spread the rumor up until a point in time where too many other agents know of the rumor. As expected, agents are more likely to hear of the rumor if they have more neighbors. (a): $(W_{\text{rank}}, \hat{W}_{\text{unif}})$; (b): $(W_{\text{unif}}, \hat{W}_{\text{unif}})$; (c): $(W_{\text{er}}, \hat{W}_{\text{ind}})$.

Furthermore, as can be seen in Figure 3.13, the L_1 error between the empirical distribution and the limiting MF system (as vectors over time)

$$\Delta\mu = \mathbb{E}\left[\sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}}\left|\frac{1}{N}\sum_{i\in[N]}\delta_{x_t^i}(x) - \int_{\mathcal{I}}\mu_t^\alpha(x)\,\mathrm{d}\alpha\right|\right] \tag{3.3.48}$$

goes to zero as the number of agents increases, showing that the finite hypergraph game is well approximated by the hypergraphon MFG for sufficiently large systems, though the error remains somewhat large due to the high variance from our sparse initialization $\mu_0(I) = 0.01$. Here, we estimated the error $\Delta\mu$ for each N over 50 realizations. Due to the $O(N^2)$ complexity of simulation and computational constraints, our experiments remain limited to the demonstrated number of agents.

We repeat the experiment in Figure 3.14 with a more dense initialization $\mu_0(A) = 0.1$ to reduce the aforementioned high contribution of variance from random initializations. Here, we observe that the resulting convergence is significantly faster.

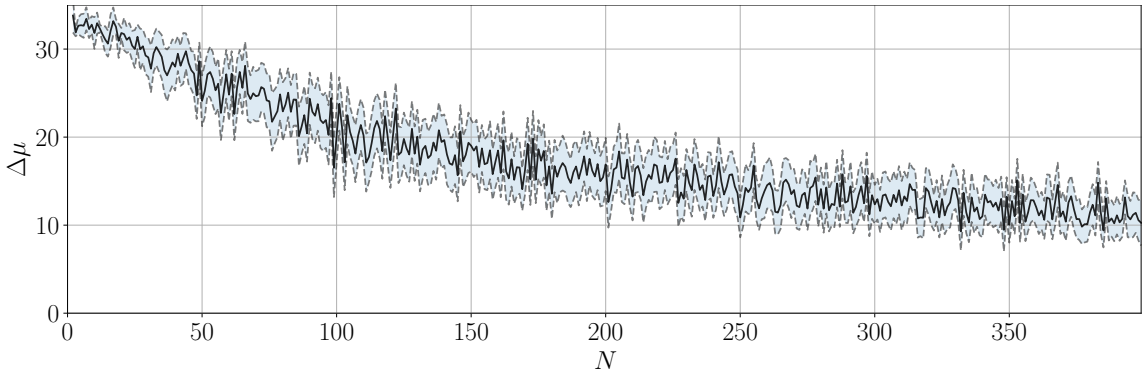


FIGURE 3.13: Convergence of the empirical MF in the limit. We compare between the fraction of aware nodes in the finite and MF system under the equilibrium policy for $(W_{\text{rank}}, \hat{W}_{\text{inv}-\text{unif}})$ from Figure C.1(a) in Appendix C.6, averaged over 50 stochastic simulations. The shaded region depicts the 95% confidence interval at each N. It can be seen that the state distributions are increasingly well approximated by the MF.

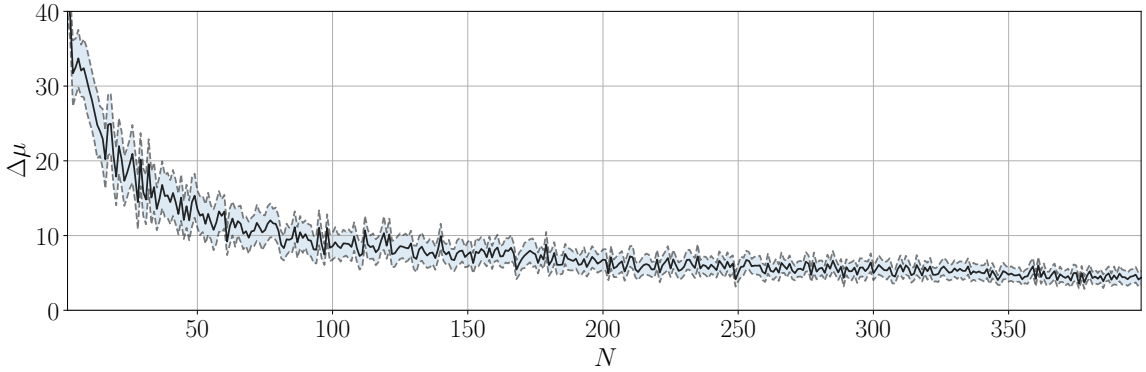


FIGURE 3.14: Convergence of the empirical MF under non-sparse initialization. We compare between the fraction of aware nodes in the finite and MF system under the equilibrium policy for $(W_{\text{rank}}, \hat{W}_{\text{inv}-\text{unif}})$ as in Figure 3.13, but with higher initial awareness. It can be seen that convergence is much faster, since the effect of random sparse initialization is avoided.

Lastly, in Figure 3.15 we demonstrate some interesting non-linear behavior for a two-layer setting where both layers consist of 3-uniform hypergraphs. Here, for the first layer we use the block hypergraphon

$$\hat{W}_{\text{block}}(\boldsymbol{\alpha}) \coloneqq \mathbf{1}_{[0,0.5]^3 \times [0,p]^3}(\alpha) + \mathbf{1}_{(0.5,1]^3 \times [0,p]^3}(\alpha),$$

for $p = 0.5$, while for the second layer we again use the inverted uniform attachment hypergraphon. In other words, we have a structure of two blocks on the first layer, while the second layer is more globally connected. Furthermore, we will initialize the rumor in the second block where $\alpha > 0.5$, i.e. $\mu_0^{\alpha}(A) = \mathbf{1}_{(0.5,1]}(\alpha)$. As we can see in Figure 3.15, in the beginning the rumor spreads in the second block $\alpha > 0.5$ where it originated from. After a while however, the rumor begins to spread faster in the first block $\alpha \leq 0.5$, since nodes with low α are significantly more interconnected on the second layer.

Overall, we can see that multi-layer hypergraphon MFGs allow for more complex behavior and modelling of connections than a single-layer graphon approach.

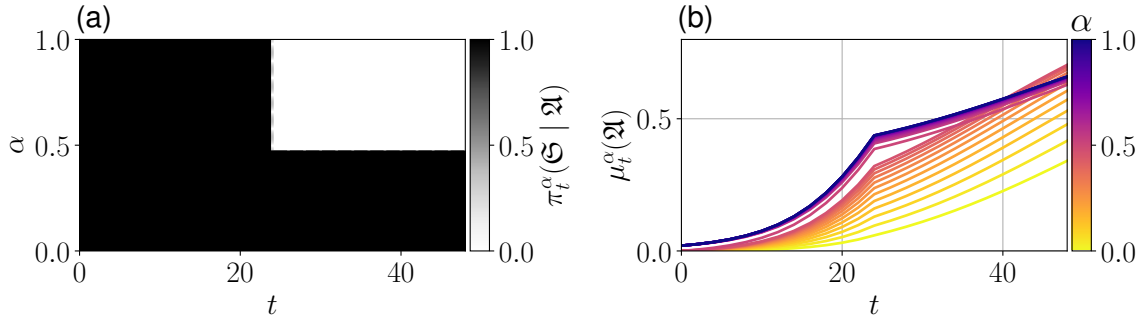


FIGURE 3.15: Equilibrium behavior for the Rumor problem with two-layer 3-uniform hypergraphs $(\hat{W}_{\text{block}}, \hat{W}_{\text{unif}})$. We observe that the rumor originates in nodes with $\alpha > 0.5$, but nodes with $\alpha \leq 0.5$ eventually catch up due to their increased connectivity. (a): The equilibrium threshold policy. (b): MF for each α.

3.3.4 *Summary*

In this section, we introduced a model for dynamical systems on hypergraphs that can describe agents with weak interaction via the graph structure. The model allows for a rigorous and simple MF description that has a complexity independent of the number of agents. We verify our approach both theoretically and empirically on a rumor spreading example. By introducing game-theoretical ideas, we thus obtain a framework for solving otherwise intractable large-scale games on hypergraphs in a tractable manner, going beyond simple pair-wise interactions.

We hope our work forms the basis for several future works, e.g., extensions to directed or weighted hypergraphs in order to generalize to arbitrary network motifs [18], adaptive networks [177], cooperative control or consideration of edge states in addition to the vertex states we have considered in this work. Furthermore, it may be of interest to consider graph models with more adjustable clustering parameters. An extension of our rumor model and theory to continuous-time models could be fruitful. Finally, so far our work remains restricted to dense graphs and deterministic limiting graphons, while in practice this is not always the case (e.g. preferential attachment graphs [158]). Here, L^p graphons [14, 172, 173] could provide a description for less dense cases, which are of great practical interest and may also be generalized to hypergraphs. We also hope that our work inspires future applications in inherently (hyper-)graphical scenarios.

3.4 BEYOND WEAK INTERACTION OF AGENTS

Standard MFGs remain limited to homogeneous agents that weakly influence each other, and cannot model major agents that strongly influence other agents, severely limiting the class of problems that can be handled. We propose a novel discrete time version of Major-Minor Mean Field Games (M3FGs), along with a learning algorithm based on FP and partitioning the probability simplex. Importantly, M3FGs generalize MFGs with common noise and can handle not only random exogeneous environment states but also major agents. A key challenge is that the MF is stochastic and not deterministic as in standard MFGs. Our theoretical investigation verifies both the M3FG model and its algorithmic solution, showing firstly the well-posedness of the M3FG model starting from a finite game of interest, and secondly convergence and approximation guarantees of the FP algorithm. Then, we empirically verify the obtained theoretical results, ablating some of the theoretical assumptions made, and show successful equilibrium learning in three example problems. Overall, we establish a learning framework for a novel and broad class of tractable games. The material presented in this section is based upon our work [3].

So far, most MFG learning frameworks remain unable to handle common noise [178], or more generally major agents. Contrary to minor agents, a major agent directly affects all minor agents and is affected by the MF of minor agents, whereas common noise also affects all minor agents, but is exogeneous and can be understood as a static major agent without actions [179]. Notably, [127] formulate an algorithm handling common noise using a continuous learning Lyapunov argument [180, 181], assuming however that the common noise is known, while [1] consider a cooperative setting. Common noise and major agents remain important in practice, as a system seldom consists only of many similar minor agents. For example, strategic agents on the market do not exist in a vacuum but must contend for instance with idiosyncratic shocks [64] or government regulators [62], while many cars on a road network [54] may be subject to traffic accidents or traffic lights. In continuous-time, such systems are known as MFGs with major and minor agents [67], and have been considered, e.g., by [182–184] for LQG systems, by [113, 185] in non-linear and partially observed settings, and more recently by [186–189]. Major agents also generalize common noise, an important problem in MFG literature [121, 127, 178]. For an additional overview, we also point to [67]. In contrast to prior work, we focus on a computational learning framework that is in discrete time. Additionally, even existing discrete-time MFG frameworks with only common noise such as by [127] have to the best of our knowledge not yet rigorously connected MFGs with the finite games of practical interest. We note that another setting with major agents has already been explored: Stackelberg MFGs. [190–192] consider a Stackelberg equilibrium instead of a Nash equilibrium, wherein a 'major' principal agent chooses their policy first and has priority (like a government or regulator); see [193, 194] for discrete time versions of the problem. Though the Stackelberg setting is of importance, it is distinct from computing Nash equilibria where major and minor agents are "on the same level": in the Stackelberg setting, minor agents only respond with a Nash equilibrium between themselves *after* the principal's policy choice. Furthermore, we are not aware of any propagation of chaos results even in discrete-time Stackelberg MFGs, for which our result also applies. The field of Stackelberg MFGs remains part of continued active research, to which our M3FG setting may also contribute, and vice versa.

OUR CONTRIBUTION. By the preceding motivation, we propose the first general discrete-time M3FG learning framework. We begin with providing a theoretical foundation of the proposed M3FG model, showing that equilibria in finite games with many agents can be approximately learned in the M3FG instead. The proof is based upon showing propagation of chaos i.e., convergence of the empirical MF, which – in contrast to its counterpart in MFGs without common noise – converges only

in distribution. We then move on to provide a learning algorithm based on FP to solve M3FGs, with convergence results and approximation guarantees for its tractable and practical, tabular variant. Empirically, our learned policies do not assume that common noise is known a priori. Due to the resulting stochastic MF, for tractable dynamic programming we allow conditioning of agent actions and policies also on the MF instead of just the agent's own state. Finally, we verify the M3FG framework on three problems, empirically supporting theoretical claims, even when the assumptions are not entirely fulfilled.

3.4.1 *Major-Minor Mean Field Games*

In this section, we begin by giving a description of considered problems and their corresponding MF system.

3.4.1.1 *Finite Agent Game*

We consider a game with N minor agents and one major agent. Let $\mathcal{X}$ and $\mathcal{U}$ be finite state and action spaces for minor agents, respectively. Let $\mathcal{X}^0$ and $\mathcal{U}^0$ be finite state and action spaces for the major agent, respectively. Let $T \in \mathbb{N}$ be a finite time horizon and let $\mathcal{T} \coloneqq \{0, 1, \dots, T-1\}$. We denote the state and the action of minor agent $i \in [N]$ at time $t \in \mathcal{T}$ by $x_t^{i,N}$ and $u_t^{i,N}$, respectively. Similarly, we denote by $x_t^{0,N}$ and $u_t^{0,N}$ the state and the action of the major agent at time t. Let μ_0 and μ_0^0 be initial probability distributions on $\mathcal{X}$ and $\mathcal{X}^0$, respectively. Define the empirical MF $\mu_t^N \coloneqq \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{x_t^{i,N}}$, where $\mathbf{1}_x$ is the indicator function equal to 1 for the argument x and 0 otherwise. The MF can be viewed as a histogram with $|\mathcal{X}|$ many bins.

We can consider several classes of policies. In this presentation, we focus on Markovian feedback policies in the following sense: the policy $\pi^{i,N}$ for minor agent i is a function of her own state, the major agent's state and the MF; the policy $\pi^{0,N}$ for the major agent is a function of her own state and the MF. We denote respectively by Π and Π_0 the sets of such minor and major agent policies.

For a given tuple of policies $(\underline{\pi}^N, \pi^{0,N}) = ((\pi^{1,N}, \dots, \pi^{N,N}), \pi^{0,N}) \in \Pi^N \times \Pi_0$, the game begins with states $x_0^{0,N} \sim \mu_0^0$, $x_0^{i,N} \sim \mu_0$ and subsequently, for $t = 0, 1, \dots, T-2$, let

$$u_t^{i,N} \sim \pi_t^{i,N}(u_t^{i,N} \mid x_t^{i,N}, x_t^{0,N}, \mu_t^N), \quad i \in [N] \tag{3.4.49a}$$
$$u_t^{0,N} \sim \pi_t^{0,N}(u_t^{0,N} \mid x_t^{0,N}, \mu_t^N), \tag{3.4.49b}$$
$$x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N), \quad i \in [N] \tag{3.4.49c}$$
$$x_{t+1}^{0,N} \sim p^0(x_{t+1}^{0,N} \mid x_t^{0,N}, u_t^{0,N}, \mu_t^N). \tag{3.4.49d}$$

where $p\colon \mathcal{X} \times \mathcal{U} \times \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ and $p^0\colon \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ are transition kernels.

In contrast to classic MFGs such as studied e.g, in [24], the minor agents' dynamics depend also on the major agent's state. An important consequence is that the minor agents' dynamics are influenced by a form of common noise. This explains why we decide to consider policies that depend on the MF μ_t^N. Furthermore, this form of common noise is not simply an exogenous source of randomness because it is influenced by the major agent's choice of policy. This makes the problem more challenging than MFGs with common noise.

Next, we define the minor and major total rewards

$$J_N^i(\underline{\pi}^N, \pi^{0,N}) = \mathbb{E}\left[\sum_{t\in\mathcal{T}} r(x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N)\right],$$
$$J_N^0(\underline{\pi}^N, \pi^{0,N}) = \mathbb{E}\left[\sum_{t\in\mathcal{T}} r^0(x_t^{0,N}, u_t^{0,N}, \mu_t^N)\right],$$

for some reward functions $r\colon \mathcal{X}\times\mathcal{U}\times\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})\to\mathbb{R}$ and $r^0\colon \mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})\to\mathbb{R}$.

In this work, we focus on the non-cooperative scenario where agents try to maximize their own objectives while anticipating the behavior of other agents. This is formalized by the solution concept of (approximate) Nash equilibria.

Definition 3.4.1. *Let $\varepsilon \geq 0$. An approximate ε-Nash equilibrium is defined as a tuple of policies $(\underline{\pi}^N, \pi^{0,N}) = ((\pi^{1,N}, \dots, \pi^{N,N}), \pi^{0,N}) \in \Pi^N \times \Pi_0$, such that we have $J_N^0(\underline{\pi}^N, \pi^{0,N}) \geq \sup_{\tilde{\pi}^0} J_N^0((\pi^{1,N}, \dots, \pi^{N,N}), \tilde{\pi}^0) - \varepsilon$ for the major agent, and furthermore $J_N^i(\underline{\pi}^N, \pi^{0,N}) \geq \sup_{\tilde{\pi}^i\in\Pi} J_N^i((\pi^1, \dots, \pi^{i-1}, \tilde{\pi}^i, \pi^{i+1}, \dots, \pi^N), \pi^{0,N}) - \varepsilon$ for all minor agents $i \in [N]$. A Nash equilibrium is an approximate 0-Nash equilibrium.*

Remark 3.4.1. *We can also consider time-dependent dynamics or rewards, multiple major agents, and infinite-horizon discounted objectives. Some results we prove below can be extended to such settings (e.g., propagation of chaos, equilibrium approximation, and FP; see also generalized infinite-horizon experiments in Appendix D.9). Similarly, we can extend the model to multiple minor agent populations with small changes, see e.g. [117]. Another possibility is to simply include types of agents into their state [105].*

3.4.1.2 *Mean Field Game*

When the number of minor agents N is large, we can approximate the game by an MFG, which corresponds formally to the limit $N\to\infty$. In an MFG, the empirical MF is replaced by a random limiting MF. Unlike standard MFGs, the limiting MF does not evolve in a deterministic way due to the influence of the major agent. Fixing major and minor agent policies π^0, π for all agents, except for a single minor agent deviating to $\hat{\pi}$, when $N\to\infty$, we obtain (intuitively by a LLN argument) the major and deviating minor agent M3FG dynamics $x_0^0 \sim \mu_0^0$, $x_0 \sim \mu_0$,

$$u_t \sim \hat{\pi}_t(u_t \mid x_t, x_t^0, \mu_t), \tag{3.4.50a}$$
$$u_t^0 \sim \pi_t^0(u_t^0 \mid x_t^0, \mu_t), \tag{3.4.50b}$$
$$x_{t+1} \sim p(x_{t+1} \mid x_t, u_t, x_t^0, u_t^0, \mu_t), \tag{3.4.50c}$$
$$x_{t+1}^0 \sim p^0(x_{t+1}^0 \mid x_t^0, u_t^0, \mu_t), \tag{3.4.50d}$$
$$\mu_{t+1} = T_t^\pi(x_t^0, u_t^0, \mu_t) \tag{3.4.50e}$$

with deterministic transitions $T_t^\pi(x^0, u^0, \mu) \coloneqq \iint p(x, u, x^0, u^0, \mu)\,\pi_t(\mathrm{d}u \mid x, x^0, \mu)\,\mu(\mathrm{d}x)$ as the conditional "expectation" of the next MF given the current major state x^0, action u^0, and random MF μ. The policy π is shared by all minor agents except one who is deviating and using $\hat{\pi}$. This means that we look for symmetric Nash equilibria where all exchangeable minor agents use the same policy, as usual in MFG literature. Still, a MFE suffices as an approximate Nash equilibrium in the finite game, which is not to say that there cannot be other heterogeneous policy tuples in the finite game that are Nash.

M3FGs now consist of *two* MDP optimality conditions, one for all minor agents and one for the major agent. An equilibrium is then optimal in each MDP simultaneously. More precisely, from the point of view of a minor agent, the goal is to optimize over $\hat{\pi}$ while (π, π^0) are fixed. This yields the minor agent MDP with state $(x_t, x_t^0, \mu_t) \in \mathcal{X} \times \mathcal{X}^0 \times \mathcal{P}(\mathcal{X})$, and action $u_t \in \mathcal{U}$, and with the objective

$$J(\hat{\pi}, \pi, \pi^0) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} r(x_t, u_t, x_t^0, u_t^0, \mu_t)\right]. \tag{3.4.51}$$

Note that, although μ_{t+1} is given by a deterministic function of (x_t^0, u_t^0, μ_t), from the point of view of a minor agent, the evolution of $(\mu_t)_t$ is stochastic since it depends on the sequence $(x_t^0, u_t^0)_t$, which is random. By definition of a Nash equilibrium, only a *single* minor agent can deviate arbitrarily to $\hat{\pi}$, and by symmetry it does not matter which "representative" minor agent deviates. Therefore there is only one MDP optimality condition for all minor agents. We also stress that since $N \to \infty$, the representative agent is insignificant and her deviation does not affect the MF.

On a similar note, from the major agent's point of view, we obtain the major agent MDP with $(\mathcal{X}^0 \times \mathcal{P}(\mathcal{X}))$-valued states (x_t^0, μ_t) and $\mathcal{U}^0$-valued actions u_t^0 of the major agent, using the same dynamics, forgetting about the (insignificant for the major agent) deviating minor agent, and optimizing instead for π^0, the corresponding major objective

$$J^0(\pi, \pi^0) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} r^0(x_t^0, u_t^0, \mu_t)\right]. \tag{3.4.52}$$

MEAN FIELD EQUILIBRIUM. The Nash equilibrium in the finite game hence corresponds to a major-minor MFE, as a fixed point of both MDPs *at once*. In other words, major and minor policies π^0, π that are optimal against themselves in the major and minor agent MDPs.

Definition 3.4.2. *A Major-Minor Mean Field Nash Equilibrium (M3FNE) is a tuple* $(\pi, \pi^0) \in \Pi \times \Pi_0$ *of policies, such that* $\pi \in \arg\max_{\pi'} J(\pi', \pi, \pi^0)$ *and* $\pi^0 \in \arg\max_{\pi'} J^0(\pi, \pi')$.

We slightly weaken the concept of optimality to *approximate* optimality, since the solution of a limiting MFG provides approximate Nash equilibria for the finite game, which are still achieved by solving for approximate M3FNE.

Definition 3.4.3. *An approximate* ε*-M3FNE is a tuple* $(\pi, \pi^0) \in \Pi \times \Pi_0$ *of policies, such that* $J(\pi, \pi, \pi^0) \geq \sup_{\pi'} J(\pi', \pi, \pi^0) - \varepsilon$ *and* $J^0(\pi, \pi^0) \geq \sup_{\pi'} J^0(\pi, \pi') - \varepsilon$.

The minimal such ε for minor and major agents are also referred to as the minor and major exploitabilities $\mathcal{E}(\pi, \pi^0)$ and $\mathcal{E}^0(\pi, \pi^0)$ of (π, π^0). Accordingly, an exploitability of 0 means that (π, π^0) is an exact M3FNE.

3.4.2 *Theoretical Foundations*

The M3FG is a theoretically rigorous formulation for large corresponding finite games. Note in particular that the MF will be stochastic due to the randomness of major agents and their states, and therefore standard results based on determinism of MFs will no longer hold. We provide such a theoretical foundation of M3FG by propagation of chaos.

CONTINUITY ASSUMPTIONS. We provide theoretical guarantees to prove that the M3FNE is an approximate Nash equilibrium in the finite game, despite having a non-deterministic MF in the limiting case, contrary to most of the existing literature [95, 195]. For this, we need some common Lipschitz continuity assumptions [104, 116].

Assumption 3.4.1. *The kernels p, p^0 are L_p, L_{p^0}-Lipschitz.*

Assumption 3.4.2. *The rewards r, r^0 are L_r, L_{r^0}-Lipschitz.*

Assumption 3.4.3. *The classes of major and minor policies Π^0, Π are equi-Lipschitz, i.e. there are $L_{\Pi^0}, L_\Pi > 0$ s.t. for all t, $\pi^0 \in \Pi^0$, $\pi \in \Pi$, we have that $\pi_t^0 \colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{U})$ and $\pi_t \colon \mathcal{X} \times \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{U})$ are L_{Π^0}, L_Π-Lipschitz.*

Here, we always consider Lipschitz continuity for all arguments using the sup metric for products, and the L_1 distance for probability measures, see e.g., Appendix D.2. We note that the Lipschitz assumption for policies – while standard – is technical. Empirically, the only piecewise Lipschitz policies obtained in Section 3.4.3 for tractability nonetheless remain close to the following approximations in the finite system. A theoretical investigation of guarantees for piecewise Lipschitz policies is left for future work.

PROPAGATION OF CHAOS. We achieve propagation of chaos "in distribution" for major and minor agents to the M3FG at rate $\mathcal{O}(1/\sqrt{N})$, which is shown inductively in Appendix D.6. Here, propagation of chaos refers to the conditional independence of minor agents, and thus convergence in the limit to the deterministic MF [126]. In contrast to MFGs with deterministic MFs, a stronger mode of convergence such as the one considered by [24] fails by stochasticity of the MF.

Theorem 3.4.1. *Consider Assumptions 3.4.1 and 3.4.3, and any equi-Lipschitz family of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{U}\times\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})}$ with shared Lipschitz constant $L_\mathcal{F}$. Then, the random variable $(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N)$ in system Eq. (3.4.49) under $((\hat{\pi}, \pi, \pi, \ldots), \pi^0)$ converges weakly, uniformly over $f \in \mathcal{F}$ and $(\hat{\pi}, \pi, \pi^0) \in \Pi \times \Pi \times \Pi^0$, to $(x_t, u_t, x_t^0, u_t^0, \mu_t)$ in system Eq. (3.4.50) under $(\hat{\pi}, \pi, \pi^0)$, for all $t \in \mathcal{T}$,*

$$\sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \left| \mathbb{E}\left[f(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N) \right] - \mathbb{E}\left[f(x_t, u_t, x_t^0, u_t^0, \mu_t) \right] \right| = \mathcal{O}(1/\sqrt{N}). \tag{3.4.53}$$

Corollary 3.4.1. *Similarly, consider Assumptions 3.4.1 and 3.4.3, and any family of equi-Lipschitz functions $\mathcal{F}^0 \subseteq \mathbb{R}^{\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})}$ with shared Lipschitz constant $L_{\mathcal{F}^0}$. Then the random variable $(x_t^{0,N}, u_t^{0,N}, \mu_t^N)$ in system Eq. (3.4.49) under $((\hat{\pi}, \pi, \pi, \ldots), \pi^0)$ converges weakly, uniformly over $f \in \mathcal{F}^0$, to (x_t^0, u_t^0, μ_t) in system Eq. (3.4.50) under $(\hat{\pi}, \pi, \pi^0)$, for all $t \in \mathcal{T}$,*

$$\sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}^0} \left| \mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N) \right] - \mathbb{E}\left[f(x_t^0, u_t^0, \mu_t) \right] \right| = \mathcal{O}(1/\sqrt{N}). \tag{3.4.54}$$

APPROXIMATE NASH EQUILIBRIUM. From propagation of chaos, the approximate Nash property of M3FNE follows, suggesting that a solution of M3FGs provides a good game-theoretic solution of interest to practical N-agent games, see Appendix D.7 for the proof based on propagation of chaos.

Corollary 3.4.2. *Consider Assumptions 3.4.1, 3.4.2 and 3.4.3, and a M3FNE* $(\pi, \pi^0) \in \Pi \times \Pi^0$. *Then, the policies* $((\pi, \dots, \pi), \pi^0)$ *constitute an* $\mathcal{O}(1/\sqrt{N})$*-Nash equilibrium in the finite game.*

Finally, existence of a M3FNE is a difficult question under policies that depend on the stochastic MF. While assuming reactive policies unconditioned on the MF could help, choosing such policies makes the design of our algorithm based on dynamic programming difficult, as policies computed via dynamic programming need to depend on the entire M3FG system state. In contrast, in usual deterministic MFGs it is sufficient to remove policy dependence on the MF, which is deterministic. For practical purposes, learning equilibria and then checking the exploitability by Theorem 3.4.3 may suffice.

3.4.3 *Fictitious Play*

To find M3FNE and solve the fixed-point problem, we formulate a FP algorithm and provide a theoretical analysis. Following the exact algorithm, as empirical contribution we provide and analyze an approximate, numerically tractable algorithm that does not assume knowledge of common noise, contrary to [127], and extend it to the setup with major and minor agents. Since the space of MFs is continuous and does not allow general exact computation of value functions, we project MFs onto a finite partition with guarantees for policy evaluation.

3.4.3.1 *Fictitious Play Algorithm*

In order to learn an M3FNE, we first propose an exact analytic algorithm based on FP [127] and provide a theoretical analysis of convergence. For this part, we will assume that the major agent's action does not affect the minor agents' transition kernel. To simplify the presentation and the analysis, we will use conditioning with respect to the sources of randomness that affect the MF, i.e., the minors' distribution. For every $t \geq 0$, let the major and minor agents' actions be determined not by the MF μ_t, but instead by the history of major states and actions, $u_t^0 \sim \pi_t^0(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)$, $x_{0:t-1}^0 \coloneqq (x_0^0, x_1^0, \dots, x_{t-1}^0)$, $u_{0:t-1}^0 \coloneqq (u_0^0, u_1^0, \dots, u_{t-1}^0)$. By induction, we can in fact view μ_t as a deterministic function of $(x_{0:t-1}^0, u_{0:t-1}^0)$ given the minor agents' policy π, since we simply have $\mu_{t+1} = T_t^\pi(x_t^0, u_t^0, \mu_t)$ recursively and deterministically. This means that for fixed policies such as a given Nash equilibrium, any policies dependent on μ_t can instead be rewritten as functions of $(x_{0:t-1}^0, u_{0:t-1}^0)$. Therefore, instead of seeing policies as functions of μ_t, we will see them as functions of the major agent randomness $(x_{0:t-1}^0, u_{0:t-1}^0)$ and we will write (slightly abusing notation) $\pi_t(x_t, x_{0:t}^0, u_{0:t-1}^0)$ and $\pi_t^0(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)$ respectively for the minor agents' and the major agent's policies. The results we prove below go beyond existing results by (i) analyzing also the major exploitability similarly to the minor exploitability, and (ii) expanding analysis of minor exploitability under presence of major agents. To this end, we formulate Assumption 3.4.4.3. and Assumption 3.4.4.4., which provide the conditions for convergence in the presence of major agents. See Appendix D.1 for more detail.

We first start by introducing the discrete time FP before analyzing it in continuous time. Here, time refers to the algorithm's current iteration and not to the time of the M3FG system, which remains discrete throughout the whole work. At any given step j of FP, we have:

$$\mu^{\bar{\pi}^j}_{t|x^0_{0:t-1},u^0_{0:t-1}} = \frac{j-1}{j}\mu^{\bar{\pi}^{j-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}} + \frac{1}{j}\mu^{\pi^{BR,j}}_{t|x^0_{0:t-1},u^0_{0:t-1}} \tag{3.4.55}$$

where we use the notation $\mu^{\pi}_{t|x^0_{0:t-1},u^0_{0:t-1}}$ for the minor state distribution at time t induced by the minor agent policy π and conditioned on the past sequence $(x^0_{0:t-1}, u^0_{0:t-1})$. Here, $\mu^{\pi^{BR,j}}_{t|x^0_{0:t-1},u^0_{0:t-1}}$ is the conditional distribution induced by the best response (BR) policy $\pi^{BR,j}$ against $\bar{\pi}^{j-1}$ and $\bar{\pi}^{0,j-1}$, i.e., $\pi^{BR,j} := \arg\max_\pi J(\pi, \bar{\pi}^{j-1}, \bar{\pi}^{0,j-1})$. The policy generating this average distribution is

$$\bar{\pi}^j_t(u|x, x^0_{0:t-1}, u^0_{0:t-1}) = \frac{\sum_{i=0}^j \mu^{\pi^{BR,i}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x)\pi^{BR,i}_t(u|x, x^0_{0:t-1}, u^0_{0:t-1})}{\sum_{i=0}^j \mu^{\pi^{BR,i}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x)}. \tag{3.4.56}$$

Meanwhile, the major agent state distribution is

$$\mu^{\bar{\pi}^{0,j}}_t = \frac{j-1}{j}\mu^{\bar{\pi}^{0,j-1}}_t + \frac{1}{j}\mu^{\pi^{0,BR,j}}_t$$

where $\bar{\pi}^{0,j}_t$ analogous to Eq. (3.4.56), but in contrast to minor agents using joint distributions $\mu^{\pi^{0,BR,i}}_t(x^0_t, x^0_{0:t-1}, u^0_{0:t-1})$ and $\pi^{0,BR,i}_t(u^0_t \mid x^0_t, x^0_{0:t-1}, u^0_{0:t-1})$.

For the convergence analysis, we study the continuous time version of above discrete time FP, as [127]. In the continuous time FP algorithm, we denote the time of the algorithm (its "iterations") with τ and we first initialize the algorithm for $\tau < 1$ with arbitrary policies for the minor agents, $\bar{\pi}^{\tau<1} = \{\bar{\pi}^{\tau<1}_t\}_{t\in\mathcal{T}}$, and major agent, $\bar{\pi}^{0,\tau<1} = \{\bar{\pi}^{0,\tau<1}_t\}_{t\in\mathcal{T}}$. For all $\tau \geq 1$, $t \in \mathcal{T}$ and $x^0_{0:t-1}, u^0_{0:t-1}$, define the FP process

$$\begin{aligned}\bar{\mu}^\tau_{t|x^0_{0:t-1},u^0_{0:t-1}} &= \frac{1}{\tau}\int_0^\tau \mu^{\pi^{BR,s}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}} ds \\ \bar{\mu}^{0,\tau}_t &= \frac{1}{\tau}\int_0^\tau \mu^{\pi^{0,BR,s}_{0:t-1}}_t ds\end{aligned} \tag{3.4.57}$$

where $\mu^{\pi^{BR,\tau}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}$ and $\mu^{\pi^{0,BR,\tau}_{0:t-1}}_t$ are conditional and joint distributions respectively, induced by the BR policies $\pi^{BR,\tau}$ and $\pi^{0,BR,\tau}$ up to time $t-1$ against $\mu^{\bar{\pi}^\tau}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x)$ and $\mu^{\bar{\pi}^{0,\tau}}_t(x^0_t, x^0_{0:t-1}, u^0_{0:t-1})$. In other words, $\pi^{BR,\tau} := \arg\min_\pi J(\pi, \bar{\pi}^\tau, \bar{\pi}^{0,\tau})$ and $\pi^{0,BR,\tau} := \arg\min_{\pi^0} J^0(\bar{\pi}^\tau, \pi^0)$.

Note that the distributions induced by the averaged policies $\{\bar{\pi}^\tau_t\}_{t\in\mathcal{T}}$ and $\{\bar{\pi}^{0,\tau}_t\}_{t\in\mathcal{T}}$ for $\tau \geq 1$ are given as

$$\begin{aligned}&\bar{\pi}^\tau_t(u|x, x^0_{0:t-1}, u^0_{0:t-1})\int_{s=0}^\tau \mu^{\pi^{BR,s}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x)ds \\ &\quad= \int_{s=0}^\tau \mu^{\pi^{BR,s}}_{t,|x^0_{0:t-1},u^0_{0:t-1}}(x)\pi^{BR,s}_t(u|x, x^0_{0:t-1}, u^0_{0:t-1})ds, \\ &\bar{\pi}^{0,\tau}_t(u^0|x^0, x^0_{0:t-1}, u^0_{0:t-1})\int_{s=0}^\tau \mu^{\pi^{0,BR,s}}_t(x^0, x^0_{0:t-1}, u^0_{0:t-1})ds \\ &\quad= \int_{s=0}^\tau \mu^{\pi^{0,BR,s}}_t(x^0, x^0_{0:t-1}, u^0_{0:t-1})\pi^{0,BR,s}_t(u^0|x^0, x^0_{0:t-1}, u^0_{0:t-1})ds,\end{aligned} \tag{3.4.58}$$

for all $t \in \mathcal{T}$ and $x^0_{0:t-1}, u^0_{0:t-1}$. For $s < 1$, $\pi^{BR,s}$ and $\pi^{0,BR,s}$ are chosen arbitrarily. The proof and the differential form of Eq. (3.4.57) and Eq. (3.4.58) can be found in Appendix D.1.

As a result, below we give a convergence analysis together with assumptions for continuous time FP, converging in both minor and major exploitability $\mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) = \max_{\pi'} J(\pi', \bar{\pi}^\tau, \bar{\pi}^{0,\tau}) - J(\bar{\pi}^\tau, \bar{\pi}^\tau, \bar{\pi}^{0,\tau})$, $\mathcal{E}^0(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) = \max_{\pi^{0\prime}} J^0(\bar{\pi}^\tau, \pi^{0\prime}) - J^0(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$, summarized as the total exploitability $\mathcal{E}_{\text{tot}}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) = \mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) + \mathcal{E}^0(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$.

Assumption 3.4.4. *1. The transition kernels are in the form of $p(x_{t+1} \mid x_t, u_t, x^0_t, u^0_t)$ and $p^0(x^0_{t+1} \mid x^0_t, u^0_t)$ for minor agents and major agent, respectively.*

2. The reward of minor and major agents are separable, i.e. for some reward functions $\tilde{r}, \bar{r}, \tilde{r}^0, \hat{r}^0, \check{r}^0$, we have

$$r(x, u, x^0, u^0, \mu) = \tilde{r}(x, x^0, u) + \bar{r}(x, x^0, \mu),$$
$$r^0(x^0, u^0, \mu) = \tilde{r}^0(x^0, u^0) + \bar{r}^0(x^0, \mu).$$

3. The game is monotone; i.e., satisfies Lasry-Lions monotonicity condition: For minor agents, we have $\forall x^0 \in \mathcal{X}^0$, $\forall \mu, \mu'$:

$$\sum_{x \in \mathcal{X}} (\mu(x) - \mu'(x))(\bar{r}(x, x^0, \mu) - \bar{r}(x, x^0, \mu')) \leq 0.$$

Meanwhile, for major agents, we have

$$\frac{d}{d\tau} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \cdot \left\langle \nabla_\mu \bar{r}^0(x^0_{t+1}, \mu^{\bar{\pi}^\tau_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}), \frac{d}{d\tau} \mu^{\bar{\pi}^\tau_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}} \right\rangle \leq 0.$$

4. We have $\tilde{\mathcal{E}}(\bar{\pi}^\tau, \pi^{0,BR,\tau}, \bar{\pi}^{0,\tau}) \leq \mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$, where we define $\tilde{\mathcal{E}}(\bar{\pi}^\tau, \pi^{0,BR,\tau}, \bar{\pi}^{0,\tau}) = J(\pi^{BR,\tau}, \bar{\pi}^\tau, \pi^{0,BR,\tau}) - J(\bar{\pi}^\tau, \bar{\pi}^\tau, \pi^{0,BR,\tau})$ with any BR policy given as $\pi^{BR,\tau} = \arg\max_\pi J(\pi, \bar{\pi}^\tau, \bar{\pi}^{0,\tau})$.

Remark 3.4.2. *Assumption 3.4.4.3. is fulfilled for major agents if $\bar{r}^0(x^0, \mu) = \bar{r}^0(x^0)$. Assumption 3.4.4.4. is satisfied for instance if $r(x, u, x^0, \mu) = r(x, u, \mu)$ and $p(x_{t+1} \mid x_t, u_t, x^0_t, u^0_t) = p(x_{t+1} \mid x_t, u_t)$. Then, we trivially have $\tilde{\mathcal{E}}(\bar{\pi}^\tau, \pi^{0,BR,\tau}, \bar{\pi}^{0,\tau}) = \mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$ by obtaining a minor agent MFG independent of the major agent.*

Theorem 3.4.2. *Under Assumption 3.4.4, the total exploitability is a strong Lyapunov function such that $\frac{d}{d\tau}\mathcal{E}_{\text{tot}}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) \leq -\frac{1}{\tau}\mathcal{E}_{\text{tot}}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$; i.e., we have $\mathcal{E}_{\text{tot}}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) = \mathcal{O}(1/\tau)$ in the continuous time FP algorithm.*

The proof of Theorem 3.4.2 can be found in Appendix D.1 and is based on a monotonic decrease of exploitability, at the same rate as standard FP in MFGs [127].

In numerical experiments, for applicability and computational tractability (due to the exponential complexity of the histories in the horizon), we condition policies on the random MF and major state instead of the histories, averaging policies uniformly instead of for each possible major state-action sequence. Further, numerically we partition and represent the (naturally continuous) MFs as described in the following, to obtain tabular Algorithm 1. Experimentally, in Section 3.4.4 we nonetheless find that the algorithm optimizes exploitability, even if Assumption 3.4.4 is not fully satisfied. The dependence of policy actions on the MF and major state has the additional advantage of allowing standard dynamic programming for major and minor MDPs, as their full MDP states include both the MF and major state.

Algorithm 1 Discrete-time, projected FP

1: Input: δ-partition $\{\mathcal{P}_i\}_{i=1\dots,M}$.
2: Initialize initial policies $\bar{\pi}_{(0)}$, $\bar{\pi}^0_{(0)}$.
3: **for** iteration $n = 0, 1, 2, \dots$ **do**
4: Compute discretized BR (as in Definition 3.4.5)

$$\pi_{(n+1)} \in \arg\max \hat{Q}_{\bar{\pi}_{(n)},\bar{\pi}^0_{(n)}},$$
$$\pi^0_{(n+1)} \in \arg\max \hat{Q}^0_{\bar{\pi}_{(n)},\bar{\pi}^0_{(n)}}.$$

5: Compute next average policies

$$\bar{\pi}_{(n+1)} \coloneqq \frac{n}{n+1}\bar{\pi}_{(n)} + \frac{1}{n+1}\pi_{(n+1)},$$
$$\bar{\pi}^0_{(n+1)} \coloneqq \frac{n}{n+1}\bar{\pi}^0_{(n)} + \frac{1}{n+1}\pi^0_{(n+1)}.$$

3.4.3.2 *Projected Mean Field*

Observe that for given current MF and major state-actions, we obtain deterministic transitions from one MF to the next. Therefore, by partitioning we can obtain deterministic transitions in-between parts of a partition of $\mathcal{P}(\mathcal{X})$, and a Bellman equation over *finite* spaces.

Definition 3.4.4. *A δ-partition $\mathcal{M} = \{\mathcal{P}_i\}_{i\in[|\mathcal{M}|]}$ is a partition of $\mathcal{P}(\mathcal{X})$, with $\|\mu - \nu\| < \delta$ for any $i \in [|\mathcal{M}|]$, $\mu, \nu \in \mathcal{P}_i$.*

Since $\mathcal{P}(\mathcal{X})$ is compact, a finite δ-partition of $\mathcal{P}(\mathcal{X})$ exists for any $\delta > 0$. We will henceforth assume for any $\delta > 0$ some δ-partition $\mathcal{M}$ of $\mathcal{P}(\mathcal{X})$ with $M = M(\delta)$ parts.

DISCRETIZED FINITE MDPS. To each part $\mathcal{P}_i$, we associate an arbitrary element $\hat{\mu}^{(i)} \in \mathcal{P}_i$ and write $\mathrm{proj}_\delta \mu$ for the δ-partition projection of MFs $\mu \in \mathcal{P}(\mathcal{X})$, i.e. whenever $\mu \in \mathcal{P}_i$ we project to the representative $\mathrm{proj}_\delta \mu = \hat{\mu}^{(i)} \in \mathcal{P}_i$.

As a result, we obtain discretized, *finite* MDP versions of the major and minor agent MDPs, where the continuous MF state is replaced by finitely many states in $\hat{\mathcal{P}}(\mathcal{X}) \coloneqq \{\hat{\mu}^{(1)}, \dots, \hat{\mu}^{(M)}\}$, evolving by discretized MF evolutions in Eq. (3.4.50), i.e. $\hat{\mu}_{t+1} = \mathrm{proj}_\delta T^\pi_t(x^0, u^0, \hat{\mu}_t)$ for any $x^0, u^0, \hat{\mu}_t$.

We can solve the discretized MDPs in a tabular manner: To compute best responses under policies (π, π^0), observe that the true action-value function Q^0_{π,π^0} of the (not discretized) major agent MDP follows the Bellman equation

$$Q^0_{\pi,\pi^0}(t, x^0, u^0, \mu) = r^0(x^0, u^0, \mu) + \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mu) \cdot \max_{u^{0\prime}} Q^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T^\pi_t(x^0, u^0, \mu)).$$

The tabular approximate action-value function $\hat{Q}^0_{\pi,\pi^0}$ for the major agent follows instead the Bellman equation of the discretized major agent MDP (letting the domain of $\hat{Q}_{\pi,\pi^0}$ be the entirety of $\mathcal{P}(\mathcal{X})$ as constants over each part $\mathcal{P}_i$),

$$\begin{aligned}\hat{Q}^0_{\pi,\pi^0}(t,x^0,u^0,\mu) &= \hat{Q}^0_{\pi,\pi^0}(t,x^0,u^0,\mathrm{proj}_\delta\mu)\\ &= r^0(x^0,u^0,\mathrm{proj}_\delta\mu) + \sum_{x^{0\prime}} p^0(x^{0\prime}\mid x^0,u^0,\mathrm{proj}_\delta\mu)\\ &\qquad \cdot \max_{u^{0\prime}} \hat{Q}^0_{\pi,\pi^0}(t+1,x^{0\prime},u^{0\prime},T^\pi_t(x^0,u^0,\mathrm{proj}_\delta\mu))\end{aligned}$$

with terminal condition zero, and the minor action-values analogously. The above can nonetheless provide a good approximation that can be computed in *tabular* form, see Appendix D.3 and empirical support in Section 3.4.4.

DISCRETIZED EQUILIBRIA. Building upon the preceding approximations, we define an approximate equilibrium as a fixed point of the discretized system.

Definition 3.4.5. *A δ-partition M3FNE is a tuple $(\pi,\pi^0)\in\hat{\Pi}\times\hat{\Pi}^0$ with $\pi\in\arg\max\hat{Q}_{\pi,\pi^0}$ and $\pi^0\in\arg\max\hat{Q}^0_{\pi,\pi^0}$ where policies in $\hat{\Pi},\hat{\Pi}^0$ are instead defined as blockwise constant over each part $\mathcal{P}_i$ of the δ-partition.*

Here, we understand $\hat{\pi}\in\arg\max\hat{Q}_{\pi,\pi^0}$ by the defining equation given as full support on optimal actions $\sum_{u\in\arg\max_{u'}\hat{Q}_{\pi,\pi^0}(t,x,u',x^0,\hat{\mu})}\hat{\pi}_t(x,x^0,\hat{\mu},u)=1$ for all $(t,x,x^0,\hat{\mu})\in\mathcal{T}\times\mathcal{X}\times\mathcal{X}^0\times\hat{\mathcal{P}}(\mathcal{X})$, and similarly for major agents, noting that $\hat{\pi}$ optimizes the preceding discretized finite MDP [69].

We note that while the discretized solutions only piecewise fulfill Assumption 3.4.3 by not being Lipschitz, in Section 3.4.4 we empirically find that the approximation of finite games and exploitability can nonetheless be accurate.

APPROXIMATION GUARANTEES. We evaluate solutions by tabular evaluation in the discretized MDP, for which we are able to obtain theoretical guarantees for evaluating the true exploitability via the approximate tabular exploitability. Under a δ-partition, define the major approximate objective

$$\hat{J}^0(\pi,\pi^0) \coloneqq \sum_{x^0}\mu^0_0(x^0)\hat{V}^{0,\pi^0}_{\pi,\pi^0}(0,x^0,\mu_0)$$

and approximate exploitability

$$\hat{\mathcal{E}}^0(\pi,\pi^0) \coloneqq \sum_{x^0}\mu^0_0(x^0)\cdot\Big(\max_{\hat{\pi}^{0\prime}\in\hat{\Pi}}\hat{V}^{0,\hat{\pi}^{0\prime}}_{\pi,\pi^0}(0,x^0,\mu_0)) - \hat{V}^{0,\pi^0}_{\pi,\pi^0}(0,x^0,\mu_0)\Big),$$

with approximate values $\hat{V}^{0,\hat{\pi}^0}_{\pi,\pi^0}$ of major deviation under (π,π^0) to $\hat{\pi}^0$, following the "discretized" Bellman equation

$$\begin{aligned}\hat{V}^{0,\hat{\pi}^0}_{\pi,\pi^0}(t,x^0,\mu) = \sum_{u^{0\prime}}\hat{\pi}^0_t(u^{0\prime}\mid x^{0\prime},\mathrm{proj}_\delta\mu)\\ \Big[r^0(x^0,u^0,\mathrm{proj}_\delta\mu) + \sum_{x^{0\prime}}p^0(x^{0\prime}\mid x^0,u^0,\mathrm{proj}_\delta\mu)\\ \hat{V}^{0,\hat{\pi}^0}_{\pi,\pi^0}(t+1,x^{0\prime},u^{0\prime},T^\pi_t(x^0,u^0,\mathrm{proj}_\delta\mu))\Big],\end{aligned}$$

and similarly for the minor agent. Note that only for the major agent, π^0 is irrelevant (replaced by $\hat{\pi}^0$). In other words, we approximate values and exploitability via the discretized finite MDPs, which has the advantage of enabling dynamic programming (backwards induction, value iteration).

By analyzing the value functions under continuity, we show in Appendix D.8 that these approximations are generally close to the true objectives and exploitabilities respectively, as the discretization becomes sufficiently fine.

Theorem 3.4.3. *Under Assumptions 3.4.1, 3.4.2 and 3.4.3, as $\delta \to 0$, approximate minor and major values tend to the exact values, and approximate exploitabilities tend to the exact exploitabilities, at rate $\mathcal{O}(\delta)$ uniformly over $(\pi, \pi^0) \in \Pi \times \Pi^0$.*

3.4.4 *Experiments*

We evaluate FP by comparing against FPI, which iterates discretized best response policies. For reproducibility, note that the algorithms used are deterministic, and details can be found in the appendix. For code, see https://github.com/tudkcui/M3FG-learning.

3.4.4.1 *Problems*

For the evaluation, we use the following problem instances for exemplary, practically applicable M3FG scenarios.

SIS EPIDEMICS CONTROL. The SIS problem is an epidemics control scenario, where each individualistic minor agent may decide whether to take costly preventative actions against becoming infected at a rate proportional to the proportion of infected. The major agent (e.g. government) is responsible for the well-being of minor agents, and can encourage preventative actions, while its state models random low- and high-infectivity seasons. The finite time horizon can be considered the time until a cure is found. The original problem without major agents has been used as a benchmark for MFG learning [9, 130].

BUFFET PROBLEM. In the Buffet problem, we consider the following scenario: At a conference with multiple buffet locations, agents desire to be at locations that are filled with food and uncrowded. However, each location depletes faster with increasing number of agents. The major agent (caterer) must keep buffets full and equally filled. The Buffet problem fulfills most assumptions (except Assumption 3.4.4.4.) and shows accordingly stable FP learning.

ADVERTISEMENT DUOPOLY MODEL. Lastly, in the advertisement model, a regulator sets the price of advertisement. Depending on the regulator's state and price of advertisement, two companies exogeneously decide on advertisement efficiencies of their subscription service. Minor agents are consumers and choose whether to change to subscriptions for the better-funded product, while the regulator avoids formation of a monopoly. Duopoly advertisement competition in a static MFG was modeled in [196].

3.4.4.2 *Numerical Results*

In the following, we provide a numerical evaluation via exploitability as the primary metric of interest, since it describes the quality of achieved equilibria. Additional experiments and parameter details are shown in Appendix D.9, including more qualitative results, the effect of alternative initializations, and analogous results for infinite-horizon discounted objectives. Beyond supporting the theoretical results, we also ablate both convergence assumptions for the algorithm and the Lipschitz policy assumption for propagation of chaos in the finite agent system.

EXPLOITABILITY CONVERGENCE. As observed in Figure 3.16, naive FPI usually fails to converge and runs into limit cycles, motivating FP. In Figure 3.17, we see that the proposed FP algorithm optimizes both approximate major and minor exploitabilities $\hat{\mathcal{E}}$, $\hat{\mathcal{E}}^0$ over its iterations. Especially for Buffet, which fulfills most of Assumption 3.4.4, learning is smooth and exploitability descends monotonically as in Theorem 3.4.2, while exploitability is nevertheless optimized in the other problems. Overall, the proposed FP algorithm improves achieved exploitabilities significantly over FPI.

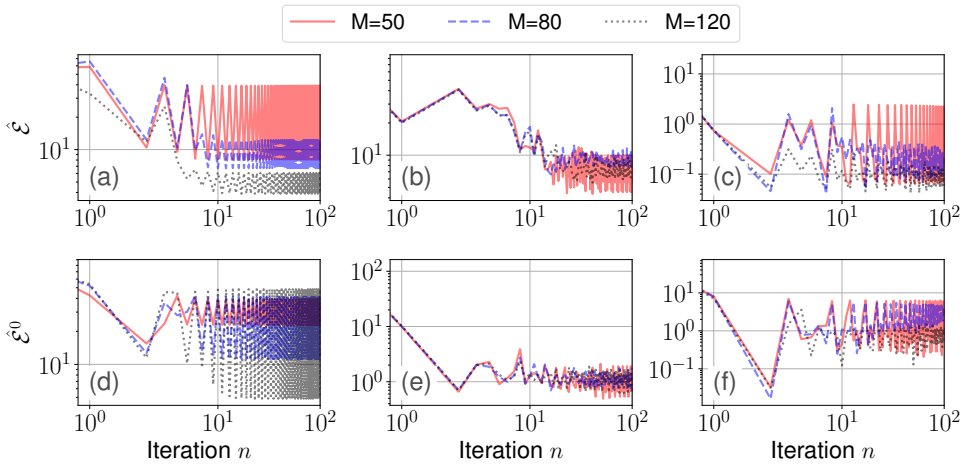


FIGURE 3.16: Non-convergence of exploitability in FPI. The approximate exploitability oscillates over iterations of FPI. (a, d): SIS, (b, e): Buffet, (c, f): Advertisement. (a-c): Minor exploitability, (d-f): major exploitability.

STABILITY OVER DISCRETIZATION. Comparing approximation results empirically over discretization bins M per dimension, i.e. using δ-partitions with $\delta \approx \frac{2}{M}$, in Figure 3.18 we observe that the FP-learned policies quickly stabilize as the discretization becomes sufficiently fine. The result supports not only the discretization approximation in Theorem 3.4.3, but also shows insensitivity of our FP algorithm to the fineness of the grid, as long as it is sufficiently fine to approximate the problem well. Hence, in the following we will use $M = 120$.

FINITE-AGENT CONVERGENCE. In Figure 3.19, the convergence of episodic returns by propagation of chaos is depicted as the number of agents $N \to \infty$. The limiting performance as the number of agents grows, quickly approaches the performance of the projected MF prediction, up to a small, negligible error from discretization and finite agents. The result supports propagation of chaos in Theorem 3.4.1 by convergence of the empirical objective to the limiting objective, despite the

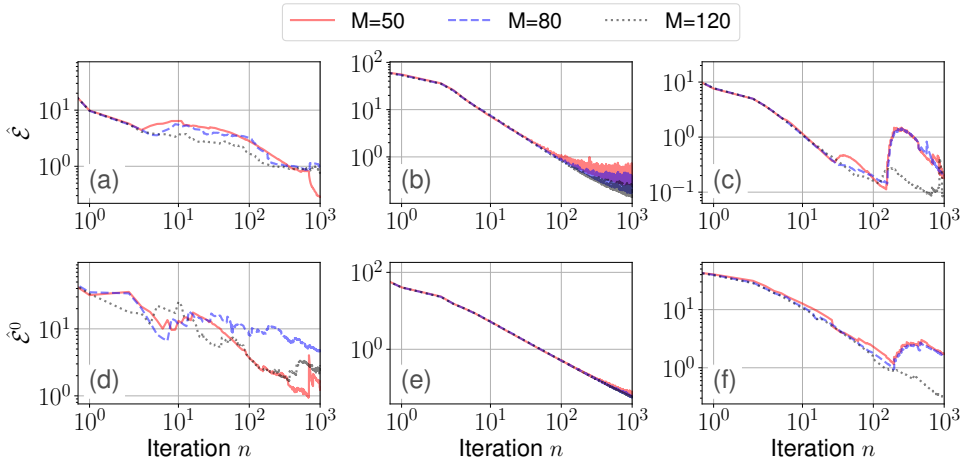


FIGURE 3.17: Convergence of exploitability in FP. The approximate exploitability is optimized via FP. (a, d): SIS, (b, e): Buffet, (c, f): Advertisement. (a-c): Minor exploitability, (d-f): major exploitability.

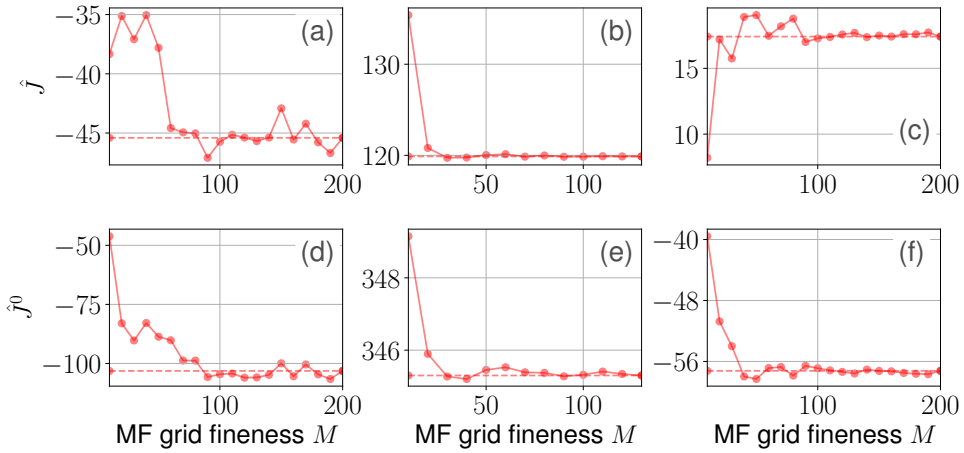


FIGURE 3.18: Stability of FP results under discretization. The final objectives of FP under discretization (dashed: right-most entry) are stable with high discretization. (a, d): SIS, (b, e): Buffet, (c, f): Advertisement. (a-c): Minor exploitability, (d-f): major exploitability.

non-Lipschitz projected MF policies. In Appendix D.9, similar results hold for (Lipschitz) uniform policies.

QUALITATIVE ANALYSIS. Lastly, we visualize the qualitative behavior obtained and find plausible equilibrium behavior, e.g., for the SIS problem. As seen in Figure 3.20, the equilibrium behavior plausibly reaches an equilibrium of infected agents, where the cost of actions equilibrates. The number of infected increases over time due to the finite horizon, discounting costs of infection beyond the horizon. Furthermore, minor agents take precautions only down to some infection threshold, at which point the expected cost of not taking precautions is higher. The major agent prevents infections in the low-infectivity regime ($x^0 = L$), while in the high-infectivity regime ($x^0 = H$) the high infection probability for minor agents already encourages preventative actions.

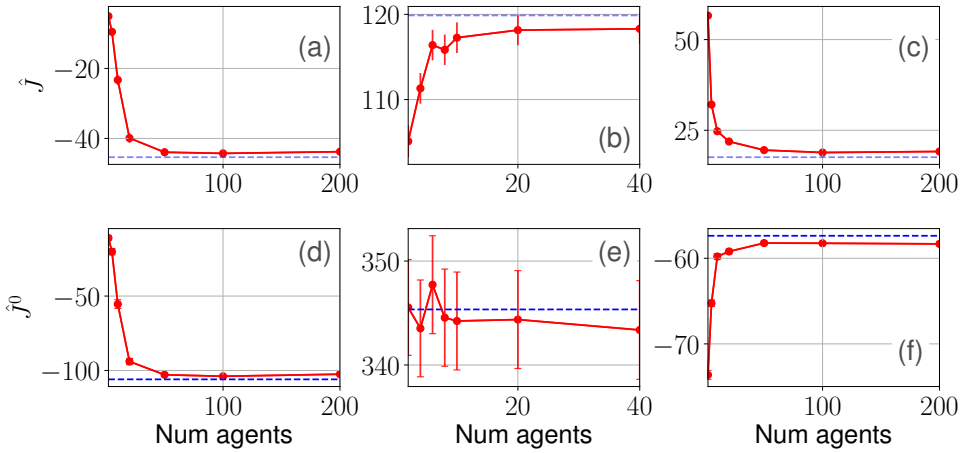


FIGURE 3.19: Convergence of finite objectives in the limit. The mean N-agent objective (red) over 1000 (or 5000 for Buffet) episodes, with 95% confidence interval, against MF predictions $\hat{J}$, $\hat{J}^0$ for FP and $M = 120$ (blue, dashed). (a, d): SIS, (b, e): Buffet, (c, f): Advertisement.

3.4.5 *Summary*

We have developed a new model and algorithm for a novel, broad class of tractable games. The framework allows scalable analysis of a large number of agents with theoretical guarantees. The proposed methods have been empirically supported through a variety of experiments. Still, for problems with multiple Nash equilibria, the FP algorithm finds only some equilibrium. Future work could address finding all or specific, e.g., socially-optimal equilibria. One could also try to relax theoretical assumptions. Lastly, since scalability of the discretization method remains an issue for larger minor state spaces, one may consider deep RL methods.

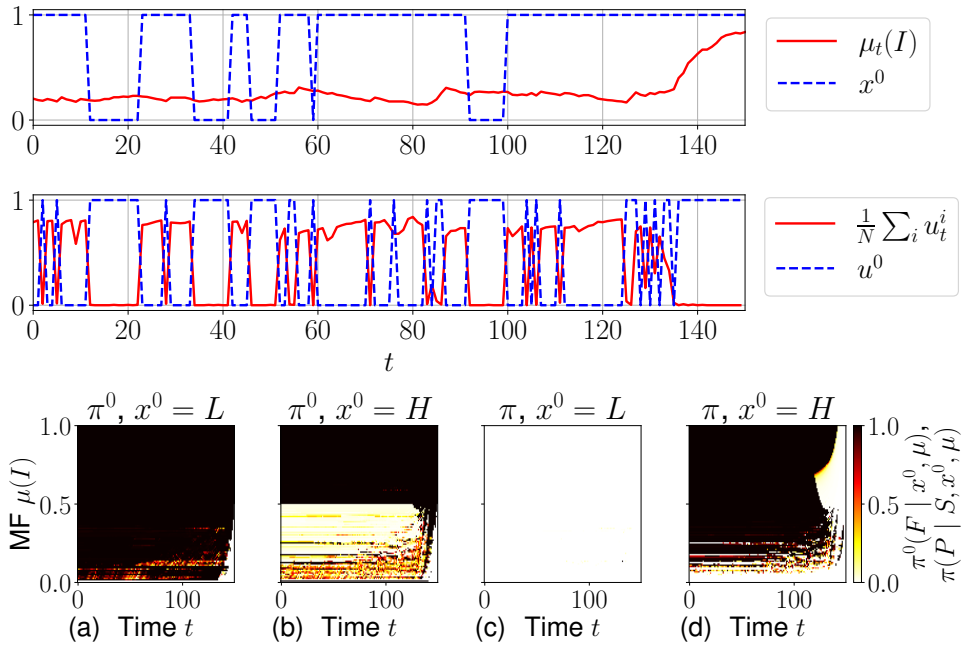


FIGURE 3.20: Example visualization of FP results in SIS. The FP-learned M3FNE in SIS, for $M = 120$. Top: example trajectory (for visualization, $L = \bar{P} = \bar{F} = 0, H = P = F = 1$, see Appendix D.9); bottom: policy.

3.5 CONCLUSION OF CHAPTER 3

In this chapter, we have first introduced methods for solving general evolutive MFGs. We have shown that basic FPI is not sufficient even in the simplest case of finite MFGs. We have thus proposed and analyzed entropy regularization as a solution to convergence, trading off with the sub-optimality of agent policies. Methodologically, we iterated relative entropy regularization via the prior descent method to further improve results, and enabled analysis of more large-scale MFGs through the use of deep RL and particle filters.

We then moved on to extend MFGs towards more general settings. First, the all-to-all interaction of agents was ameliorated by integration of graph structures, through the limit theory of large graphs and graphons. The resulting GMFG model was analyzed and algorithmically reduced to standard MFGs with approximation guarantees. We also extended the graphical setting beyond pairwise interaction, by considering multilayer hypergraphs as a generalization, where agents may have "hyperedges" between multiple other agents. The previously developed regularization approach was extended to these generalized settings under discretization, with approximate optimality guarantees under standard Lipschitz assumptions.

Finally, we addressed strong interaction in the presence of many minor agents and a single major player, through the framework of M3FGs. The added difficulty through the stochasticity of the MF was addressed by considering general policies conditioned on the current MF. Some approximation properties of the discretization for evaluation of exploitability were shown. We also analyzed and proposed a practical FP algorithm by discretizing the space of MFs and simplified weighting of best responses. The applicability of the developed framework was verified on various examples with finite and infinite time horizons.

Overall, we have addressed our RQs I and II of learning and model generality, in order to increase the applicability of MFGs in the competitive case. In this chapter, we have shown how to generalize MFG models and algorithms to more practical and general models, and we hope that it shows the reader how to perform their own generalizations for particular scenarios. Alternatively, while general algorithms still remain to be considered for solving MFGs under no additional assumptions such as monotonicity or regularization, we hope that our contributions have extended the usefulness of MFGs in practice by further extending the range of problems that can be modelled by MFGs and allowing direct off-the-shelf usage of our developed frameworks. In the next chapter, we will consider the cooperative case of MFC instead, which is also of great practical relevance.

4

COOPERATIVE MEAN FIELD CONTROL

In this chapter, we study the cooperative case of MFC for tractable large-scale MARL in the presence of many agents. We begin our exposition with a special case of near-static MFC with evolving environment states and static MF. There, a first result is shown on the near-optimality of MFC solutions over heterogeneous policies in the finite control problem. The framework is demonstrated using PPO. We then move on to more general MFC, which is extended to major agents similarly as in the competitive MFG case, in order to increase the generality and flexibility of MFC. Basic theoretical properties are shown. There, we also propose and analyze MFC-based MARL on the finite MARL problem instead of the limiting MFC problem. Finally, we tackle the challenge of partial observability, which is especially important in practical large-scale systems. Theory and algorithms are extended to this case, which allows for solving hard Dec-POMDPs through the MFC approach. The analysis is performed over equi-Lipschitz classes of policies, which are shown to allow stationary near-optimal solutions.

4.1 STATIC MEAN FIELD CONTROL

MARL methods have shown remarkable potential in solving complex multi-agent problems but mostly lack theoretical guarantees. In particular, their scalability to many agents can be limited due to the combinatorial nature of MARL [40]. Recently, MFC and MFGs have been established as a tractable solution for large-scale multi-agent problems with many agents. In this work, driven by a motivating load balancing problem, we consider a discrete-time MFC model with common environment states. We rigorously establish approximate optimality as the number of agents grows in the finite agent case and find that a DPP holds, resulting in the existence of an optimal stationary policy. As exact solutions are difficult in general due to the resulting continuous action space of the limiting MFC MDP, we apply established deep RL methods to solve the associated MFC problem. The performance of the learned MFC policy is compared to typical MARL approaches and is found to converge to the MF performance for sufficiently many agents, verifying the obtained theoretical results and reaching competitive solutions. The material presented in this section is based on [8].

MF theory applied to the cooperative setting is known as MFC, where one assumes that many agents cooperate to achieve Pareto optima [124, 197]. MFC has various applications e.g. in smart heating [53] or portfolio management [198]. The dimensions of the MFC problem are independent of the specific number of agents, making it more tractable. However, solving the MFC problem has the challenge of time-inconsistency due to the non-Markovian nature of the problem [197, 199, 200]. A recent way of handling this inherent time-inconsistency problem is to use an enlarged state-action space [103, 104, 121, 201]. We similarly apply this technique by lifting up the state-action space into its probability measure space, since it will enable usage of established RL methods. In this section we extend the theory of discrete-time MFC by considering additional environment states. An advantage of discrete-time models is applicability of a plethora of RL solutions. Our model can be considered a special case of the MFC equivalent of M3FG [113, 202] in Section 3.4 with trivial major agent policy, which had not been formulated yet. While the model in this section uses a simplifying assumption of trivial agent dynamics, it allows us to directly show approximate optimality of MFC over any heterogeneous tuple of policies in the finite system. We have since extended the consideration of environment states via the Major-Minor Mean Field Control (M3FC) framework in Section 4.2.

OUR CONTRIBUTION. The main contributions of this work are: (i) We propose a new discrete-time MFC formulation that transforms large-scale multi-agent control problems with common environment states into a simple MDP with lifted state-action space; (ii) we rigorously show approximate optimality for sufficiently large systems as well as existence of an optimal stationary policy through a DPP, and (iii) associated with this standard discrete-time MDP with continuous action space, we verify our theoretical findings empirically using modern RL techniques. As a result, we outperform existing baselines for the many-agent case and obtain a methodology to solve large multi-agent control problems such as the following.

4.1.1 *A Motivating Load Balancing Scenario*

While the concept of MF limits has been used in queuing systems before, it has mostly been used for the state of the buffer fillings of queues or the number of servers/queues [57, 203]. In this work we use MFs to represent the state of a large amount of schedulers while modeling the queues exactly. See also Figure 4.1 for a visualization of the problem. Note that in principle, our model could be

used for any similar resource allocation problem such as allocation of many firefighters to houses on fire.

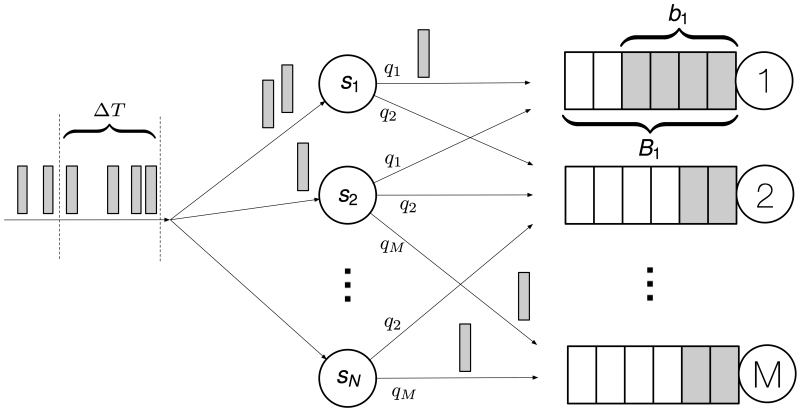


FIGURE 4.1: Overview of the queuing system. Many schedulers (middle) obtain packets at a fixed rate (left) that must be assigned to one of the accessible queues (right) such that total packet drops are minimized.

Consider a queuing system with N agents called schedulers, $[s_1, \dots, s_N]$, and M parallel servers, each with its own finite FIFO queue. Denote the queue filling by $b_i \in \{0, \dots, B_i\}$, $i = 1, \dots, M$ where B_i is the maximum buffer space for the i-th queue. At any time step t, the state $x_t^{i,N} \in \mathcal{X}$ of a scheduler is the set of queues it has access to. The agent state space $\mathcal{X}$ therefore consists of all combinations of queue access where every agent has access to at least one of the queues. The environment state is the current buffer filling $x^0 = [b_1, \dots, b_M]$, where b_j is the buffer filling of queue j.

In discrete-time, the number of job arrivals to be assigned at each time step t is Poisson distributed with rate $\lambda \Delta T$ and the number of serviced jobs for each server is Poisson distributed with rate $\beta \Delta T$, where $\Delta T > 0$ can be considered the time span between each synchronization of schedulers. As an approximation, we assume that all queue departures in a time slot happen before the new arrivals, and newly arrived jobs thus cannot be serviced in the same time slot.

We split the total number of job packets which arrive in some time step ΔT uniformly at random amongst the schedulers. The jobs assigned to each scheduler need to be sent out immediately. Each scheduler decides which of the accessible queues it sends its arrived jobs to during each time step. If a job is mapped to a full buffer, it is lost and a penalty c_d is incurred. The goal of the system is therefore to minimize the number of job drops. At each step of the decision making, we assume that the state of the environment x^0 and their own accessible queues are known to the schedulers.

We can model the dynamics of the environment state dependent on the empirical state-action distribution of all schedulers: Consider agents choosing some choice of queues as their action, where inaccessible queues are treated as randomly picking a destination. In that case, to assign a packet to its destination queue, it is clearly sufficient to consider the empirical distribution: Sampling from the empirical distribution, using the sampled action and, if inaccessible, resampling an accessible queue provides the desired behavior.

4.1.2 *Static Mean Field Control with Major States*

In this section, we formulate a N-agent model that in the limit of $N \to \infty$ results in a more tractable MFC problem. Importantly, we will then show approximate optimality and a DPP for the MFC problem, allowing for application of RL.

Notation. *Let $\mathcal{A}$ be a finite set. We equip $\mathcal{A}$ with the discrete metric and denote the set of real-valued functions on $\mathcal{A}$ by $\mathbb{R}^{\mathcal{A}}$, For $f \in \mathbb{R}^{\mathcal{A}}$ let $\|f\|_\infty = \max_{a\in\mathcal{A}} f(a)$. Denote by $|\mathcal{A}|$ the cardinality of $\mathcal{A}$. Denote by $\mathcal{P}(\mathcal{A}) = \{p \in \mathbb{R}^{\mathcal{A}} \colon p(a) \geq 0, \sum_{a\in\mathcal{A}} p(a) = 1\}$ the space of probability simplices, equivalent to the probability measures on $\mathcal{A}$. Equip $\mathcal{P}(\mathcal{A})$ with the l_1-norm $\|\mu - \nu\|_1 = \sum_{a\in\mathcal{A}} |\mu(a) - \nu(a)|$. For readability, we uncurry occurrences of multiple parentheses, e.g. $\pi_t(x_t^0)(x) \equiv \pi_t(x_t^0, x)$. Define $\mu(f) \coloneqq \sum_{a\in\mathcal{A}} f(a)\mu(a)$ for any $\mu \in \mathcal{P}(\mathcal{A})$, $f \colon \mathcal{A} \to \mathbb{R}$.*

4.1.2.1 *Finite Agent Model*

Let $\mathcal{X}$, $\mathcal{U}$ be a finite state and action space respectively. Let $\mathcal{X}^0$ be a finite environment state space. For any $N \in \mathbb{N}$, at each time $t = 0, 1, \ldots$, the states and actions of agent $i = 1, \ldots, N$ are random variables denoted by $x_t^{i,N} \in \mathcal{X}$ and $u_t^{i,N} \in \mathcal{U}$. Analogously, the environment state is a random variable denoted by $x_t^{0,N} \in \mathcal{X}^0$. Define the empirical state-action distribution $\mathbb{G}_t^N = \frac{1}{N}\sum_{i=1}^N \delta_{(x_t^{i,N}, u_t^{i,N})} \in \mathcal{P}(\mathcal{X}\times\mathcal{U})$. For each agent i, we consider locally Markovian policies $\pi^i = \{\pi_t^i\}_{t\geq 0} \in \Pi_N$ from the space of admissible Markov policies Π_N where $\pi_t^i \colon \mathcal{X}^0 \times \mathcal{X} \to \mathcal{P}(\mathcal{U})$. Further, we define the policy profile $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^N) \in \Pi_N^N$.

Acting only on local and environment information may seem like a strong restriction. However, other agent states are uninformative under continuity assumptions as $N \to \infty$ as the interaction between agents will be restricted to the increasingly deterministic empirical state-action distribution.

Let $\mu_0 \in \mathcal{P}(\mathcal{X})$ be the initial agent state distribution, $\mu_0^0 \in \mathcal{P}(\mathcal{X}^0)$ the initial environment state distribution and $p^0 \colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}\times\mathcal{U}) \to \mathcal{P}(\mathcal{X}^0)$ a transition kernel. The random variables shall follow $x_0^{0,N} \sim \mu_0^0$ and subsequently

$$x_t^{i,N} \sim \mu_0(x_t^{i,N}), \tag{4.1.1}$$

$$u_t^{i,N} \sim \pi_t^i(x_t^{0,N}, x_t^{i,N}), \tag{4.1.2}$$

$$x_{t+1}^{0,N} \sim p^0(x_t^{0,N}, \mathbb{G}_t^N), \tag{4.1.3}$$

where for simplicity of further analysis, the agent states are always sampled according to μ_0.

Remark 4.1.1. *While the above is a strong dynamics assumption, our formulation is nonetheless sufficient for the load balancing problem. In principle, any results should similarly hold under appropriate assumptions for nontrivial agent state dynamics by considering MF and environment state together. As this will significantly complicate analysis, an according extension of theoretical results is left to future works.*

Let us introduce another notation. First, define the space of decision rules $\mathcal{H} \coloneqq \{h \colon \mathcal{X} \to \mathcal{P}(\mathcal{U})\}$. Then a one-step policy profile $\boldsymbol{h} = (h^1, \ldots, h^N) \in \mathcal{H}^N$ is an N-fold decision rule. Our major example of a one-step policy profile is $(\pi_t^1(x^0), \ldots, \pi_t^N(x^0))$ for fixed $t \geq 0$, fixed $x^0 \in \mathcal{X}^0$ and potentially different policies for the N agents. For given agent state distribution μ_0 and a one-step policy profile $\boldsymbol{h} \in \mathcal{H}^N$ let $x^{i,N} \sim \mu_0, u^{i,N} \sim h^i(x^{i,N})$, s.t. $(x^{i,N}, u^{i,N})_{i=1,\ldots,N}$ are independent.

Then, consider a random measure $\mathbb{G}_{\boldsymbol{h}}^N \in \mathcal{P}(\mathcal{X} \times \mathcal{U})$ or equivalently its random probability mass function $\mathbb{G}_{\boldsymbol{h}}^N \colon \mathcal{X} \times \mathcal{U} \to [0,1]$,

$$\mathbb{G}_{\boldsymbol{h}}^N(x,u) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{x,u}(x^{i,N}, u^{i,N}) \,. \tag{4.1.4}$$

Define $\mathcal{G}^N(\mu_0, \boldsymbol{h})$ as the distribution of $\mathbb{G}_{\boldsymbol{h}}^N$, so $\mathcal{G}^N(\mu_0, \boldsymbol{h})$ is a distribution over the set $\mathcal{P}(\mathcal{X} \times \mathcal{U})$ and $\mathbb{G}_{\boldsymbol{h}}^N \sim \mathcal{G}^N(\mu_0, \boldsymbol{h})$. Consider the primary example $\boldsymbol{h} = (\pi_t^1(x^0), \dots, \pi_t^N(x^0))$. In contrast to the empirical distribution $\mathbb{G}_t^N$ that depends on a random $x_t^{0,N}$, the random probability mass function $\mathbb{G}_{\boldsymbol{h}}^N$ has $x_t^{0,N} = x^0$ fixed. By $\mathbb{E}[\mathbb{G}_{\boldsymbol{h}}^N]$ we denote in the following the entry-wise expectation $\{\mathbb{E}[\mathbb{G}_{\boldsymbol{h}}^N(x,u)]\}_{(x,u)\in\mathcal{X}\times\mathcal{U}}$.

Let $\gamma \in (0,1)$ be the discount factor and $r \colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X} \times \mathcal{U}) \to \mathbb{R}$ a reward function. The goal is to maximise the discounted accumulated reward

$$J^N(\boldsymbol{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^{0,N}, \mathbb{G}_t^N)\right] \tag{4.1.5}$$

which generalizes optimizing an average per-agent reward

$$J^N(\boldsymbol{\pi}) = \sum_{i=1}^{N} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \tilde{r}(x_t^{i,N}, x_t^{0,N}, \mathbb{G}_t^N)\right] \tag{4.1.6}$$

for some shared $\tilde{r} \colon \mathcal{X} \times \mathcal{X}^0 \times \mathcal{P}(\mathcal{X} \times \mathcal{U}) \to \mathbb{R}$ through choosing the reward function as $r(x_t^{0,N}, \mathbb{G}_t^N) \equiv \sum_{x\in\mathcal{X}} \tilde{r}(x, x_t^{0,N}, \mathbb{G}_t^N) \sum_{u\in\mathcal{U}} \mathbb{G}_t^N(x,u)$.

As the optimality concept in this work, we therefore define approximate Pareto optimality.

Definition 4.1.1 (Pareto optimality)**.** *For $\varepsilon > 0$, $\boldsymbol{\pi}^\varepsilon \in \Pi_N^N$ is ε-Pareto optimal if and only if*

$$J^N(\boldsymbol{\pi}^\varepsilon) \geq \sup_{\boldsymbol{\pi}} J^N(\boldsymbol{\pi}) - \varepsilon \,. \tag{4.1.7}$$

A visualization of this model can be found in Figure 4.2.

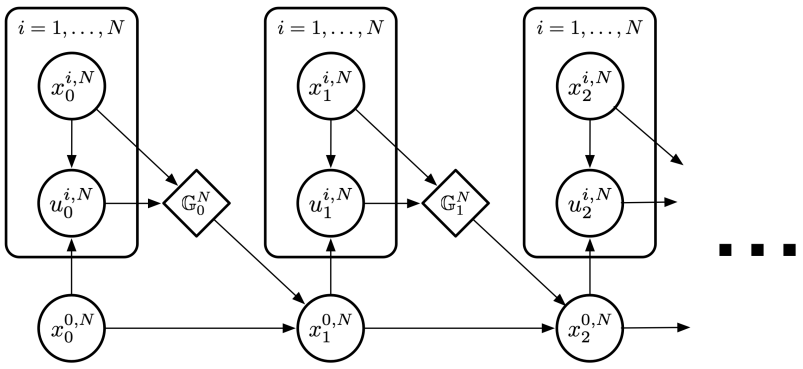


FIGURE 4.2: Overview of the multi-agent system as a probabilistic graphical model using plate notation [204], where circles and diamonds indicate stochastic and deterministic nodes respectively. Each agent i chooses an action $u_t^{i,N}$ conditional on the environment state $x_t^{0,N}$ and local agent state $x_t^{i,N}$, influencing the next environment state $x_{t+1}^{0,N}$ only via their empirical distribution $\mathbb{G}_t^N$. Agent states are assumed i.i.d. for simplicity of analysis.

4.1.2.2 *Mean Field Model*

As $N \to \infty$, we formally obtain the following MFC MDP, which will be rigorously justified in the sequel. At each time $t = 0, 1, \ldots$, the environment state is a random variable denoted by $x_t^0 \in \mathcal{X}^0$. We consider Markovian upper-level policies $\pi = \{\pi_t\}_{t \geq 0} \in \Pi$ from the space of such policies Π where $\pi_t \colon \mathcal{X}^0 \to \mathcal{H}$. We equip both $\mathcal{H}$ and Π with the supremum metric. As mentioned, the population state distribution is fixed to $\mu_0 \in \mathcal{P}(\mathcal{X})$ at all times. The random state-action distribution is therefore given by

$$\mathbb{G}_t \coloneqq \mathbb{G}(\mu_0, \pi_t(x_t^0)) \tag{4.1.8}$$

where $\mathbb{G} \colon \mathcal{P}(\mathcal{X}) \times \mathcal{H} \to \mathcal{P}(\mathcal{X} \times \mathcal{U})$ is defined by

$$\mathbb{G}(\mu, h)(x, u) \coloneqq h(x, u)\mu(x) \tag{4.1.9}$$

for any $x \in \mathcal{X}, u \in \mathcal{U}$. The random environment state variables therefore follow $x_0^0 \sim \mu_0^0$ and subsequently

$$x_{t+1}^0 \sim p^0(x_t^0, \mathbb{G}_t)\,. \tag{4.1.10}$$

Analogously, the objective becomes

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^0, \mathbb{G}_t)\right]\,. \tag{4.1.11}$$

We require the following simple continuity assumption to obtain meaningful results in the limit as $N \to \infty$.

Assumption 4.1.1 (Continuity of r and p^0)**.** *The functions r and p^0 are continuous, i.e. for all $x^0 \in \mathcal{X}^0$ and $\mathbb{G}_n \to \mathbb{G} \in \mathcal{P}(\mathcal{X} \times \mathcal{U})$ we have*

$$r(x^0, \mathbb{G}_n) \to r(x^0, \mathbb{G}), \quad p^0(x^0, \mathbb{G}_n) \to p^0(x^0, \mathbb{G})\,. \tag{4.1.12}$$

By compactness of $\mathcal{P}(\mathcal{X} \times \mathcal{U})$, we have boundedness.

Proposition 4.1.1. *Under Assumption 4.1.1, r is bounded by some R, i.e. for any $x^0 \in \mathcal{X}^0$, $\mathbb{G} \in \mathcal{P}(\mathcal{X} \times \mathcal{U})$ we have*

$$|r(x^0, \mathbb{G})| \leq R\,. \tag{4.1.13}$$

Our first goal will be to show that as $N \to \infty$, the optimal solution to the MFC is approximately Pareto optimal in the finite N case. This will motivate solving the MFC problem.

4.1.3 *Approximate Optimality under Heterogeneous Policy Tuples*

We first show the following lemma on uniform convergence in probability of empirical state-action distributions to their state-action-wise average for fixed one-step policy profiles.

Lemma 4.1.1. *Let $x^0 \in \mathcal{X}^0$ and $\boldsymbol{h} \in \mathcal{H}^N$ be an arbitrary one-step policy profile. Let $\mathbb{G}^N \sim \mathcal{G}(\mu_0, \boldsymbol{h})$. Then*

(i) $\mathbb{E}\left[\|\mathbb{G}^N - \mathbb{E}[\mathbb{G}^N]\|_1^2\right] \leq \frac{|\mathcal{X}|^2|\mathcal{U}|^2}{4N}$

(ii) $\mathbb{P}\left(\|\mathbb{G}^N - \mathbb{E}[\mathbb{G}^N]\|_1 \geq \varepsilon\right) \leq \frac{|\mathcal{X}|^2|\mathcal{U}|^2}{4\varepsilon^2 N}$

Proof. By Chebyshev's inequality, (i) implies (ii). It remains to prove (i). Let $x^{i,N} \sim \mu_0$ be i.i.d. and $u^{i,N} \sim \pi^i(x^0, x^{i,N})$, s.t. $(x^{i,N}, u^{i,N})_{i=1,\dots,N}$ are independent. Then by the sub-additivity of $\mathbb{E}[(\cdot)^2]^{\frac{1}{2}}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbb{G}^N - \mathbb{E}[\mathbb{G}^N]\|_1^2\right]^{\frac{1}{2}} \\
&\quad= \mathbb{E}\left[\left(\sum_{x\in\mathcal{X},u\in\mathcal{U}}\left|\frac{1}{N}\sum_{i=1}^N \mathbf{1}_{x,u}(x^{i,N},u^{i,N}) - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \mathbf{1}_{x,u}(x^{i,N},u^{i,N})\right]\right|\right)^2\right]^{\frac{1}{2}} \\
&\quad\leq \sum_{x\in\mathcal{X},u\in\mathcal{U}}\left(\mathbb{V}\left[\frac{1}{N}\sum_{i=1}^N \mathbf{1}_{x,u}(x^{i,N},u^{i,N})\right]\right)^{\frac{1}{2}} \\
&\quad= \sum_{x\in\mathcal{X},u\in\mathcal{U}}\left(\frac{1}{N^2}\sum_{i=1}^N \mathbb{V}\left[\mathbf{1}_{x,u}(x^{i,N},u^{i,N})\right]\right)^{\frac{1}{2}} \\
&\quad\leq \sum_{x\in\mathcal{X},u\in\mathcal{U}}\left(\frac{1}{N^2}\sum_{i=1}^N \frac{1}{4}\right)^{\frac{1}{2}} = \frac{|\mathcal{X}||\mathcal{U}|}{2\sqrt{N}}
\end{aligned}
$$

using the trivial variance bound $\frac{1}{4}$ for indicator functions. □

To achieve approximate optimality of MF solutions in the N-agent case, we first define how to obtain an N-agent policy $\boldsymbol{\pi}^N \in \Pi_N^N$ from a MF policy $\hat{\pi} \in \Pi$ by

$$\boldsymbol{\pi}^N(\hat{\pi}) = (\pi^1, \dots, \pi^N) \text{ with } \pi_t^i(x^0, x) = \hat{\pi}_t(x^0)(x)$$

for all $i = 1, \dots, N$, i.e. all agents with state $x \in \mathcal{X}$ will follow the action distribution $\hat{\pi}_t(x_t^0)(x)$ at times $t \geq 0$.

Theorem 4.1.1. *Under Assumption 4.1.1, we have uniform convergence of the N-agent objective to the MF objective as $N \to \infty$, i.e.*

$$\lim_{N\to\infty}\sup_{\pi\in\Pi}\left|J^N(\boldsymbol{\pi}^N(\pi)) - J(\pi)\right| = 0\,. \tag{4.1.14}$$

Proof. We have by definition

$$\sup_{\pi\in\Pi}\left|J^N(\boldsymbol{\pi}^N(\pi)) - J(\pi)\right| \tag{4.1.15}$$

$$= \sup_{\pi\in\Pi}\left|\sum_{t=0}^{\infty}\gamma^t\mathbb{E}\left[r(x_t^{0,N},\mathbb{G}_t^N) - r(x_t^0,\mathbb{G}_t)\right]\right| \tag{4.1.16}$$

$$\leq \sum_{t=0}^{\infty}\gamma^t\sup_{\pi\in\Pi}\left|\mathbb{E}\left[r(x_t^{0,N},\mathbb{G}_t^N) - r(x_t^0,\mathbb{G}_t)\right]\right|. \tag{4.1.17}$$

To obtain the desired result, we first show for any $t \geq 0$ that $\sup_{\pi\in\Pi}\|\mathcal{L}(x_t^{0,N}) - \mathcal{L}(x_t^0)\|_1 \to 0$ implies $\mathcal{L}(x_t^{0,N},\mathbb{G}_t^N) \to \mathcal{L}(x_t^0,\mathbb{G}_t)$ weakly uniformly over all $\pi \in \Pi$. Note that $\sup_{\pi\in\Pi}\|\mathcal{L}(x_t^{0,N}) - \mathcal{L}(x_t^0)\|_1 \to 0$ by definition implies

$$\sup_{\pi\in\Pi}\left|\mathcal{L}(x_t^{0,N})(x^0) - \mathcal{L}(x_t^0)(x^0)\right| \to 0$$

for any $x^0 \in \mathcal{X}^0$. For the joint law, consider any $f\colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}\times\mathcal{U}) \to \mathbb{R}$, continuous and bounded by $|f| \leq F$. Then

$$\begin{aligned}
&\sup_{\pi\in\Pi}\left|\mathcal{L}(x_t^{0,N},\mathbb{G}_t^N)(f) - \mathcal{L}(x_t^0,\mathbb{G}_t)(f)\right| \\
&\quad= \sup_{\pi\in\Pi}\left|\mathbb{E}\left[f(x_t^{0,N},\mathbb{G}_t^N)\right] - \mathbb{E}\left[f(x_t^0,\mathbb{G}_t)\right]\right| \\
&\quad\leq \sup_{\pi\in\Pi}\sum_{x^0\in\mathcal{X}^0}\left|\mathbb{E}\left[f(x_t^{0,N},\mathbb{G}_t^N)\;\middle|\; x_t^{0,N}=x^0\right]\mathcal{L}(x_t^{0,N})(x) - f(x^0,\mathbb{G}(\mu_0,\pi_t(x^0))\,\mathcal{L}(x_t^0)(x^0)\right| \\
&\quad\leq \sum_{x^0\in\mathcal{X}^0}\sup_{\pi\in\Pi}\left|\mathbb{E}\left[f(x_t^{0,N},\mathbb{G}_t^N)\;\middle|\; x_t^{0,N}=x^0\right]\right|\sup_{\pi\in\Pi}\left|\mathcal{L}(x_t^{0,N})(x^0) - \mathcal{L}(x_t^0)(x^0)\right| \\
&\qquad+ \sum_{x^0\in\mathcal{X}^0}\sup_{\pi\in\Pi}\left|f(x^0,\mathbb{G}(\mu_0,\pi_t(x^0)) - \mathbb{E}\left[f(x_t^{0,N},\mathbb{G}_t^N)\;\middle|\; x_t^{0,N}=x^0\right]\right|\sup_{\pi\in\Pi}\mathcal{L}(x_t^0)(x^0),
\end{aligned}$$

where the first sum goes to zero by assumption and boundedness of f. For the second term, consider arbitrary fixed $x^0 \in \mathcal{X}^0$. Write $\mathbb{G}_\pi$ short for $\mathbb{G}(\mu_0,\pi_t(x^0))$ and introduce $\mathbb{G}_\pi^N \sim \mathcal{G}(\mu_0,(\pi_t(x^0),\dots,\pi_t(x^0)))$ for all $N,\pi\in\Pi$. So in contrast to $\mathbb{G}_t^N$ that depends on a random $x_t^{0,N}$, the random probability mass function $\mathbb{G}_\pi^N$ has $x_t^{0,N} = x^0$ fixed. Then

$$\begin{aligned}
&f(x^0,\mathbb{G}(\mu_0,\pi_t(x^0))) - \mathbb{E}\left[f(x_t^{0,N},\mathbb{G}_t^N)\;\middle|\; x_t^{0,N}=x^0\right] \\
&\quad= f(x^0,\mathbb{G}_\pi) - \mathbb{E}\left[f(x^0,\mathbb{G}_\pi^N)\right]
\end{aligned}$$

We observe that for any $(x,u)\in\mathcal{X}\times\mathcal{U}$

$$\mathbb{E}[\mathbb{G}_\pi^N(x,u)] = \mathbb{G}_\pi(x,u). \tag{4.1.18}$$

For this purpose, let $x^{i,N} \sim \mu_0$ be i.i.d. and $u^{i,N} \sim \pi_t(x^0,x^{i,N})$, s.t. $(x^{i,N},u^{i,N})_{i=1,\dots,N}$ are independent. Then for any $(x,u)\in\mathcal{X}\times\mathcal{U}$ we have

$$\begin{aligned}
\mathbb{E}[\mathbb{G}_\pi^N(x,u)] &= \frac{1}{N}\sum_{i=1}^N\mathbb{E}\left[\mathbf{1}_{x,u}(x^{i,N},u^{i,N})\right] \\
&= \frac{1}{N}\sum_{i=1}^N\mu_0(x)\pi_t(x^0,x,u)
\end{aligned}$$

$$= \mathbb{G}_\pi(x,u)\,.$$

Let $\varepsilon > 0$ arbitrary. By compactness of $\mathcal{P}(\mathcal{X}\times\mathcal{U})$, the function $f(x^0,\cdot)\colon \mathcal{P}(\mathcal{X}\times\mathcal{U}) \to \mathbb{R}$ is uniformly continuous. Consequently, there exists $\delta > 0$ such that for all $\pi \in \Pi$

$$\begin{aligned}\|\mathbb{G}_\pi - \mathbb{G}_\pi^N\|_1 < \delta \\ \implies \left|f(x^0,\mathbb{G}_\pi) - f(x^0,\mathbb{G}_\pi^N)\right| < \frac{\varepsilon}{2}\,.\end{aligned}$$

By Lemma 4.1.1 (ii) and Eq. (4.1.18) there exists $N' \in \mathbb{N}$ such that for $N > N'$ and for all $\pi \in \Pi$ we have

$$\mathbb{P}\left(\|\mathbb{G}_\pi - \mathbb{G}_\pi^N\|_1 \geq \delta\right) \leq \frac{\varepsilon}{4F}\,.$$

As a result, we have

$$\begin{aligned}&\mathbb{E}\left[\left|f(x^0,\mathbb{G}_\pi) - f(x^0,\mathbb{G}_\pi^N)\right|\right] \\ &\quad\leq \mathbb{P}\left(\left|f(x^0,\mathbb{G}_\pi) - f(x^0,\mathbb{G}_\pi^N)\right| \geq \frac{\varepsilon}{2}\right)\cdot 2F + 1\cdot\frac{\varepsilon}{2} \\ &\quad\leq \mathbb{P}\left(\|\mathbb{G}_\pi - \mathbb{G}_\pi^N\|_1 \geq \delta\right)\cdot 2F + \frac{\varepsilon}{2} \\ &\quad\leq \frac{\varepsilon}{4F}\cdot 2F + \frac{\varepsilon}{2} = \varepsilon\,.\end{aligned}$$

Since ε was arbitrary, and no choices depended on $\pi \in \Pi$, we have the desired convergence of the second term

$$\lim_{N\to\infty}\sup_{\pi\in\Pi}\left|f(x,\mathbb{G}(\mu_0,\pi_t(x))) - \mathbb{E}\left[f(x_t^{0,N},\mathbb{G}_t^N)\;\middle|\; x_t^{0,N} = x\right]\right| = 0\,.$$

We can now show $\mathcal{L}(x_t^{0,N},\mathbb{G}_t^N) \to \mathcal{L}(x_t^0,\mathbb{G}_t)$ weakly uniformly over all $\pi \in \Pi$ by induction over all t, which by Assumption 4.1.1 will imply

$$\sup_{\pi\in\Pi}\left|\mathbb{E}\left[r(x_t^{0,N},\mathbb{G}_t^N) - r(x_t^0,\mathbb{G}_t)\right]\right| \to 0 \tag{4.1.19}$$

for all $t \geq 0$ and hence the desired statement by the dominated convergence theorem applied to Eq. (4.1.17).

At $t = 0$, we trivially have $\mathcal{L}(x_t^{0,N}) = \mu_0^0 = \mathcal{L}(x_t^0)$ and therefore $\mathcal{L}(x_0^{0,N},\mathbb{G}_0^N) \to \mathcal{L}(x_0^0,\mathbb{G}_0)$ uniformly by the prequel. Assume that the induction assumption holds at time t, then at time $t+1$ we have

$$\begin{aligned}&\|\mathcal{L}(x_{t+1}^{0,N}) - \mathcal{L}(x_{t+1}^0)\|_1 \\ &\quad= \sum_{x^0\in\mathcal{X}^0}\left|\mathcal{L}(x_{t+1}^{0,N})(x^0) - \mathcal{L}(x_{t+1}^0)(x^0)\right| \\ &\quad= \sum_{x^0\in\mathcal{X}^0}\left|\mathbb{E}\left[p^0(x^0 \mid x_t^{0,N},\mathbb{G}_t^N)\right] - \mathbb{E}\left[p^0(x^0 \mid x_t^0,\mathbb{G}_t)\right]\right| \to 0\end{aligned}$$

uniformly by Assumption 4.1.1 and induction assumption. □

To extend to optimality over arbitrary asymmetric policy tuples, we show that the performance of policy tuples is close to the averaged policy as $N \to \infty$.

Theorem 4.1.2. *Under Assumption 4.1.1, as $N \to \infty$ we have similar performance of any policy tuple $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N) \in \Pi_N^N$ and its average policy $\hat{\pi}(\boldsymbol{\pi}) \in \Pi$ defined by $\hat{\pi}_t(x^0)(u \mid x) = \frac{1}{N}\sum_{i=1}^N \pi_t^i(u \mid x^0, x)$ in the N-agent case, i.e. with shorthand $\hat{\pi} = \hat{\pi}(\boldsymbol{\pi})$ we have*

$$\lim_{N\to\infty} \sup_{\boldsymbol{\pi}\in\Pi_N^N} \left| J^N(\pi^1, \dots, \pi^N) - J^N(\boldsymbol{\pi}^N(\hat{\pi})) \right| = 0\,. \tag{4.1.20}$$

Proof. Let $\boldsymbol{\pi} \in \Pi_N^N$ arbitrary. Again, we have by definition

$$\sup_{\boldsymbol{\pi}\in\Pi_N^N} \left| J^N(\pi^1, \dots, \pi^N) - J^N(\boldsymbol{\pi}^N(\hat{\pi})) \right| \tag{4.1.21}$$

$$\leq \sum_{t=0}^{\infty} \gamma^t \sup_{\boldsymbol{\pi}\in\Pi_N^N} \left| \mathbb{E}\left[r(x_t^{0,N}, \mathbb{G}_t^N) - r(\hat{x}_t^{0,N}, \hat{\mathbb{G}}_t^N) \right] \right| \tag{4.1.22}$$

by introducing random variables $\hat{x}_t^{0,N}$, $\hat{\mathbb{G}}_t^N$, $\hat{x}_t^{i,N}$, $\hat{u}_t^{i,N}$, $i = 1, \dots, N$ induced by instead applying the averaged policy tuple $\boldsymbol{\pi}^N(\hat{\pi})$ in Eq. (4.1.2). By dominated convergence, it is sufficient to show term-wise convergence to zero in Eq. (4.1.22).

Fix $t \geq 0$. As in the proof of Theorem 4.1.1, we show that $\sup_{\pi\in\Pi} \|\mathcal{L}(x_t^{0,N}) - \mathcal{L}(\hat{x}_t^{0,N})\|_1 \to 0$ implies $\sup_{\boldsymbol{\pi}\in\Pi_N^N} \left| \mathcal{L}(x_t^{0,N}, \mathbb{G}_t^N)(f) - \mathcal{L}(\hat{x}_t^{0,N}, \hat{\mathbb{G}}_t^N)(f) \right| \to 0$ for any $f \colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X} \times \mathcal{U}) \to \mathbb{R}$ continuous and bounded, since

$$\begin{aligned}
&\sup_{\boldsymbol{\pi}\in\Pi_N^N} \left| \mathcal{L}(x_t^{0,N}, \mathbb{G}_t^N)(f) - \mathcal{L}(\hat{x}_t^{0,N}, \hat{\mathbb{G}}_t^N)(f) \right| \\
&\quad = \sup_{\boldsymbol{\pi}\in\Pi_N^N} \left| \mathbb{E}\left[f(x_t^{0,N}, \mathbb{G}_t^N) \right] - \mathbb{E}\left[f(\hat{x}_t^{0,N}, \hat{\mathbb{G}}_t^N) \right] \right| \\
&\quad \leq \sup_{\boldsymbol{\pi}\in\Pi_N^N} \sum_{x^0\in\mathcal{X}^0} \left| \mathbb{E}\left[f(x_t^{0,N}, \mathbb{G}_t^N) \;\middle|\; x_t^{0,N} = x^0 \right] \right| \left| \mathcal{L}(x_t^{0,N})(x^0) - \mathcal{L}(\hat{x}_t^{0,N})(x^0) \right| \\
&\qquad + \sup_{\boldsymbol{\pi}\in\Pi_N^N} \sum_{x^0\in\mathcal{X}^0} \left| \mathbb{E}\left[f(x^0, \hat{\mathbb{G}}_t^N) \;\middle|\; \hat{x}_t^{0,N} = x^0 \right] - \mathbb{E}\left[f(x^0, \mathbb{G}_t^N) \;\middle|\; x_t^{0,N} = x^0 \right] \right| \left| \mathcal{L}(\hat{x}_t^{0,N})(x^0) \right.
\end{aligned}$$

where the first sum goes to zero by assumption and boundedness of f. For the second term, consider arbitrary fixed $x^0 \in \mathcal{X}^0$, $\boldsymbol{\pi} \in \Pi_N^N$. Then introduce random variables $\mathbb{G}_{\boldsymbol{\pi}}^N \sim \mathcal{G}^N(\mu_0, (\pi_t^1(x^0), \dots, \pi_t^N(x^0)))$ and $\mathbb{G}_{\hat{\pi}}^N \sim \mathcal{G}^N(\mu_0, (\hat{\pi}_t(x^0), \dots, \hat{\pi}_t(x^0)))$ for every $N \in \mathbb{N}$ and $\boldsymbol{\pi} \in \Pi_N^N$. Then we have

$$\begin{aligned}
&\mathbb{E}\left[f(x^0, \hat{\mathbb{G}}_t^N) \;\middle|\; \hat{x}_t^{0,N} = x^0 \right] - \mathbb{E}\left[f(x^0, \mathbb{G}_t^N) \;\middle|\; x_t^{0,N} = x^0 \right] \\
&\qquad = \mathbb{E}\left[f(x^0, \mathbb{G}_{\hat{\pi}}^N \right] - \mathbb{E}\left[f(x^0, \mathbb{G}_{\boldsymbol{\pi}}^N) \right]\,.
\end{aligned}$$

We observe that for any $(x, u) \in \mathcal{X} \times \mathcal{U}$:

$$\mathbb{E}[\mathbb{G}_{\hat{\pi}}^N(x, u)] = \mathbb{E}[\mathbb{G}_{\boldsymbol{\pi}}^N(x, u)]\,. \tag{4.1.23}$$

For this purpose, let $x^{i,N} \sim \mu_0$ and $u^{i,N} \sim \pi_t^i(x^0, x^{i,N})$ as well as $\hat{x}^{i,N} \sim \mu_0$ and $\hat{u}^{i,N} \sim \hat{\pi}_t(x^0, x^{i,N})$, s.t. $(x^{i,N}, u^{i,N})_{i=1,\dots,N}$ and $(\hat{x}^{i,N}, \hat{u}^{i,N})_{i=1,\dots,N}$ are independent, respectively. Then for any $(x, u) \in \mathcal{X} \times \mathcal{U}$:

$$\mathbb{E}[\mathbb{G}_{\hat{\pi}}^N(x, u)] = \frac{1}{N}\sum_{i=1}^N \mathbb{P}(\hat{x}^{i,N} = x, \hat{u}^{i,N} = u)$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N} \mu_0(x)\hat{\pi}(x,u) = \frac{1}{N}\sum_{i=1}^{N} \mu_0(x)\frac{1}{N}\sum_{j=1}^{N} \pi^j(x,u) \\
&= \frac{1}{N}\sum_{j=1}^{N} \mu_0(x)\pi^j(x,u) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}(x^{i,N}=x, u^{i,N}=u) \\
&= \mathbb{E}[\mathbb{G}^N_{\boldsymbol{\pi}}(x,u)]\,.
\end{aligned}
$$

Then by Eq. (4.1.23), sub-additivity of $\mathbb{E}[(\cdot)^2]^{\frac{1}{2}}$ and Lemma 4.1.1 (i),

$$
\begin{aligned}
&\mathbb{E}[\|\mathbb{G}^N_{\hat{\pi}} - \mathbb{G}^N_{\boldsymbol{\pi}}\|_1^2]^{\frac{1}{2}} \\
&\quad\le \mathbb{E}\left[\left(\|\mathbb{G}^N_{\hat{\pi}} - \mathbb{E}[\mathbb{G}^N_{\hat{\pi}}]\|_1 + \|\mathbb{G}^N_{\boldsymbol{\pi}} - \mathbb{E}[\mathbb{G}^N_{\boldsymbol{\pi}}]\|_1\right)^2\right]^{\frac{1}{2}} \\
&\quad\le \mathbb{E}[\|\mathbb{G}^N_{\hat{\pi}} - \mathbb{E}[\mathbb{G}^N_{\hat{\pi}}]\|_1^2]^{\frac{1}{2}} + \mathbb{E}[\|\mathbb{G}^N_{\boldsymbol{\pi}} - \mathbb{E}[\mathbb{G}^N_{\boldsymbol{\pi}}]\|_1^2]^{\frac{1}{2}} \\
&\quad\le 2\frac{|\mathcal{X}|\cdot|\mathcal{U}|}{\sqrt{4N}} = \frac{|\mathcal{X}|\cdot|\mathcal{U}|}{\sqrt{N}}\,.
\end{aligned}
$$

Chebyshev's inequality implies

$$
\mathbb{P}\left(\|\mathbb{G}^N_{\hat{\pi}} - \mathbb{G}^N_{\boldsymbol{\pi}}\|_1 \ge \varepsilon\right) \le \frac{|\mathcal{X}|^2|\mathcal{U}|^2}{\varepsilon^2 N} \tag{4.1.24}
$$

independent of $\boldsymbol{\pi} \in \Pi_N^N$.

Then analogously to the proof of Theorem 4.1.1,

$$
\sup_{\boldsymbol{\pi}\in\Pi_N^N} \left| \mathbb{E}\left[f(x^0, \hat{\mathbb{G}}_t^N) \;\middle|\; \hat{x}_t^{0,N} = x^0 \right] - \mathbb{E}\left[f(x^0, \mathbb{G}_t^N) \;\middle|\; x_t^{0,N} = x^0 \right] \right| \to 0
$$

can be concluded, showing the desired implication.

We now show by induction over all t that for any $t \ge 0$, and any $f\colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}\times\mathcal{U}) \to \mathbb{R}$ continuous and bounded, $\sup_{\boldsymbol{\pi}\in\Pi_N^N} \left|\mathcal{L}(x_t^{0,N}, \mathbb{G}_t^N)(f) - \mathcal{L}(\hat{x}_t^{0,N}, \hat{\mathbb{G}}_t^N)(f)\right| \to 0$ which by Assumption 4.1.1 will again imply that Eq. (4.1.22) goes to zero.

At $t = 0$, we have by definition $\mathcal{L}(x_0^{0,N}) = \mu_0^0 = \mathcal{L}(\hat{x}_0^{0,N})$. This implies that $\sup_{\boldsymbol{\pi}\in\Pi_N^N} \left|\mathcal{L}(x_0^{0,N}, \mathbb{G}_0^N)(f) - \mathcal{L}(\hat{x}_0^{0,N}, \hat{\mathbb{G}}_0^N)(f)\right| \to 0$ for any $f\colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}\times\mathcal{U}) \to \mathbb{R}$ by the prequel. Assuming the induction assumption holds at time t, then at time $t+1$

$$
\begin{aligned}
&\|\mathcal{L}(x_{t+1}^{0,N}) - \mathcal{L}(\hat{x}_{t+1}^{0,N})\|_1 \\
&\quad= \sum_{x^0\in\mathcal{X}^0} \left| \mathbb{E}\left[p^0(x^0 \mid x_t^{0,N}, \mathbb{G}_t^N) - p^0(x^0 \mid \hat{x}_t^{0,N}, \hat{\mathbb{G}}_t^N) \right] \right| \to 0
\end{aligned}
$$

uniformly by induction assumption and continuity and boundedness of p^0, which implies the desired statement. □

Corollary 4.1.1. *Under Assumption 4.1.1, for any $\varepsilon > 0$ there exists $N(\varepsilon)$ such that for all $N > N(\varepsilon)$ a policy π^* optimal in the MFC MDP – that is, $J(\pi^*) = \sup_{\pi\in\Pi} J(\pi)$ – is ε-Pareto optimal in the N-agent case, i.e.*

$$
J^N(\boldsymbol{\pi}^N(\pi^*)) \ge \sup_{\boldsymbol{\pi}} J^N(\boldsymbol{\pi}) - \varepsilon\,. \tag{4.1.25}
$$

Proof. By Theorem 4.1.1 and Theorem 4.1.2, there exists $N' \in \mathbb{N}$ such that for average policy $\hat{\pi}$ of $\boldsymbol{\pi}$ and all $N > N'$ we have

$$\begin{aligned}
&\sup_{\boldsymbol{\pi}} \left(J^N(\boldsymbol{\pi}) - J^N(\boldsymbol{\pi}^N(\pi^*))\right) \\
&\quad \leq \sup_{\boldsymbol{\pi}} \left(J^N(\boldsymbol{\pi}) - J^N(\boldsymbol{\pi}^N(\hat{\pi}))\right) \\
&\qquad + \sup_{\boldsymbol{\pi}} \left(J^N(\boldsymbol{\pi}^N(\hat{\pi})) - J(\hat{\pi})\right) \\
&\qquad + \sup_{\boldsymbol{\pi}} \left(J(\hat{\pi}) - J(\pi^*)\right) \\
&\qquad + \sup_{\boldsymbol{\pi}} \left(J(\pi^*) - J^N(\boldsymbol{\pi}^N(\pi^*))\right) \\
&\quad < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + 0 + \frac{\varepsilon}{3} = \varepsilon\,.
\end{aligned}$$

Reordering terms gives the desired inequality. □

4.1.4 *A Standard Dynamic Programming Principle and Reinforcement Learning*

The following DPP for the MFC MDP is a standard result, for which the MDP state will be only the environment state, see e.g. [104, 121].

Define action-value function $Q\colon \mathcal{X}^0 \times \mathcal{H} \to \mathbb{R}$,

$$Q(x^0, h) \coloneqq \sup_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^0, \mathbb{G}(\mu_0, \pi_t(x_t^0))) \,\middle|\, x_0^0 = x^0, \pi_0(x^0) = h\right]. \tag{4.1.26}$$

Note that by boundedness of r, we trivially have

$$|Q| \leq \frac{R}{1-\gamma}\,.$$

As we have an MDP with finite state space $\mathcal{X}^0$, the following Bellman equation will hold, see [68].

Theorem 4.1.3. *The Bellman equation*

$$Q(x^0, h) = r(x^0, \mathbb{G}(\mu_0, h)) + \gamma \mathbb{E}_{\tilde{x}^0 \sim p^0(x^0, \mathbb{G}(\mu_0, h))}\left[\sup_{\tilde{h} \in \mathcal{H}} Q(\tilde{x}^0, \tilde{h})\right] \tag{4.1.27}$$

holds for all $x^0 \in \mathcal{X}^0, h \in \mathcal{H}$.

In the following, we obtain existence of an optimal stationary policy by compactness of $\mathcal{H}$ and continuity of Q, which shall be inherited from the continuity of r and p^0.

Lemma 4.1.2. *The unique function that satisfies the Bellman equation is given by Q. Further, if there exists $h_{x^0} \in \arg\max_{h \in \mathcal{H}} Q(x^0, h)$ for any $x^0 \in \mathcal{X}^0$, then the policy π^* with $\pi_t^*(x^0) = h_{x^0}$ is an optimal stationary policy.*

Proof. For uniqueness, define the space of $\frac{R}{1-\gamma}$-bounded functions $\mathcal{Q} \coloneqq \{f\colon \mathcal{X}^0 \times \mathcal{H} \to [-\frac{R}{1-\gamma}, \frac{R}{1-\gamma}]\}$ and the Bellman operator $B\colon \mathcal{Q} \to \mathcal{Q}$ defined by

$$(BQ)(x^0, h) = r(x^0, \mathbb{G}(\mu_0, h)) + \gamma \mathbb{E}_{\tilde{x}^0 \sim p^0(x^0, \mathbb{G}(\mu_0, h))} \left[\sup_{\tilde{h} \in \mathcal{H}} Q(\tilde{x}^0, \tilde{h}) \right] . \tag{4.1.28}$$

We show that $\mathcal{Q}$ is a complete metric space under the supremum norm. Let $(Q_n)_{n\in\mathbb{N}}$ be a Cauchy sequence of functions $Q_n \in \mathcal{Q}$. Then by definition, for any $\varepsilon > 0$ there exists $n' \in \mathbb{N}$ such that for all $n, m > n'$ we have

$$\begin{aligned} &\|Q_n - Q_m\|_\infty < \varepsilon \\ \implies \forall x^0 \in \mathcal{X}^0, h \in \mathcal{H}\colon &\left|Q_n(x^0, h) - Q_m(x^0, h)\right| < \varepsilon \end{aligned}$$

such that for all $x^0 \in \mathcal{X}^0, h \in \mathcal{H}$ there exists a value $c_{x^0,h} \in [-\frac{R}{1-\gamma}, \frac{R}{1-\gamma}]$ for which $Q_n(x^0, h) \to c_{x^0,h}$. Define the function $Q' \in \mathcal{Q}$ by $Q'(x^0, h) = c_{x^0,h}$, then we have

$$\begin{aligned} &\left|Q_n(x^0, h) - Q'(x^0, h)\right| \\ &\quad = \lim_{m\to\infty} \left|Q_n(x^0, h) - Q_m(x^0, h)\right| < \varepsilon \end{aligned}$$

for all $x^0 \in \mathcal{X}^0, h \in \mathcal{H}, n > n'$, and hence $Q_n \to Q' \in \mathcal{Q}$ as $n \to \infty$. This implies completeness of $(\mathcal{Q}, \|\cdot\|_\infty)$.

We now show that B is a contraction under the supremum norm, i.e.

$$\|BQ_1 - BQ_2\|_\infty \leq C\|Q_1 - Q_2\|_\infty$$

for some $C < 1$. Define the shorthand $\tilde{x}^0 \sim p^0(x^0, \mathbb{G}(\mu_0, h))$. We have

$$\begin{aligned} &\|BQ_1 - BQ_2\|_\infty \\ &\quad = \sup_{x^0\in\mathcal{X}^0, h\in\mathcal{H}} \left|BQ_1(x^0, h) - BQ_2(x^0, h)\right| \\ &\quad \leq \sup_{x^0\in\mathcal{X}^0, h\in\mathcal{H}} \gamma \mathbb{E}_{\tilde{x}^0} \left[\left| \sup_{\tilde{h}\in\mathcal{H}} Q_1(\tilde{x}^0, \tilde{h}) - \sup_{\tilde{h}\in\mathcal{H}} Q_2(\tilde{x}^0, \tilde{h}) \right| \right] \\ &\quad \leq \sup_{x^0\in\mathcal{X}^0, h\in\mathcal{H}} \gamma \|Q_1 - Q_2\|_\infty \end{aligned}$$

with $\gamma < 1$. Therefore, by Banach fixed point theorem, B has the unique fixed point Q.

For optimality, define the policy action-value function Q^π for $\pi \in \Pi$ as the fixed point of $B^\pi\colon \mathcal{Q} \to \mathcal{Q}$ defined by

$$(B^\pi Q)(x^0, h) = r(x^0, \mathbb{G}(\mu_0, h)) + \gamma \mathbb{E}_{\tilde{x}^0} \left[Q(\tilde{x}^0, \pi(\tilde{x}^0)) \right] .$$

From this, we immediately have

$$\begin{aligned} Q^{\pi^*}(x^0, h) &= r(x^0, \mathbb{G}(\mu_0, h)) + \gamma \mathbb{E}_{\tilde{x}^0} \left[Q(\tilde{x}^0, \pi^*(\tilde{x}^0)) \right] \\ &= r(x^0, \mathbb{G}(\mu_0, h)) + \gamma \mathbb{E}_{\tilde{x}^0} \left[\sup_{\tilde{h}\in\mathcal{H}} Q(\tilde{x}^0, \tilde{h}) \right] \\ &= Q(x^0, h) \end{aligned}$$

which implies that π^* is optimal, see also [68]. □

Lemma 4.1.3. *The action-value function Q is continuous.*

Proof. We will show as $x_n^0 \to x^0 \in \mathcal{X}^0$ and $h_n \to h \in \mathcal{H}$,

$$Q(x_n^0, h_n) \to Q(x^0, h)\,.$$

By the Bellman equation, we immediately have

$$\begin{aligned}
&|Q(x_n^0, h_n) - Q(x^0, h)| \\
&\quad \le \left|r(x_n^0, \mathbb{G}(\mu_0, h_n)) - r(x^0, \mathbb{G}(\mu_0, h))\right| \\
&\qquad + \left|\gamma \sum_{\tilde{x}^0 \in \mathcal{X}^0} \left(p^0(\tilde{x}^0 \mid x_n^0, \mathbb{G}(\mu_0, h_n)) - p^0(\tilde{x}^0 \mid x^0, \mathbb{G}(\mu_0, h))\right) \sup_{\tilde{h} \in \mathcal{H}} Q(\tilde{x}^0, \tilde{h})\right| \\
&\quad \le \left|r(x_n^0, \mathbb{G}(\mu_0, h_n)) - r(x^0, \mathbb{G}(\mu_0, h))\right| \\
&\qquad + \frac{\gamma R}{1-\gamma} \sum_{\tilde{x}^0 \in \mathcal{X}^0} \left|p^0(\tilde{x}^0 \mid x_n^0, \mathbb{G}(\mu_0, h_n)) - p^0(\tilde{x}^0 \mid x^0, \mathbb{G}(\mu_0, h))\right| \to 0
\end{aligned}$$

since $r, p^0, \mathbb{G}$ are continuous and Q is bounded. □

Corollary 4.1.2. *There exists an optimal stationary policy $\pi^*\colon \mathcal{X}^0 \to \mathcal{H}$ such that $Q^{\pi^*} = Q$.*

Proof. By Lemma 4.1.3, Q is continuous. Furthermore, $\mathcal{H}$ is compact. By the extreme value theorem, there exists $h_{x^0} \in \arg\max_{h \in \mathcal{H}} Q(x^0, h)$ for any $x^0 \in \mathcal{X}^0$. By Lemma 4.1.2, there exists an optimal stationary policy π^*. □

Since $\mathcal{H}$ is continuous, general exact solutions are difficult. Instead, we apply RL with stochastic policies on the MFC MDP to find an optimal stationary policy.

4.1.5 *Experiments*

We compare the empirical performance of the MF solution in the aforementioned load balancing problem. Since there exist few theoretical guarantees for tractable MARL methods [40], we compare our approach (MF) to empirically effective independent learning (IL) [86], i.e. applying single-agent RL to each separate agent (NA), as well as the well-known Join-Shortest-Queue (JSQ) algorithm [203], where agents choose the shortest queue accessible and otherwise randomly. To make independent learning more tractable, we also share policy parameters between all agents using parameter sharing (PS) [90] and train each policy via the PPO algorithm [73] using the RLlib 1.2.0 Pytorch implementation [76] for $400,000$ time steps in the N-agent cases and 2 million time steps in the MF case, which is sufficient for convergence of MF and N-agent policies up to $N = 4$, after which N-agent training becomes unstable under the shared hyperparameters in Table 4.1 and continues to fail even with more time steps.

For policies and critics, we use separate feedforward networks with two hidden layers of 256 nodes and $\tanh$ activations. In the MF case the policy outputs parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$ of a diagonal Gaussian distribution over actions, which are sampled and clipped between 0 and 1. We normalize each of these output values such that they give the probability of assigning to an accessible queue given some agent state, i.e. a shared lower-level decision rule $h \in \mathcal{H}$ for all agents. A visualization of this process can be found in Figure 4.3.

TABLE 4.1: Parameters and hyperparameters used in the experiments.

Symbol	Function	Value
c_d	Packet drop penalty	1
M	Number of queues	2
B_i	Queue buffer sizes	5
ΔT	Time step size	$0.5\,\mathrm{s}$
λ	Packet arrival rate	$(3M-1)\ \mathrm{s}^{-1}$
β	Queue servicing rate	$3\,\mathrm{s}^{-1}$
γ	Discount factor	0.99
	PPO	
l_r	Learning rate	5×10^{-5}
λ_{PPO}	GAE coefficient	0.2
β_{PPO}	Initial KL coefficient	0.2
d_{targ}	KL target	0.01
ε	Clip parameter	0.3
B	Training batch size	4000
B_m	SGD mini-batch size	128
k	SGD iterations per batch	30

Note that we use stochastic policies as required by stochastic policy gradient methods, though we can easily obtain a deterministic policy if necessary by simply using the mean parameter of the Gaussian distribution. In the N-agent case, we output queue assignment probabilities for each of the agents via a standard softmax final layer. Invalid assignments to queues that are not accessible by an agent are treated as randomly sampling one from all accessible queues.

As can be seen in Figure 4.4 for μ_0 given such that the probability of access to both queues is 0.6 and otherwise uniformly random, the MF solution reaches its MF performance in the N-agent case as N grows large. This validates our theoretical findings empirically. Our solution further appears to outperform NA and PS for sufficiently many agents, as IL approaches increasingly fail to learn due to the credit assignment problem.

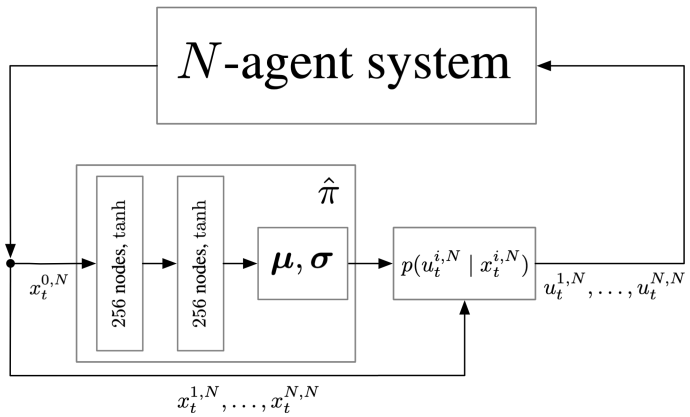


FIGURE 4.3: Overview of MFC application in N-agent systems: Conditional on the environment state $x_t^{0,N}$, the upper-level MF policy $\hat{\pi}$ outputs a sampled, shared lower-level policy for all agents i, from which random actions $u_t^{i,N}$ are sampled conditional on local agent states $x_t^{i,N}$.

Moreover, our best learned policy is close to JSQ and competitive with slight irregularities at $b_0 = 0$. Observe in Figure 4.4 that the MF policy gives an interpretable solution. NA and PS are trained separately for each N, while the MF policy is trained only once and used for all N. As N grows, the MF policy performance becomes increasingly close to the MFC MDP and competitive with JSQ, while NA and PS begin to fail learning due to the credit assignment problem. (b): MF policy probabilities of assigning to queue 1 against buffer fillings b_0, b_1 for agents with access to both queues, averaged over 500 samples. As a queue becomes more filled, the optimal solution will be more likely to avoid assignment of packets to that queue.

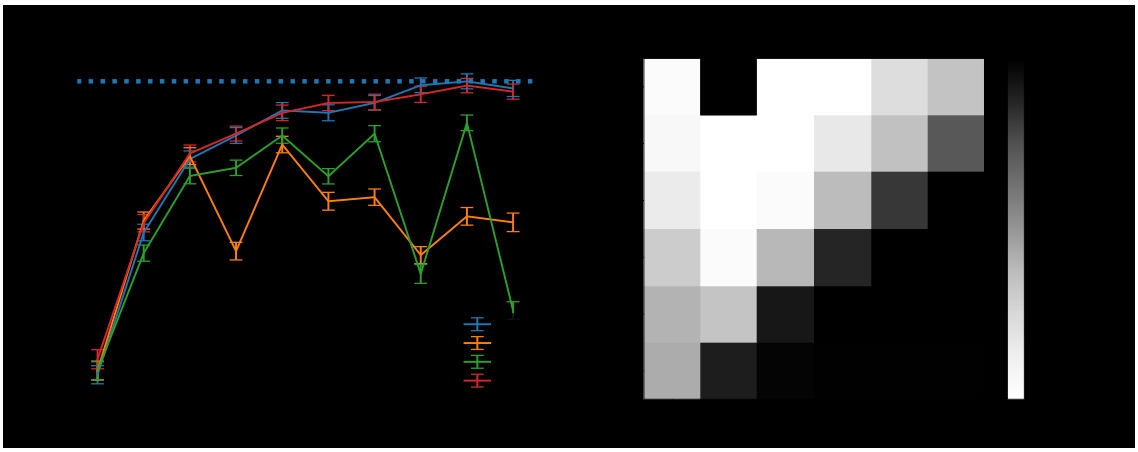


FIGURE 4.4: Qualitative evaluation of learned balancing policy. (a): Cumulative reward average over 500 runs with 95% confidence interval achieved against number of agents N. The dotted line indicates cumulative reward of MF in the MFC MDP.

4.1.6 *Summary*

In this work, we have formulated a discrete-time MFC model with common environment states motivated by a load balancing problem. We have rigorously shown approximate optimality as $N \to \infty$ and applied RL to solve the MFC MDP. Empirically, we obtain competitive results for sufficiently many agents and validate our theoretical results. For future work, it could be interesting to consider partial observability of the system for schedulers, or methods to scale to large numbers of queues. Potential extensions are manifold and include, e.g., major-minor systems, partial observability. We note that since the appearing of this work, separate work [128] has relatedly shown the sufficiency of heterogeneous policies in standard MFC without external states, as well as two-team generalizations.

4.2 TOWARDS STRONG INTERACTION IN MEAN FIELD CONTROL

Recent MARL using MFC provides a tractable and rigorous approach to otherwise difficult cooperative MARL. However, the strict MFC assumption of many independent, weakly-interacting agents is too inflexible in practice. We generalize MFC to instead simultaneously model many similar and few complex agents – as M3FC. Theoretically, we give approximation results for finite agent control, and verify the sufficiency of stationary policies for optimality together with a DPP. Algorithmically, we propose Major-Minor Mean Field Multi-Agent Reinforcement Learning (M3FMARL) for finite agent systems instead of the limiting system. The algorithm is shown to approximate the policy gradient of the underlying M3FC MDP. Finally, we demonstrate its capabilities experimentally in various scenarios. We observe a strong performance in comparison to state-of-the-art policy gradient MARL methods. The material presented in this section is based upon our work [1].

MEAN FIELD CONTROL FOR MARL. Aggregated interaction models such as MFGs [22, 23] and MFC [107, 124, 205] simplify MARL in the limit of infinite agents, with problem complexity independent of the exact number of agents. The result is tractability by avoiding exponentially large joint state-action spaces [40]. This has led to scalable MARL via MFC [107, 201]. And indeed, in applications such aggregation is commonly found on some level, e.g., in chemical reaction networks for aggregate molecule mass [206], related mass-action epidemics models [49], or traffic where congestion depends on the number of travelling cars [54], to name just a few. See also epidemics control [207], drone swarms [208], self organization [209], and many more financial [64] or engineering scenarios [65].

LIMITATIONS OF STANDARD MFC. However, the strict assumption of only *minor* agents – i.e. independent, homogeneous agents that can be summarized by their distribution (MF) – limits applicability. In practice, systems often consist of more than homogeneous agents, and hence one must extend standard MFC towards *major* agents or environment states that are not aggregated. For instance, in modelling car traffic on road networks [54, 61], when considering only the distribution of cars (*minor* agents) on the network, one cannot model *major* agents or environment states, such as traffic lights or road conditions respectively. Another example is the logistics scenario in Figure 4.5 and in the experiments, where many drones on a moving truck collect many packages.

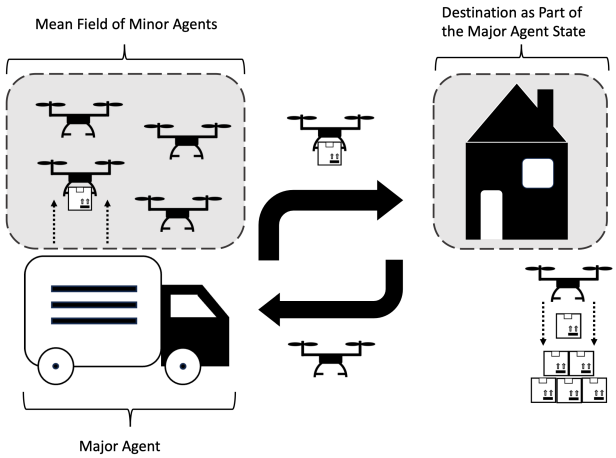


FIGURE 4.5: Logistics example for major-minor MFC: Many drones are modelled as minor agent MF, while truck and package destinations are modelled by a major agent. (See Foraging problem in Section 4.2.3.1)

For this purpose, a first step in the continuous-time MFG literature is to consider common noise [127, 178], in order to relax the unconditional independence of minor agents. Some more recent works consider such common noise also in discrete-time MFC [107, 121, 210, 211], or equivalently, global environment states [212]. Essentially, this extension allows MFC to also model random environment effects such as the arrival of new packages in the logistics example (Figure 4.5). [107] provide a reformulation of MARL into single-agent RL and consider algorithms for the resulting MDP. [210] give approximation theorems and approximate optimality in the finite system by the limiting MFC solution with common noise, and [121, 211] quantify the rates of convergence explicitly. See also Table 4.2 for a brief comparison between existing works. In comparison, for the common noise setting, we contribute a new approximation analysis of MFC-based MARL algorithms, where in contrast to prior work, we learn directly with *finite agents*.

TABLE 4.2: A comparison of recent related works and a subset of their results on *discrete-time* MFC. *prop. chaos*: propagation of chaos; *opt. policy*: existence of optimal (stationary) policies; *common noise*: presence thereof; *non-finite*: non-finite state-actions, e.g. compact; *major agent*: presence thereof; *RL*: RL algorithm ($^+$: learns / is analyzed on finite MARL problems).

Ref.	*prop. chaos*	*opt. policy*	*common noise*	*non-finite*	*major agent*	*RL*
[107]	✗	✓	✓	✓	✗	✓
[104, 201]	✓	✓	✗	✗	✗	✓
[210]	✓	✓	✓	✓	✗	✗
[105, 212]	✓	✗	✓	✗	✗	✓
[121, 211]	✓	✓	✓	✓	✗	✗
our work	✓	✓	✓	✓	✓	✓$^+$

More importantly however, a second contribution is to consider major *agents*. Major agents generalize common noise or environmental states, and take actions that have a non-negligible effect on the system. So far, major agents have only been considered in *continuous*-time, *non*-cooperative MFGs [113, 185, 202, 213]. To the best of our knowledge, no such *discrete*-time, *cooperative* framework has been formulated yet. In this work, we investigate such a MARL framework.

CONTRIBUTION. Existing MFC cannot model general agents and many aggregated agents simultaneously. In essence, we generalize the solution spaces of single-agent RL and MFC-based MARL – frameworks for cooperative MARL as depicted in Figure 4.6. This provides both tractability

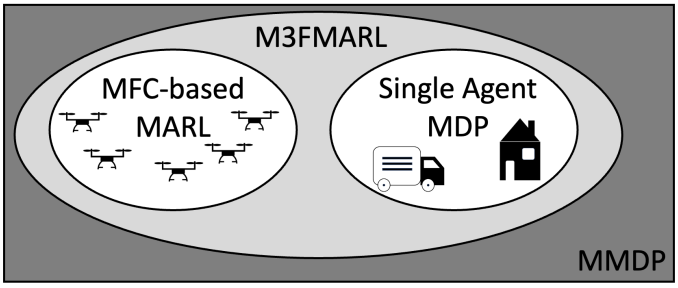


FIGURE 4.6: Comparison of solution spaces. Our M3FC-based MARL generalizes MFC-based MARL and standard single-agent RL in the solution space of general MARL solutions, reducing the otherwise combinatorial nature of MARL [40] to a tractable but still general setting.

for many aggregated agents and generality for arbitrary general agents. Our contribution is briefly summarized into (i) formulating the first discrete-time MFC model with major agents, together with establishing its theoretical properties; (ii) providing a MFC-based MARL algorithm, which in contrast to prior work learns on the finite problem of interest; and (iii) we perform a significant empirical evaluation, also obtaining positive comparisons of MFC-based MARL against state of the art, whereas prior works on MFC were limited to verifying algorithms on one or two examples.

4.2.1 *Major-Minor Mean Field Control*

To begin, in this section we extend standard MFC by modelling the presence of a major agent. The generalization to more than one major agent is straightforward. This leads to our discrete-time M3FC model. Overall, we obtain a formulation that allows standard MARL handling of major agents, while tractably handling many minor agents via MFC-based techniques.

Notation: By $\mathbb{E}_X$ we denote conditional expectations given X. The space of probability measures $\mathcal{P}(\mathcal{X})$ on compact metric spaces $\mathcal{X}$ is equipped with the 1-Wasserstein distance, unless noted otherwise [214]. Note compactness of $\mathcal{P}(\mathcal{X})$ on compact $\mathcal{X}$ by Prokhorov's theorem [215]. Hence, we sometimes use the uniformly (not Lipschitz) equivalent metric $d_\Sigma(\mu,\mu') \coloneqq \sum_{m=1}^{\infty} 2^{-m} |\int f_m \,\mathrm{d}(\mu-\mu')|$, for some sequence of continuous $f_m \colon \mathcal{X} \to [-1,1]$ [216, Theorem 6.6].

4.2.1.1 *Finite-Agent System*

Consider N (minor) agents $i \in [N] \coloneqq \{1,\dots,N\}$ with compact metric state and action spaces $\mathcal{X}, \mathcal{U}$, equipped with random states and actions $x_t^{i,N}$ and $u_t^{i,N}$ at times $t \in \mathbb{N}$, where initial states $x_0^{i,N} \sim \mu_0$ are independently sampled from some initial distribution $\mu_0 \in \mathcal{P}(\mathcal{X})$. In addition to standard MFC, we also consider a single major agent, though the framework can be extended to multiple. Consider major agent state and action spaces, $\mathcal{X}^0, \mathcal{U}^0$ and state-actions $x_t^{0,N}$, $u_t^{0,N}$, with the major agent formally indexed by $i = 0$. Given all actions, the agent states evolve according to kernels p, p^0 depending on (i) the agent's own state-actions, (ii) the major state-actions, and (iii) the empirical MF, i.e. the $\mathcal{P}(\mathcal{X})$-valued empirical state distribution $\mu_t^N \coloneqq \frac{1}{N}\sum_{i=1}^{N} \delta_{x_t^{i,N}}$. This means that minor agents affect other agents only at rate $\frac{1}{N}$. In practice, we identify minor agents as all agents that matter through their MF μ_t^N. Any remaining agents are major, such that the problem-specific stratification into major and minor agents is always possible.

By symmetry, the system state at any time t is therefore entirely given by $(x_t^{0,N}, \mu_t^N)$. Accordingly, in MFC we share policies between all minor agents. We consider time-variant policies $\pi \in \Pi$, $\pi^0 \in \Pi^0$ from some classes of major and minor policies Π, Π^0 that depend on an agent's own state and $(x_t^{0,N}, \mu_t^N)$ at all times t. Overall, for all $i \in [N]$ and $t \in \mathbb{N}$, the finite MFC system follows

$$u_t^{i,N} \sim \pi_t(u_t^{i,N} \mid x_t^{i,N}, x_t^{0,N}, \mu_t^N), \tag{4.2.29a}$$
$$u_t^{0,N} \sim \pi_t^0(u_t^{0,N} \mid x_t^{0,N}, \mu_t^N), \tag{4.2.29b}$$
$$x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N), \tag{4.2.29c}$$
$$x_{t+1}^{0,N} \sim p^0(x_{t+1}^{0,N} \mid x_t^{0,N}, u_t^{0,N}, \mu_t^N)\,. \tag{4.2.29d}$$

The goal is then to maximize the infinite-horizon discounted objective $J^N(\pi,\pi^0) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^{0,N}, u_t^{0,N}, \mu_t^N)\right]$ over minor and major policies (π,π^0), with discount $\gamma \in (0,1)$

and reward function $r\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. While an optimal behavior could be learned using standard MARL policy gradient methods, for improved tractability we introduce the following M3FC model in the case of many minor agents.

Remark 4.2.1. *The model is as expressive as in existing MFC [105, 201], as it also includes (i) joint state-action MFs $\nu_t \in \mathcal{P}(\mathcal{X}\times\mathcal{U})$, by splitting time steps in two and defining new states in $\mathcal{X}\cup\mathcal{X}\times\mathcal{U}$, (ii) average rewards over all agents, and (iii) random rewards r_t^i by $r(\mu_t^N) \equiv \frac{1}{N}\sum_{i=1}^N \mathbb{E}[r_t^i \mid x_t^{i,N}, \mu_t^N]$. A finite horizon is handled analogously (without optimal stationary policies).*

4.2.1.2 *Mean Field Control Limit*

By the introduction of the MF limit, we obtain a large, more tractable subclass of cooperative multi-agent control problems, which may otherwise suffer from the curse of many agents (combinatorial joint state-action space, [40]). We introduce the MF limit by formally taking $N \to \infty$: The finite-agent control problem is replaced by a higher-dimensional single-agent MDP – the M3FC MDP. By symmetry, we summarize minor agents into their probability law, the MF $\mu_t \equiv \mathcal{L}(x_t^{i,N}) \in \mathcal{P}(\mathcal{X})$. It replaces its empirical analogue μ_t^N by a LLN. Thus, by definition, the MF μ_t evolves forward as

$$\mu_{t+1} = T(x_t^0, u_t^0, \mu_t, \mu_t \otimes \pi_t(\mu_t)) = \iint p(x, u, x_t^0, u_t^0, \mu_t)\pi_t(\mathrm{d}u \mid x, \mu_t)\mu_t(\mathrm{d}x), \tag{4.2.30}$$

with $\pi_t(\mu_t) \coloneqq \pi_t(\cdot \mid \cdot, \mu_t)$, product measures $\mu_t \otimes \pi_t(\mu_t)$ of measure μ_t and kernel $\pi_t(\mu_t)$ on $\mathcal{X}\times\mathcal{U}$, and deterministic dynamics for the MF, $T(x^0, u^0, \mu, h) \coloneqq \iint p(\cdot \mid x, u, x^0, u^0, \mu)h(\mathrm{d}x, \mathrm{d}u)$.

Therefore, the state of the limiting system consists only of the MF μ_t and major state x_t^0. As a result, we obtain the limiting *M3FC MDP*

$$h_t \sim \hat{\pi}_t(h_t \mid x_t^0, \mu_t), \tag{4.2.31a}$$
$$u_t^0 \sim \pi_t^0(u_t^0 \mid x_t^0, \mu_t), \tag{4.2.31b}$$
$$\mu_{t+1} = T(x_t^0, u_t^0, \mu_t, h_t), \tag{4.2.31c}$$
$$x_{t+1}^0 \sim p^0(x_{t+1}^0 \mid x_t^0, u_t^0, \mu_t) \tag{4.2.31d}$$

with objective $J(\hat{\pi}, \pi^0) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(x_t^0, u_t^0, \mu_t)\right]$ and transition dynamics for the MF $T(x^0, u^0, \mu, h) \coloneqq \iint p(\cdot \mid x, u, x^0, u^0, \mu)h(\mathrm{d}x, \mathrm{d}u)$. Here, we identify $\mu_t \otimes \pi_t(\mu_t) \equiv h_t \in \mathcal{H}(\mu_t)$ in the compact set $\mathcal{H}(\mu) \subseteq \mathcal{P}(\mathcal{X}\times\mathcal{U})$ of *desired joint state-action distributions with first marginal* μ as part of the action of the M3FC MDP.

In other words, the action of the M3FC MDP is (h_t, u_t^0) where h_t replaces all the minor agent actions by a LLN. Accordingly, minor agent policies are replaced by MFC policies $\hat{\pi}$ mapping from current μ_t to desired state-action distribution h_t. The limiting M3FC model abstracts away all the minor agents in the finite system, and considers only the MF and the major agents, as visualized in Figure 4.7. The reason for writing joint h_t is mostly technical, as for deterministic $\hat{\pi}$, we write $\pi_t = \Phi(\hat{\pi}_t)$ to reobtain agent policies μ_t-a.e. uniquely by disintegration [217] of $h_t = \hat{\pi}_t(\mu_t)$ into $\mu_t \otimes \pi_t'$ with decision rule $\pi_t' \in \mathcal{P}(\mathcal{U})^{\mathcal{X}}$ and using $\pi_t(\mu_t) \equiv \pi_t'$. Inversely, any $\pi \in \Pi$ is represented in the MFC MDP by deterministic $\hat{\pi}_t = \Phi^{-1}(\pi)_t = \mu_t \otimes \pi_t$.

Remark 4.2.2. *Strictly speaking, in finite-agent control one could jointly select actions $(u_t^{0,N}, u_t^{1,N}, \ldots, u_t^{N,N})$ given joint states $(x_t^{0,N}, x_t^{1,N}, \ldots, x_t^{N,N})$. But intuitively, (i) joint states reduce to $(x_t^{0,N}, \mu_t^N)$, while (ii) joint actions are replaced by the LLN and sampling actions. Optimality of MFC solutions over larger classes of heterogeneous or joint policies is plausible, but to the best of our knowledge, general result are still limited. See also Appendix E.17.*

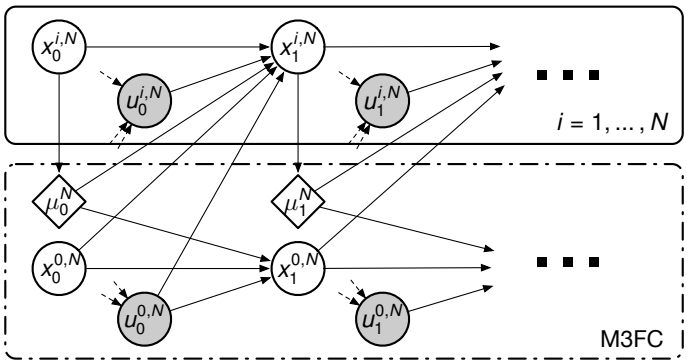


FIGURE 4.7: The dynamics Eq. (4.2.29) as a probabilistic graphical model, with actions in grey (inputs omitted for readability). Diamonds denote deterministic functions. M3FC abstracts minor agents $i \in [N]$ by a LLN, considering only their MF as variables in the dotted box.

For the unfamiliar reader, in Appendix E.2 we recap basic deterministic MFC without major agents or common noise. There, we recap Lipschitz approximation theorems and the DPP in compact spaces.

COMMON NOISE AND GLOBAL STATES. In the classical sense [121, 127], common noise is given by random noise $\epsilon_t^0 \sim p_\epsilon(\epsilon_t^0)$ sampled from a fixed distribution p_ϵ, and affects all minor agents at once, $x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, \epsilon_t^0, \mu_t^N)$. This allows to model systems with stochastic MFs and inter-agent correlation, and has added difficulty to the theoretical analysis [178]. Of similar interest are also "major" global states $x_t^{0,N}$, which need not be sampled from fixed distributions but evolve dynamically (for MFC with finite global states, see e.g. [212]).

Both common noise and global states are contained in the M3FC model by using a trivial major agent without actions. We also note that, in general, common noise is equivalent to global states, as global states can be integrated into the minor state conditioned on the common noise. However, for computational purposes the separation of global states and minor agent states can be helpful, as the simplex $\mathcal{P}(\mathcal{X})$ over minor states can be kept smaller for methods based on discretization of the simplex.

4.2.1.3 *Dynamic Programming*

As a first step, it is well known that stationary (time-independent) policies suffice for optimality in infinite-horizon discounted MDPs. In the following, this property is also verified for the M3FC MDP. For the following technical results, we assume standard Lipschitz conditions [104, 105, 116].

Assumption 4.2.1. *The transition kernels p, p^0 and rewards r are Lipschitz with constants L_p, L_{p^0}, L_r.*

Assumption 4.2.1 is true, e.g., in finite spaces if transition matrix entries of P are Lipschitz in the $|\mathcal{X}|$-dimensional MF vector. The sufficiency of stationary policies is obtained by the DPP, which can also be used to compute exact optimal policies in the M3FC MDP. We use the value function V^* as the fixed point of the Bellman equation, $V^*(x^0, \mu) = \max_{(h,u^0)\in\mathcal{H}(\mu)\times\mathcal{U}^0} r(x^0, u^0, \mu) + \gamma \mathbb{E}_{y^0\sim p^0(y^0|x^0,u^0,\mu)} V^*(y^0, T(x^0, u^0, \mu, h))$.

Theorem 4.2.1. *Under Assumption 4.2.1, there exist optimal stationary, deterministic policies* $\hat{\pi}$, π^0 *for the M3FC MDP Eq.* (4.2.31) *by choosing* $(\hat{\pi}(x^0,\mu),\pi^0(x^0,\mu))$ *from the maximizers of* $\arg\max_{(h,u^0)\in\mathcal{H}(\mu)\times\mathcal{U}^0} r(x^0,u^0,\mu)+\gamma\mathbb{E}_{y^0\sim p^0(y^0|x^0,u^0,\mu)}V^*(y^0,T(x^0,u^0,\mu,h))$.

Remark 4.2.3. *We obtain existence of optimal deterministic stationary minor and major policies* $\hat{\pi}$, π^0 *via optimal joint policies* $\tilde{\pi}\equiv\hat{\pi}\otimes\pi^0$, $(h_t,u_t^0)\sim\tilde{\pi}((h_t,u_t^0)\mid x_t^0,\mu_t)$.

The results follow from classical MDP theory [69]. Thus, we may solve M3FC problems through the DPP, or approximately by using policy gradients with *stationary* policies for the M3FC MDP, which has naturally continuous actions.

4.2.1.4 *Finite Agent Convergence*

Next, in order to show the approximate optimality of M3FC solutions, we first obtain **propagation of chaos** [218] – convergence of empirical MFs to the limiting MF. The result theoretically backs the reduction of multi-agent control to *single-agent* MDPs, as there is no loss of optimality in the finite problem by considering the M3FC problem. We assume standard Lipschitz conditions on policies [104, 105, 116].

Assumption 4.2.2. *The classes of policies* Π, Π^0 *are equi-Lipschitz sets of policies, i.e. there exists* $L_\Pi>0$ *such that for all* t *and* $\pi\in\Pi$, $\pi_t\in\mathcal{P}(\mathcal{U})^{\mathcal{X}\times\mathcal{P}(\mathcal{X})}$ *is* L_Π*-Lipschitz, and similarly for major policies* $\pi^0\in\Pi^0$.

We note that Lipschitz policies are natural, as we usually parametrize policies in a Lipschitz manner; in particular, neural networks allow Lipschitz analysis [116, 219, 220]. The result is that the limiting system approximates large finite systems.

Theorem 4.2.2. *Fix any family of equi-Lipschitz functions* $\mathcal{F}\subseteq\mathbb{R}^{\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})}$ *with shared Lipschitz constant* $L_\mathcal{F}$. *Under Assumptions 4.2.1 and 4.2.2,* $(x_t^{0,N},u_t^{0,N},\mu_t^N)$ *converges weakly to* (x_t^0,u_t^0,μ_t), *uniformly over* $f\in\mathcal{F}$, $(\pi,\pi^0)\in\Pi\times\Pi^0$, $\hat{\pi}=\Phi^{-1}(\pi)$ *at all times* $t\in\mathbb{N}$,

$$\sup_{f,\pi,\pi^0}\left|\mathbb{E}\left[f(x_t^{0,N},u_t^{0,N},\mu_t^N)-f(x_t^0,u_t^0,\mu_t)\right]\right|\to 0. \tag{4.2.32}$$

Further, the convergence rate is $\mathcal{O}(1/\sqrt{N})$ *if* $|\mathcal{X}|<\infty$.

The above motivates M3FC by the following **near optimality** result of M3FC MDP solutions in the finite system, as it suffices to optimize over stationary M3FC policies.

Corollary 4.2.1. *Under Assumptions 4.2.1 and 4.2.2, optimal deterministic M3FC MDP policies* $(\hat{\pi}^*,\pi^{0*})\in\arg\max_{(\hat{\pi},\pi^0)}J(\hat{\pi},\pi^0)$ *with* $\Phi(\hat{\pi}^*)\in\Pi$ *yield* ε*-optimal* $(\Phi(\hat{\pi}^*),\pi^{0*})$ *with* $\varepsilon\to 0$ *as* $N\to\infty$ *in the finite system,* $J^N(\Phi(\hat{\pi}^*),\pi^{0*})\geq\sup_{(\pi,\pi^0)\in\Pi\times\Pi^0}J^N(\pi,\pi^0)-\varepsilon$.

Therefore, one may solve difficult finite-agent MARL by detouring over the corresponding M3FC MDP as depicted in Figure 4.8, reducing to an MDP of a complexity independent of the number of agents N, which we solve in Section 4.2.2.

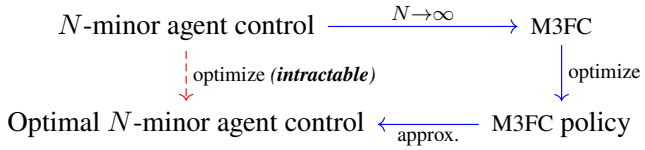


FIGURE 4.8: Approximation of intractable N-agent control by M3FC (blue path), the solution of which is near-optimal for large N.

4.2.2 *Major-Minor Mean Field Multi-Agent Reinforcement Learning*

As indicated in the prequel and in Figure 4.6, MARL via M3FC generalizes both single-agent RL and MARL via MFC in the searched policy solution space. Therefore, in M3FC one only optimizes over a tractable, smaller solution space of a single minor and major policy Π, Π^0. At the same time, the framework is highly general and handles arbitrary major agents with many minor agents simultaneously. The reduction of MARL problems to a fixed-complexity single-agent M3FC MDP is the key. In this section, we develop MARL algorithms based on the M3FC framework.

Recalling the motivation of MFC, it is crucial to find tractable sample-based MARL techniques for both complex problems where other methods fail, and for problems where we have no access to the dynamics or reward model. Relating to the former, RL has been applied before to solve MFC given that we know the MFC model equations [105, 107, 116]. However, regarding the latter, we should instead use the MFC formalism to give rise to novel *MARL* algorithms.

While literature usually focused analysis on the former, in our work we analyze the proposed algorithm not on limiting M3FC MDPs, but on the more interesting finite M3FC system. In particular, if the M3FC MDP is known, one can instantiate finite systems of any size for training. We consider the following perspective: By Theorem 4.2.2, the M3FC MDP is approximated well by the finite system. Therefore, we can solve the limiting M3FC MDP by applying our proposed algorithm directly to finite M3FC systems.

Since we know by Theorem 4.2.1 that stationary policy suffice, we solve the M3FC MDP Eq. (4.2.31) using stationary policies and single-agent RL techniques but on its finite multi-agent instance Eq. (4.2.29), the combination of which we aptly refer to as M3FMARL. The result is Algorithm 2, where we directly apply RL to multi-agent systems Eq. (4.2.29) by observing next states $(x_{t+1}^{0,N}, \mu_{t+1}^N)$ and rewards $r_t^N := r(x_t^{0,N}, u_t^{0,N}, \mu_t^N)$. The algorithm can be understood as a kind of hierarchical algorithm, as M3FC MDP actions specify behavior for all minor agents at once.

Algorithm 2 M3FMARL

for $n = 0, 1, \ldots$ **do**
  for $t = 0, \ldots, B_{\text{len}} - 1$ **do**
    Sample M3FC action from RL policy, i.e.
    $u_t \equiv (u_t^{0,N}, \pi_t') \sim \tilde{\pi}^\theta(u_t \mid x_t^{0,N}, \mu_t^N)$.
    for $i = 1, \ldots, N$ **do**
      Sample i-th minor action $u_t^{i,N} \sim \pi_t'(u_t^{i,N} \mid x_t^{i,N})$.
    Execute $u_t^{0,N}, u_t^{1,N}, \ldots$ for reward r_t^N, next state $(x_{t+1}^{0,N}, \mu_{t+1}^N)$, termination $d_{t+1} \in \{0, 1\}$.
  Perform an update on $\tilde{\pi}^\theta$ using $B = ((x_t^{0,N}, \mu_t^N), u_t, r_t^N, d_{t+1}, (x_{t+1}^{0,N}, \mu_{t+1}^N))_{t \geq 0}$.

4.2.2.1 *M3FC-based Policy Gradients*

The proposed algorithm can be theoretically motivated. As shown in the following, finite-agent Policy Gradients (PGs) estimate the true limiting M3FC MDP PG. First, note that finite state-actions $\mathcal{X}, \mathcal{U}$ lead to continuous M3FC MDP actions $\mathcal{H}(\mu)$, while continuous $\mathcal{X}, \mathcal{U}$ even yield infinite-dimensional $\mathcal{H}(\mu)$. Therefore, we have at least continuous MDPs, complicating value-based learning.

For this reason, we mainly consider PG methods to solve M3FC-type MARL problems. We parametrize M3FC MDP solutions via RL policies $\tilde{\pi}^\theta$ with parameters θ, outputting $\xi \in \Xi$ from some compact parameter space Ξ with a Lipschitz map $\Gamma(\xi) = \pi'_t$ to L_Π-Lipschitz minor agent decision rules π'_t (formally, $h_t = \mu_t \otimes \pi'_t$). Assuming the Lipschitzness of the policy network and its gradient in all arguments, on which there has been a great number of recent literature (see e.g. [219, 220] and references therein), we formulate Assumption 4.2.3.

Assumption 4.2.3. *The parameter map Γ, joint policy $\tilde{\pi}^\theta$ and log-gradient $\nabla_\theta \log \tilde{\pi}^\theta$ (or gradient $\nabla_\theta \tilde{\pi}^\theta$) are L_Γ, $L_{\tilde{\pi}}$, $L_{\nabla\tilde{\pi}}$-Lipschitz and uniformly bounded.*

Then, we can apply the PG theorem [221] for the M3FC MDP. The M3FC MDP Eq. (4.2.31) essentially substitutes many-agent systems Eq. (4.2.29), which are natural approximations of the M3FC MDP by Theorem 4.2.2. Therefore, we show that M3FMARL (Algorithm 2) – single-agent PG on the multi-agent M3FC system – approximates the true PG of the limiting M3FC MDP, in the case of many minor agents. In other words, M3FMARL solves MARL by approximately solving the single-agent M3FC MDP using policy gradients.

Theorem 4.2.3. *Under Assumptions 4.2.1, 4.2.2 and 4.2.3, the approximate PG of joint policy $\tilde{\pi}^\theta$ computed on the finite M3FC system Eq.* (4.2.29) *in Algorithm 2 uniformly tends to the true PG of the M3FC MDP Eq.* (4.2.31)*, as $N \to \infty$.*

Importantly, the underlying MDP complexity is *independent* of the number of minor agents. Therefore, we would expect Algorithm 2 to be able to perform well in M3FC-type problems, possibly compared to straightforward MARL where each agent is handled separately. Intuitively, for many agents, the reward signal for any single agent can become uninformative: A cooperative, "averaged" reward remains almost unaffected by a single agent's actions. This well-known credit assignment issue is therefore solved by the hierarchical structure of M3FC, as credit is assigned to M3FC actions, which affect all minor agents at once and hence receive aggregated credit. Another advantage is that MFC profits from any advances in single-agent RL.

4.2.2.2 *Implementation Details*

We use the PPO algorithm [73] to obtain a M3FC policy π_{RL}, instantiating the Major-Minor Mean Field PPO (M3FPPO) algorithm as an instance of M3FMARL, Algorithm 2. Other PG algorithms (Advantage Actor Critic (A2C), leading to Major-Minor Mean Field Advantage Actor Critic (M3FA2C)) are also compared in our experiments. We parametrize MFs in $\mathcal{P}(\mathcal{X})$ and joint distributions in $\mathcal{H}(\mu_t^N)$. In practice, for finite $\mathcal{X}$, $\mathcal{U}$, the parametrization of $\mathcal{P}(\mathcal{X})$ is immediate by finite-dimensional vectors $\mu_t^N \in \mathcal{P}(\mathcal{X})$. For M3FC actions, consider – in addition to the major agent action – the matrix $\xi \in [-1,1]^{\mathcal{X}\times\mathcal{U}}$, which is mapped to probabilities of minor actions in any minor state $\pi'_t(u \mid x) \coloneqq Z^{-1}(\xi_{xu} + 1 + \epsilon)$, for small $\epsilon = 10^{-10}$ and normalizer Z. For continuous $\mathcal{X}, \mathcal{U}$, we instead partition $\mathcal{X}$ into M bins and represent μ_t^N as a histogram, mapping $\xi \in [-1,1]^{M\times 2}$ instead

to diagonal Gaussian means and standard deviations, $\mu_{\mathcal{X}_i} \in \mathcal{U}$, $\sigma_{\mathcal{X}_i} \in [\epsilon, 0.25 + \epsilon]$, for each of M bins $\mathcal{X}_i \subseteq \mathcal{X}$. Major actions $u_t^{0,N}$ follow categorical or diagonal Gaussian distributions for discrete or continuous $\mathcal{U}^0$.

We use two hidden layers of 256 nodes and $\tanh$ activations for the neural networks of the policies. The neural network policy outputs parameters of a diagonal Gaussian over the major action u^0 and matrices U as discussed above. In the discrete Beach scenario below, the neural network instead outputs a categorical distribution using a final softmax layer. We used no GPUs and around 300,000 CPU core hours on Intel Xeon Platinum 9242 CPUs. Optimal transport costs are computed using POT [222]. Our M3FC MDP implementation follows the gym interface [223], while the implementation of MARL as in the following fulfills RLlib interfaces [76]. The RL implementations in our work are based on MARLlib 1.0 [91] (MIT license), which uses RLlib 1.8 [76] (Apache-2.0 license) with hyperparameters in Table E.1, and otherwise default settings.

4.2.2.3 *Comparison to MARL*

The M3FMARL algorithm falls into the paradigm of CTDE [40], as we sample a single central M3FC MDP action during training, but enable decentralized execution by sampling π'_t separately on each agent instead. For instance, when converged to a deterministic M3FC policy (of which an optimal one is guaranteed to exist by Theorem 4.2.1), the M3FC action is always trivially equal for all agents.

Since we also consider continuous minor agent action spaces in our experiments, we compare against PG methods for MARL. In particular, we firstly consider IPPO, as PPO with independent learning [86] and parameter sharing [90], and secondly also MAPPO with centralized critics. The latter has repeatedly shown strong state-of-the-art performance in cooperative MARL [42, 87, 88]. We also separate major and minor agent policies for improved performance of IPPO / MAPPO. For comparison, we use the same observations for the policy input as in M3FMARL. The policy network architectures match, and the same PPO implementation and hyperparameters are shared with M3FPPO in Table E.1. Minor agents are additionally allowed to observe their own states. More details can be found in Appendix E.18.

4.2.3 *Experiments*

In this section, we demonstrate the performance of M3FPPO on illustrative, practical problems. Unless noted otherwise, we use $M = 49$ bins ($M = 7$ in *Potential*), train for around 24 hours, and train M3FPPO on the finite-agent system Eq. (4.2.29) with $N = 300$ minor agents unless noted otherwise (similar results for less agents in Appendix E.18). Full descriptions and additional experiments and discussions are in Appendix E.18.

4.2.3.1 *Problems*

To verify the usefulness of M3FMARL whenever the M3FC model Eq. (4.2.29) is accurate, we consider 5 benchmark tasks that fulfill the M3FC modelling assumptions. To begin, the simple two Gaussian (**2G**) problem has no major agent and is equipped with a time-dependent major state: A periodic, time-variant mixture of two Gaussians μ_t^* – the major state – is noisily observed analogously to μ_t^N via $M = 49$ bins. Minor agents should then track the mixture distribution over time, which can find application for example in UAV-based cellular coverage of dynamic users [224]. In the

Formation problem, we extend such formation control with major agents. In addition to 2G, one added major agent tracks a moving target. Meanwhile, minor agents instead track a formation around the dynamic major agent, see e.g. [225] for applications. The **Beach** bar process is a studied classic [127, 226], where minor agents minimize their distances to a bar and additionally avoid crowded areas. Here, the bar moves on a discrete torus. The **Foraging** problem is archetypal of swarm intelligence [227], and has agents forage randomly generated foraging areas. In particular, we can consider the logistics scenario depicted in Figure 4.5, where a major package truck moves in a restricted space (roads) while minor drones collect packages for urban parcel delivery [228]. Drones fill up at package "foraging" areas, and unload near the major agent. Lastly, in the **Potential** problem, minor agents can generate a potential landscape, the gradient of which pushes the major agent – e.g., a large object affected by magnetic active matter [229] – to be delivered to a variable target.

4.2.3.2 *Evaluation*

In Figure 4.9, we see that M3FPPO learning is stable, as M3FPPO reduces hard-to-analyze MARL to single-agent RL, avoiding pathologies of MARL such as non-stationarity of multi-agent learning, or the combinatorial complexity over numbers of agents.

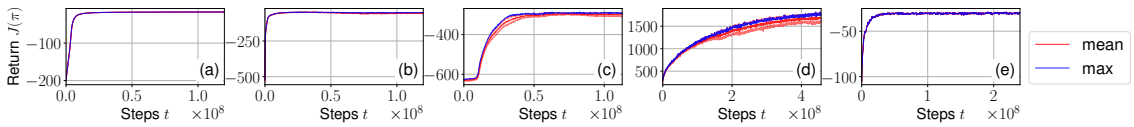


FIGURE 4.9: Training curves (mean episode return) of M3FPPO (red), with shaded standard deviation, and maximum (blue) over all three trials (two for Foraging). (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

In Figure 4.10, we find similar success in directly training M3FPPO for small N instead of transferring from high N. We conclude that M3FPPO remains applicable even with as few as 5 agents.

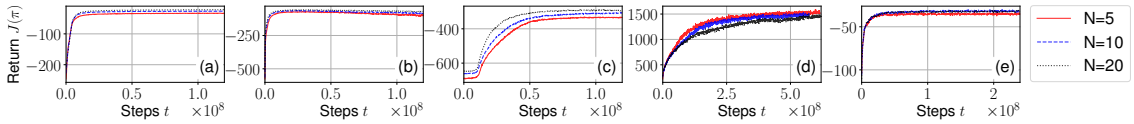


FIGURE 4.10: Training curves (mean episode return vs. time steps) of M3FPPO, trained on the finite systems with $N \in \{5, 10, 20\}$. (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

M3FPPO usually compares well against its A2C variant (M3FA2C) and IPPO / MAPPO, see Table 4.3 and Section E.18.2. Meanwhile, IPPO / MAPPO under the same hyperparameters as M3FPPO (large batch sizes, see Table E.1) can be more unstable and lead to worse results, see Figure 4.11.

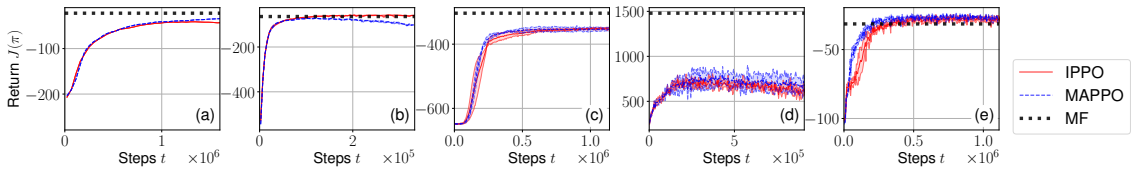


FIGURE 4.11: Comparing IPPO / MAPPO vs. results of M3FPPO (MF, ours), as in Figure 4.9 (no maxima, $N = 20$).

TABLE 4.3: Comparison of mean episode returns between best trained policies of standard MARL and M3FMARL methods on a system with $N = 20$ agents ($\pm$ 95% confidence interval, for a number of episodes as in Figure 4.13).

Problem	IPPO	MAPPO	M3FA2C	M3FPPO
2G	-43.9 $\pm$ 1.1	-26.0 $\pm$ 0.5	-30.6 $\pm$ 0.6	**-22.2** $\pm$ 0.56
Formation	**-51.1** $\pm$ 2.4	-101.1 $\pm$ 7.1	-79.2 $\pm$ 3.1	-63.9 $\pm$ 4.2
Beach	-350.3 $\pm$ 3.4	-342.9 $\pm$ 4.7	-424.8 $\pm$ 5.5	**-303.5** $\pm$ 3.4
Foraging	735.3 $\pm$ 46.4	803.9 $\pm$ 54.6	1398.0 $\pm$ 57.1	**1479.4** $\pm$ 36.3
Potential	-27.1 $\pm$ 1.4	**-26.7** $\pm$ 1.7	-50.4 $\pm$ 5.5	-31.3 $\pm$ 1.3

QUALITATIVE BEHAVIOR. In Figure 4.12, we observe successfully trained behavior in Beach and Foraging: In Beach, M3FPPO learns to accumulate up to 70% of agents on the bar, as more agents on the space lead to a suboptimal reduction in rewards. In Foraging, we find that agents successfully deplete foraging areas shown in the bottom left, moving on afterwards. Further, M3FPPO successfully learns to form mixtures of Gaussians in 2G, a Gaussian around a moving major agent successfully tracking its target in Formation, and similar success in pushing the major agent towards its target in Potential, see Section E.18.3.

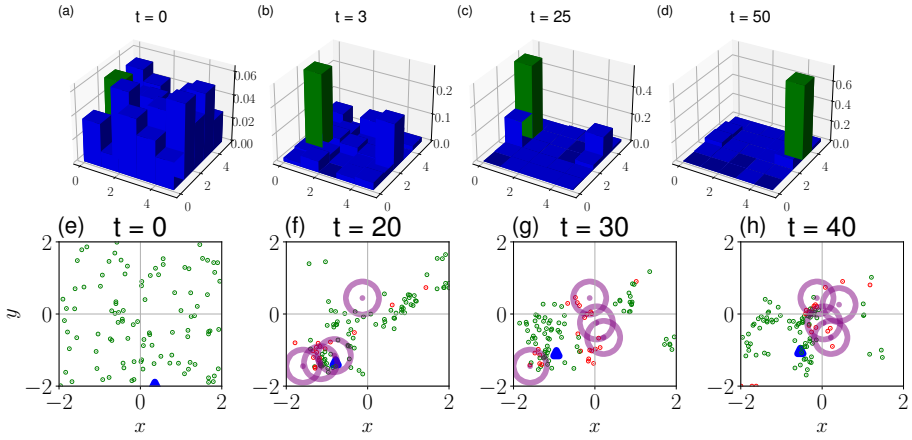


FIGURE 4.12: Qualitative visualization of M3FC in Beach (a-d), Foraging (e-h). (a-d): empirical MF, major agent & target in green; (e-h): blue / green triangle: major agent / target; green / red dots: less- / more-than-half encumbered minor agents; purple: current foraging areas.

QUANTITATIVE SUPPORT OF THEORY. In Figure 4.13, we *transfer* the trained M3FPPO policy to $N = 2, \ldots, 50$, comparing against the performance in the limit ($N = 500$). As N grows, the performance converges to the limit, supporting Theorem 4.2.2 and Corollary 4.2.1. Any sufficiently large system has the same limiting performance as predicted by the theory. We thus have empirical support for scalability, and also transferability between varying numbers of minor agents.

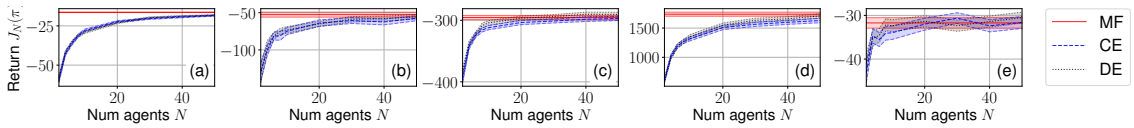


FIGURE 4.13: Mean episode return of M3FC policies in finite systems as in Figure 4.9, (a-c) 100, (d) 300 or (e) 500 trials, 95% confidence interval. MF: CE for $N = 500$; CE / DE: centralized / decentralized execution.

COMPARISON TO MARL. Comparing Figures 4.9 and 4.11 and Table 4.3, we see that (i) by experience sharing, standard MARL can be more sample-efficient, as each step gives N samples instead of just one; and (ii) M3FPPO matches or outperforms IPPO and MAPPO, despite having significantly less control over minor agent actions: All minor agents in a bin (with similar minor agent states) use the same action distributions, which suffices for strong results.

DECENTRALIZED EXECUTION. Lastly, decentralized execution by agent-wise randomization – i.e. sampling M3FC actions per agent instead of a single shared, correlated M3FC action – has little to no effect, and can even marginally improve performance, see e.g., Beach in Figure 4.13(c). Figure 4.13 verifies the performance of M3FMARL as a CTDE method.

4.2.4 *Summary*

We have proposed a generalization of MDPs and MFC, enabling tractable state-of-the-art MARL on general many-agent systems, with both theoretical and empirical support. Beyond the current model and its optimality guarantees, one could work on refined approximations [230], and local interactions [92]. Algorithmically, M3FC MDP actions $\mathcal{H}(\mu)$ could move beyond binning $\mathcal{X}$ to gain performance, e.g. via kernels as in the following Section 4.3. Lastly, one may try to quantify convergence to the rate $\mathcal{O}(1/\sqrt{N})$ for non-finite $\mathcal{X}$, as the current proof strategy would need hard-to-verify or unrealistic d_Σ-Lipschitzness.

4.3 MEAN FIELD CONTROL UNDER PARTIAL INFORMATION

MARL remains a challenge in terms of decentralization, partial observability and scalability to many agents. Meanwhile, collective behavior requires resolution of the aforementioned challenges, and remains of importance to many state-of-the-art applications such as active matter physics, self-organizing systems, opinion dynamics, and biological or robotic swarms. Here, MARL via MFC offers a potential solution to scalability, but fails to consider decentralized and partially observable systems. In this work, we enable decentralized behavior of agents under partial information by proposing novel models for Decentralized Partially-Observable Mean Field Control (Dec-POMFC), a broad class of problems with permutation-invariant agents allowing for reduction to tractable single-agent MDP with single-agent RL solution. We provide rigorous theoretical results, including a dynamic programming principle, together with optimality guarantees for Dec-POMFC solutions applied to finite swarms of interest. Algorithmically, we propose Dec-POMFC-based policy gradient methods for MARL via centralized training and decentralized execution, together with policy gradient approximation guarantees. In addition, we improve upon state-of-the-art histogram-based MFC by kernel methods, which is of separate interest also for fully observable MFC. We evaluate numerically on representative collective behavior tasks such as adapted Kuramoto and Vicsek swarming models, being on par with state-of-the-art MARL. Overall, our framework takes a step towards RL-based engineering of artificial collective behavior via MFC. The material presented in this section is based upon our work [2].

COLLECTIVE BEHAVIOR AND PARTIAL OBSERVABILITY. Of practical interest is the design of simple local interaction rules in order to fulfill global, cooperative objectives by emergence of global behavior [231]. For example, intelligent and self-organizing robotic swarms provide engineering applications such as Internet of Things or precision farming, for which a general design framework remains elusive [232, 233]. Other domains include group decision-making and opinion dynamics [234], biomolecular self-assembly [235] and active matter [236, 237] such as self-propelled nano-particles [238], microswimmers [239], etc. [231]. Overall, there is a need for scalable MARL under strong decentralization and partial information.

SCALABLE AND PARTIALLY OBSERVABLE MARL. Despite its many applications, decentralized cooperative control remains a difficult problem in MARL [40], especially if coupled with the simultaneous requirement of scalability. Recent scalable MARL methods include graphical decompositions [92, 240] amongst others [40]. However, most remain limited to full observability [240]. One line of algorithms applies pairwise MF approximations over neighbors [120], which has yielded decentralized, partially observable extensions [241, 242]. Relatedly, MARL based on MFGs (non-cooperative) and MFC (cooperative) focus on a broad class of systems with many exchangeable agents. While the theory for MFG is developed [79, 195, 243], to the best of our knowledge, neither MFC-based MARL algorithms nor discrete-time MFC have been proposed under *partial information and decentralization*, except in special linear-quadratic cases [244, 245]. Further, MFGs have been useful for analyzing emergence of collective behavior [129, 209], but less for "engineering" collective behavior to achieve global objectives as in MFC, which is our focus. This is in contrast to *rational, selfish* agents, as a decomposition of global objectives into per-agent rewards is non-trivial [246, 247]. Beyond scalability to many agents, general MFC for MARL is also not yet scalable to *high-dimensional* state-actions due to discretization of the simplex [104, 107], except in linear-quadratic models [248, 249]. Instead, we consider general discrete-time MFC and scale to higher dimensions via kernels. We note that our model has a similar flavor to TD-POMDPs [250], as the MF also abstracts influence from all other agents. However, both Dec-POMFC and TD-POMDP

address different types of problems, as the latter considers local per-agent states, while the MF in the former is both a globally, jointly defined state between all agents and influenced by all agents.

OUR CONTRIBUTION. A *tractable* framework for *cooperative* control, that can handle *decentralized, partially observable* systems, is missing. By the preceding motivation, we propose such a framework as illustrated in Figure 4.14. Our contributions may be summarized as (i) proposing the first discrete-time MFC model with decentralized and partially observing agents; (ii) providing accompanying approximation theorems, reformulations to a tractable single-agent MDP, and novel optimality results over equi-Lipschitz policies; (iii) establishing a MARL algorithm with policy gradient guarantees; and (iv) presenting kernel-based MFC parametrizations of separate interest for general, higher-dimensional MFC. The algorithm is verified on classical collective swarming behavior models, and compared against standard MARL. Overall, our framework steps toward tractable RL-based engineering of artificial collective behavior for large-scale multi-agent systems.

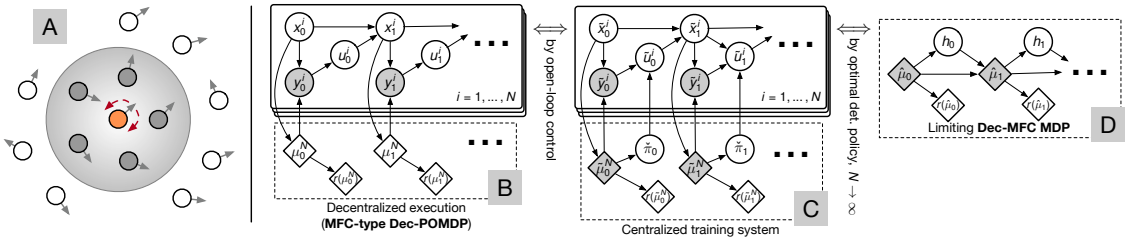


FIGURE 4.14: The partially-observable MFC model. A: Partially-observable Vicsek problem: agents must align headings (arrows), but observe only partial information (e.g. heading distribution in grey circle for orange agent). B: The decentralized model as a graphical model (grey: observed variables). C: In centralized training, we also observe the MF, guiding the learning of upper-level actions $\check{\pi}$. D: The solved limiting MDP.

4.3.1 *Decentralized Partially Observable Mean Field Control*

In this section, we introduce the motivating finite MFC-type decentralized partially observable control problem, as a special case of cooperative, general Dec-POMDPs [43, 46]. We then proceed to simplify in three steps of (i) taking the infinite-agent limit, (ii) relaxing partial observability during training, and (iii) correlating agent actions during training, in order to arrive at a tractable MDP with optimality guarantees, see also Figures 4.14 and 4.15. For proofs see Appendices F.1 to F.16.

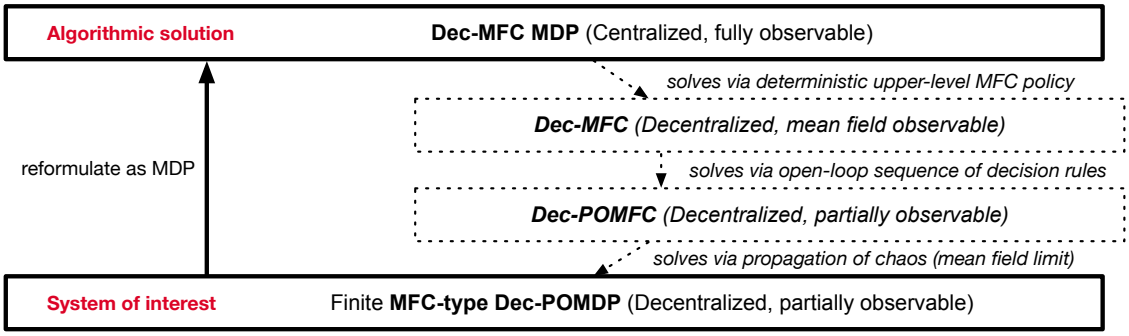


FIGURE 4.15: Reformulation of MFC-type Dec-POMDPs as an MDP. Three steps of approximation (MF limit, open-loop control, and MDP reformulation) allow us to reformulate the broad class of MFC-type Dec-POMDP to a tractable Decentralized Mean-Field-Observable Mean Field Control (Dec-MFC) MDP.

In a nutshell, Dec-POMDPs are hard, and hence we *reformulate* into the Dec-POMFC, for which we develop a new theory for optimality of Dec-POMFC solutions in the finite Dec-POMDP. The solution of Dec-POMFC itself also remains hard, because its MDP is not just continuous, but *infinite-dimensional* for continuous state-actions. The MDP is later addressed in Section 4.3.2 by (i) kernel parametrizations and (ii) approximate policy gradients on the finite Dec-POMDP (Theorem 4.3.3).

4.3.1.1 *MFC-Type Cooperative Multi-Agent Control*

To begin, we define the finite Dec-POMDP of interest, which is assumed to be MFC-type. In other words, (i) agents are **permutation invariant**, i.e. only the overall distribution of agent states matters, and (ii) agents observe only part of the system. We assume agents $i \in [N] \coloneqq \{1, \dots, N\}$ endowed with random states x_t^i, observations y_t^i and actions u_t^i at times $t \in \mathcal{T} \coloneqq \mathbb{N}$ from compact metric state, observation and action spaces $\mathcal{X}, \mathcal{Y}, \mathcal{U}$ (finite or continuous). Agent dynamics depend on other agents only via the empirical MF $\mu_t^N \coloneqq \frac{1}{N} \sum_{i \in [N]} \delta_{x_t^i}$. Policies are memory-less and shared by all agents, archetypal of collective behavior under simple rules [251], and of interest to compute-constrained agents, including e.g. nano-particles or small robots. Optionally, memory and history-dependence can be integrated into the state, see Appendix F.2. Agents act according to policy $\pi \in \Pi$ from a class $\Pi \subseteq \mathcal{P}(\mathcal{U})^{\mathcal{Y} \times \mathcal{T}}$ of policies, with spaces of probability measures $\mathcal{P}(\cdot)$, equipped with the 1-Wasserstein metric W_1 [214]. Starting with initial distribution μ_0, $x_0^i \sim \mu_0$, the **MFC-type Dec-POMDP** dynamics are

$$y_t^i \sim p_{\mathcal{Y}}(y_t^i \mid x_t^i, \mu_t^N), \quad u_t^i \sim \pi_t(u_t^i \mid y_t^i), \quad x_{t+1}^i \sim p(x_{t+1}^i \mid x_t^i, u_t^i, \mu_t^N) \tag{4.3.33}$$

for all $(i, t) \in [N] \times \mathcal{T}$, with transition kernels $p \colon \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$, $p_{\mathcal{Y}} \colon \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$, objective $J^N(\pi) = \mathbb{E}[\sum_{t \in \mathcal{T}} \gamma^t r(\mu_t^N)]$ to maximize over $\pi \in \Pi$ under reward function $r \colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, and discount factor $\gamma \in (0, 1)$. Results generalize to finite horizons, average per-agent rewards $r_{\mathrm{per}} \colon \mathcal{X} \to \mathbb{R}$, $r(\mu_t^N) = \int r_{\mathrm{per}} \mathrm{d}\mu_t^N$, and joint state-observation-action MFs via enlarged states.

Since general Dec-POMDPs are hard [46], our model establishes a tractable special case of high generality. Standard MFC already covers a broad range of applications, e.g. see surveys for finance [64] and engineering [65] applications, which can now be handled under partial information. In addition, many classical, inherently partially observable models are covered by MFC-type Dec-POMDPs, such as the Kuramoto or Vicsek models in Section 4.3.3, where many-agent convergence is known as propagation of chaos [126].

4.3.1.2 *Limiting Mean Field Control System*

In order to achieve tractability for large multi-agent systems, the first step is to take the infinite-agent limit. By a LLN, this allows us to describe large systems only by the MF μ_t. Consider a representative agent as in Eq. (4.3.33) with states $x_0 \sim \mu_0, x_{t+1} \sim p(x_{t+1} \mid x_t, u_t, \mu_t)$, observations $y_t \sim p_{\mathcal{Y}}(y_t \mid x_t, \mu_t)$ and actions $u_t \sim \pi_t(u_t \mid y_t)$. Then, its state probability law replaces the empirical state distribution, informally $\mu_t = \mathcal{L}(x_t) \equiv \lim_{N \to \infty} \mu_t^N$. Looking only at the MF, we hence obtain the Dec-POMFC system

$$\mu_{t+1} = \mathcal{L}(x_{t+1}) = T(\mu_t, \pi_t) \coloneqq \iiint p(x, u, \mu_t) \pi_t(\mathrm{d}u \mid y) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu_t) \mu_t(\mathrm{d}x) \tag{4.3.34}$$

by deterministic transitions $T \colon \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{U})^{\mathcal{Y}} \to \mathcal{P}(\mathcal{X})$ and objective $J(\pi) = \sum_{t=0}^{\infty} \gamma^t r(\mu_t)$.

Approximation guarantees. Under mild continuity assumptions, the Dec-POMFC model in Eq. (4.3.34) constitutes a good approximation of large-scale MFC-type Dec-POMDP in Eq. (4.3.33) with many agents.

Assumption 4.3.1a. *The transitions p, $p_{\mathcal{Y}}$ and rewards r are Lipschitz with constants L_p, $L_{p_{\mathcal{Y}}}$, L_r.*

Assumption 4.3.1b. *The class of policies Π is the set of all L_Π-Lipschitz policies for some $L_\Pi > 0$, i.e. for all $t \in \mathcal{T}$ and $\pi \in \Pi$, we have that $\pi_t \colon \mathcal{Y} \to \mathcal{P}(\mathcal{U})$ is L_Π-Lipschitz. Alternatively, we may assume unrestricted policies and that (i) observations only depend on the agent state, and (ii) $\mathcal{X}$ is finite.*

Lipschitz continuity of the model is commonly assumed [104, 105, 195], and in general at least (uniform) continuity is required: Consider a counterexample with uniform initial μ_0 over states A, B. If dynamics, observations, or rewards jump between regimes at $\mu(A) = \mu(B) = 0.5$, the finite system will randomly experience all regimes, while limiting MFC experiences only the regime at $\mu(A) = \mu(B) = 0.5$. Meanwhile, Lipschitz policies are not only standard in MFC literature [105, 116] by neural networks (NNs) [220], but also fulfilled for finite $\mathcal{Y}$ trivially without loss of generality ($L_\Pi \coloneqq \operatorname{diam}(\mathcal{U})$), and for continuous $\mathcal{Y}$ by kernel parametrizations in Section 4.3.2. We extend MFC approximation theorems [1, 104, 105] to partial observations and compact spaces.

Theorem 4.3.1. *Fix an equicontinuous family of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{P}(\mathcal{X})}$. Under Assumptions 4.3.1a to 4.3.1b, the MF converges in the sense of $\sup_{\pi\in\Pi} \sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_t^N) - f(\mu_t)\right|\right] \to 0$ at all times $t \in \mathcal{T}$.*

The approximation rate is $\mathcal{O}(1/\sqrt{N})$ for finite state-actions, using equi-Lipschitz $\mathcal{F}$ (Appendix F.1). Hence, the easier Dec-POMFC simplifies otherwise hard Dec-POMDPs. Indeed, we later show that such optimal Lipschitz Dec-POMFC policies are guaranteed to exist via closedness of joint-measures under equi-Lipschitz kernels (Appendix F.8), see Propositions 4.3.1 and 4.3.2 and Theorem 4.3.2 later.

Corollary 4.3.1. *Under Assumptions 4.3.1a to 4.3.1b, any optimal Dec-POMFC policy $\pi \in \arg\max_{\pi'\in\Pi} J(\pi')$ is ε-optimal in the MFC-type Dec-POMDP, $J^N(\pi) \geq \sup_{\pi'\in\Pi} J^N(\pi') - \varepsilon$, with $\varepsilon \to 0$ as $N \to \infty$.*

4.3.1.3 *Rewriting Policies with Mean Field Observations*

Now introducing the next system for reduction to an MDP, writing $\bar{\mu}$, $\bar{\pi}$ etc., let policies depend also on $\bar{\mu}_t$, i.e. policies "observe" the MF. While we could reason that agents might observe the MF or use filtering to estimate it [252], more importantly, the limiting MF is *deterministic*. Therefore, w.l.o.g. we obtain the Dec-MFC dynamics

$$\bar{\mu}_{t+1} = T(\bar{\mu}_t, \bar{\pi}_t(\bar{\mu}_t)) \coloneqq \iiint p(x, u, \bar{\mu}_t)\bar{\pi}_t(\mathrm{d}u \mid y, \bar{\mu}_t)p_{\mathcal{Y}}(\mathrm{d}y \mid x, \bar{\mu}_t)\bar{\mu}_t(\mathrm{d}x), \tag{4.3.35}$$

with shorthand $\bar{\pi}_t(\bar{\mu}_t) = \bar{\pi}_t(\cdot \mid \cdot, \bar{\mu}_t)$, initial $\bar{\mu}_0 = \mu_0$ and according objective $\bar{J}(\bar{\pi}) = \sum_{t=0}^{\infty} \gamma^t r(\bar{\mu}_t)$ to optimize over (now MF-dependent) policies $\bar{\pi} \in \bar{\Pi} \subseteq \mathcal{P}(\mathcal{U})^{\mathcal{Y}\times\mathcal{P}(\mathcal{X})\times\mathcal{T}}$.

Deterministic open-loop control transforms optimal Dec-MFC policies $\bar{\pi} \in \arg\max_{\bar{\pi}'\in\bar{\Pi}} \bar{J}(\bar{\pi}')$ into optimal *Dec-POMFC* policies $\pi \in \arg\max_{\pi\in\Pi} J(\pi)$ with decentralized execution, and vice

versa: For given $\bar{\pi}$, compute deterministic MFs $(\bar{\mu}_0, \bar{\mu}_1, \dots)$ via Eq. (4.3.35) and let $\pi = \Phi(\bar{\pi})$ by $\pi_t(\mathrm{d}u \mid y) = \bar{\pi}(\mathrm{d}u \mid y, \bar{\mu}_t)$. Analogously, represent $\pi \in \Pi$ by $\bar{\pi} \in \bar{\Pi}$ with constant $\bar{\pi}_t(\nu) = \pi_t$ for all ν.

Proposition 4.3.1. *For any $\bar{\pi} \in \bar{\Pi}$, define $(\bar{\mu}_0, \bar{\mu}_1, \dots)$ as in Eq.* (4.3.35)*. Then, for $\pi = \Phi(\bar{\pi}) \in \Pi$, we have $\bar{J}(\bar{\pi}) = J(\pi)$. Inversely, for any $\pi \in \Pi$, let $\bar{\pi}_t(\bar{\nu}) = \pi_t$ for all $\bar{\nu}$, then again $\bar{J}(\bar{\pi}) = J(\pi)$.*

Corollary 4.3.2. *Optimal Dec-MFC policies $\bar{\pi} \in \arg\max_{\bar{\pi}' \in \bar{\Pi}} \bar{J}(\bar{\pi}')$ yield optimal Dec-POMFC policies $\Phi(\bar{\pi})$, i.e. $J(\Phi(\bar{\pi})) = \sup_{\pi' \in \Pi} J(\pi')$.*

Knowing initial μ_0 is often realistic, as deployment is commonly for well-defined problems of interest. Even then, knowing μ_0 is not strictly necessary (Section 4.3.3). In contrast to standard deterministic open-loop control, (i) agents have stochastic dynamics and observations, and (ii) agents randomize actions instead of playing a trajectory, still leading to quasi-deterministic MFs by the LLN.

4.3.1.4 *Reduction to Dec-MFC MDPs*

Lastly, we reformulate as an MDP with more tractable theory and algorithms, writing $\hat{\mu}$, $\hat{\pi}$ etc. The recent MFC MDP [103, 107, 201] reformulates *fully observable* MFC as MDPs with higher-dimensional state-actions. Similarly, we reduce Dec-MFC to an MDP with joint state-observation-action distributions as its MDP actions. The **Dec-MFC MDP** has states $\hat{\mu}_t \in \mathcal{P}(\mathcal{X})$ and actions $h_t \in \mathcal{H}(\hat{\mu}_t) \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{U})$ in the set of joint $h_t = \hat{\mu}_t \otimes p_{\mathcal{Y}}(\hat{\mu}_t) \otimes \check{\pi}_t$ under any L_Π-Lipschitz policy $\check{\pi}_t \in \mathcal{P}(\mathcal{U})^{\mathcal{Y}}$. Here, $\nu \otimes K$ is the product measure of measure ν and kernel K, and νK is the measure $\nu K = \int K(\cdot \mid x)\nu(\mathrm{d}x)$. For $\check{\pi}_t \in \mathcal{P}(\mathcal{U})^{\mathcal{Y}}$, $\mu_{xy} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, we write $\mu_{xy} \otimes \check{\pi}_t$ by letting $\check{\pi}_t$ constant on $\mathcal{X}$. In other words, the desired joint h_t results from all agents replacing the previous system's policy $\bar{\pi}_t$ by lower-level policy $\check{\pi}_t$, which may be reobtained from h_t a.e.-uniquely by disintegration [253]. Equivalently, identify $\mathcal{H}(\mu)$ with μ and classes of $\check{\pi}_t$ yielding the same joint, and in practice we parametrize $\check{\pi}_t$. Thus, we obtain the MDP dynamics

$$h_t \sim \hat{\pi}(\hat{\mu}_t), \quad \hat{\mu}_{t+1} = \hat{T}(\hat{\mu}_t, h_t) \coloneqq \iiint p(x, u, \hat{\mu}_t) h_t(\mathrm{d}x, \mathrm{d}y, \mathrm{d}u) \tag{4.3.36}$$

for Dec-MFC MDP policy $\hat{\pi} \in \hat{\Pi}$ and objective $\hat{J}(\hat{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(\hat{\mu}_t)\right]$. The Dec-MFC MDP policy $\hat{\pi}$ is "upper-level", as we sample h_t from $\hat{\pi}$, to apply the lower-level policy $\check{\pi}_t[h_t]$ to all agents.

GUIDANCE BY MF DEPENDENCE. Intuitively, the MF *guides* policy search in potentially hard, decentralized problems, and reduces to a single-agent MDP where we make some existing theory compatible. First, we formulate a DPP, i.e. exact solutions by Bellman's equation for the value function $V(\mu) = \sup_{h \in \mathcal{H}(\mu)} r(\mu) + \gamma V(\hat{T}(\mu, h))$ [69]. Here, a central theoretical novelty is closedness of joint measures under equi-Lipschitz policies (Appendix F.8). Concomitantly, we obtain optimality of stationary deterministic $\hat{\pi}$. For technical reasons, only here we assume Hilbertian $\mathcal{Y}$ (e.g. finite or Euclidean) and finite $\mathcal{U}$.

Assumption 4.3.2. *The observations $\mathcal{Y}$ are a metric subspace of a Hilbert space. Actions $\mathcal{U}$ are finite.*

Theorem 4.3.2. *Under Assumptions 4.3.1a to 4.3.1b and Assumption 4.3.2, there exists an optimal stationary, deterministic policy $\hat{\pi}$ for the Dec-MFC MDP, with $\hat{\pi}(\mu) \in \arg\max_{h \in \mathcal{H}(\mu)} r(\mu) + \gamma V(\hat{T}(\mu, h))$.*

DECENTRALIZED EXECUTION. Importantly, guidance by MF is only for training and not execution. An optimal upper-level policy $\hat{\pi} \in \arg\max_{\hat{\pi}' \in \hat{\Pi}} \hat{J}(\hat{\pi})$ is optimal also for the initial system, if it is deterministic, and an optimal one exists by Theorem 4.3.2. The lower-level policies $\bar{\pi}_t \equiv \check{\pi}_t$ are obtained by inserting the sequence of MFs $\hat{\mu}_0, \hat{\mu}_1, \ldots$ into $\hat{\pi}$, and remain non-stationary stochastic policies.

Proposition 4.3.2. *For deterministic $\hat{\pi} \in \hat{\Pi}$, let $\hat{\mu}_t$ as in Eq.* (4.3.36) *and $\bar{\pi} = \Psi(\hat{\pi})$ by $\bar{\pi}_t(\nu) = \check{\pi}_t$ for all ν, then $\hat{J}(\hat{\pi}) = \bar{J}(\bar{\pi})$. Inversely, for $\bar{\pi} \in \bar{\Pi}$, let $\hat{\pi}_t(\nu) = \nu \otimes p_{\mathcal{Y}}(\nu) \otimes \bar{\pi}_t(\nu)$ for all ν, then $\hat{J}(\hat{\pi}) = \bar{J}(\bar{\pi})$.*

Note that the determinism of the *upper-level policy* is strictly necessary: A simple counterexample is a problem where agents should choose to aggregate to one state. If the upper-level policy randomly chooses between moving all agents to either A or B, then a corresponding random agent policy splits agents and fails to aggregate. At the same time, randomization of *agent actions* remains necessary for optimality, as the problem of equally spreading would require uniformly random agent actions.

COMPLEXITY. Tractability of multi-agent control heavily depends on information structure [254]. General Dec-POMDPs have doubly-exponential complexity (NEXP) [46] and are harder than fully observable control (PSPACE, [81]). In contrast, Dec-POMFC surprisingly imposes little additional complexity over standard MFC, as the MFC MDP remains deterministic in the absence of common noise correlating agents [178]. An analysis with common noise is possible, e.g., if observing the MF, but out of scope.

4.3.2 *Partially Observable Mean Field Multi-Agent Reinforcement Learning*

All that remains is to solve Dec-MFC MDPs. As we obtain continuous Dec-MFC MDP states and actions even for finite $\mathcal{X}, \mathcal{Y}, \mathcal{U}$, and infinite-dimensional ones for continuous $\mathcal{X}, \mathcal{Y}, \mathcal{U}$, a value-based approach can be hard. Our PG approach allows finding simple policies for collective behavior, with emergence of global intelligent behavior described by rewards r, under arbitrary (Lipschitz) policies. For generality, we use NN upper-level and kernel lower-level policies. While lower-level (Lipschitz, [220]) NNs policies could be considered akin to hypernetworks [255], the resulting distributions over NN parameters as MDP actions are too high-dimensional and failed in our experiments. We directly solve finite-agent MFC-type Dec-POMDPs by solving the Dec-MFC MDP in the background. Indeed, the **theoretical optimality** of Dec-MFC MDP solutions is guaranteed over Lipschitz policies in Π.

Corollary 4.3.3. *Under Assumptions 4.3.1a to 4.3.1b, a deterministic Dec-MFC solution $\hat{\pi} \in \arg\max_{\hat{\pi}'} \hat{J}(\hat{\pi}')$ is ϵ-optimal in the Dec-POMDP with $\epsilon \to 0$ as $N \to \infty$, $J^N(\Phi(\Psi(\hat{\pi}))) \geq \sup_{\pi' \in \Pi} J^N(\pi') - \epsilon$.*

HISTOGRAM VS. KERNEL PARAMETRIZATIONS. Except for linear-quadratic algorithms [245, 248, 249], the only approach to learning MFC in continuous spaces $\mathcal{X} \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$ (and here $\mathcal{Y}$) is by partitioning and "discretizing" [104, 107]. Existing Q-Learning with kernel regression [104] is for *finite* states $\mathcal{X}$ with kernels on $\mathcal{P}(\mathcal{X})$, and learns on the MFC MDP. We allow *continuous* $\mathcal{Y}$ by kernels on $\mathcal{Y}$ itself, and learn on the finite-agent system. Unfortunately, partitions fail Lipschitzness

and hence approximation guarantees, even in standard MFC. Instead, we use kernel representations for MFs μ_t^N and lower-level policies $\check{\pi}_t$.

We represent $\mathcal{P}(\mathcal{X})$-valued MDP states μ_t^N not by counting agents in each bin, but instead mollify around each center $x_b \in \mathcal{X}$ of $M_{\mathcal{X}}$ bins $b \in [M_{\mathcal{X}}]$ using kernels. The result is Lipschitz and approximates histograms arbirtarily well [256, Theorem 1]. Hence, we obtain input logits $I_b = \int \kappa(x_b, x)\mu_t^N(\mathrm{d}x) = \frac{1}{N}\sum_{i\in[N]} \kappa(x_b, x_t^i)$ for some kernel $\kappa\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $b \in [M_{\mathcal{X}}]$. Output logits constitute mean and log-standard deviation of a diagonal Gaussian over parameter representations $\xi \in \Xi$ of $\check{\pi}_t$. We obtain Lipschitz $\check{\pi}_t$ by representing $\check{\pi}_t$ via $M_{\mathcal{Y}}$ points $y_b \in \mathcal{Y}$ such that $\check{\pi}_t(u \mid y) = \sum_{b\in[M_{\mathcal{Y}}]} \kappa(y_b, y)p_b(u) / \sum_{b\in[M_{\mathcal{Y}}]} \kappa(y_b, y)$. Here, we consider L_λ-Lipschitz maps λ from parameters $\xi \in \Xi$ to distributions $p_b = \lambda_b(\xi) \in \mathcal{P}(\mathcal{U})$ with compact parameter space Ξ, and for kernels choose Radial Basis Function (RBF) kernels $\kappa(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$ with some bandwidth $\sigma^2 > 0$.

Proposition 4.3.3. *Under RBF kernels κ, for any ξ and continuous $\mathcal{Y}$, the lower-level policies $\Lambda(\xi)(u \mid y) \coloneqq \sum_{b\in[M_{\mathcal{Y}}]} \kappa(y_b, y)\lambda_b(u \mid \xi) / \sum_{b\in[M_{\mathcal{Y}}]} \kappa(y_b, y)$ are L_Π-Lipschitz in y, as in Assumption 4.3.1b, if $\sigma^2 \exp^2\left(-\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})^2\right) \geq \frac{1}{L_\Pi}\operatorname{diam}(\mathcal{Y})\operatorname{diam}(\mathcal{U})\max_{y\in\mathcal{Y}}\|y\|$, and such $\sigma^2 > 0$ exists.*

Proposition 4.3.3 ensures Assumption 4.3.1b. To achieve optimality by Corollary 4.3.3, deterministic policies commonly result from convergence of stochastic PGs by taking action means, or can also be guaranteed using deterministic PGs [257, 258]. Beyond allowing for (i) theoretical guarantees, and (ii) finer control over agent actions, another advantage of kernels is (iii) the improved complexity over histograms. Even a histogram with only 2 bins per dimension requires 2^d bins in d-dimensional spaces, while kernel representations may place e.g. 2 points per dimension, improving upon the otherwise exponential complexity, see also Appendix F.17 for empirical support.

DIRECT MULTI-AGENT REINFORCEMENT LEARNING ALGORITHM. Applying RL to the Dec-MFC MDP is satisfactory for solutions only under known MFC models. Importantly, we do not always have access to the model, and even if we do, parametrizing MFs in arbitrary compact $\mathcal{X}$ is hard. Instead, it is more practical and tractable to train on a finite system. Our direct MARL approach hence trains on a finite N-agent MFC-type Dec-POMDP of interest, in a model-free manner. In order to exploit the underlying MDP, our algorithm assumes *during training* that (i) the MF is observed, and (ii) agents can correlate actions (e.g. centrally, or sharing seeds). Therefore, the finite system Eq. (4.3.33) is adjusted for training by correlating agent actions on a single centrally sampled lower-level policy $\check{\pi}_t$. Now write $\hat{\pi}^\theta(\xi_t \mid \tilde{\mu}_t^N)$ as density over parameters $\xi_t \in \Xi$ under a base measure (discrete, Lebesgue). Substituting ξ_t as actions parametrizing h_t in the MDP Eq. (4.3.36), e.g. by using RBF kernels, yields the **centralized training** system as seen in Figure 4.14 for stationary policy $\hat{\pi}^\theta$ parametrized by θ,

$$\begin{aligned} &\check{\pi}_t = \Lambda(\tilde{\xi}_t), \quad \tilde{\xi}_t \sim \hat{\pi}^\theta(\tilde{\mu}_t^N), \\ \tilde{y}_t^i \sim p_{\mathcal{Y}}(\tilde{y}_t^i \mid \tilde{x}_t^i, \tilde{\mu}_t^N), \quad &\tilde{u}_t^i \sim \check{\pi}_t(\tilde{u}_t^i \mid \tilde{y}_t^i), \quad \tilde{x}_{t+1}^i \sim p(\tilde{x}_{t+1}^i \mid \tilde{x}_t^i, \tilde{u}_t^i, \tilde{\mu}_t^N), \quad \forall i \in [N]. \end{aligned} \tag{4.3.37}$$

POLICY GRADIENT APPROXIMATION. Since we train on a finite system, it is not immediately clear whether centralized training really yields the PG for the underlying Dec-MFC MDP, also in existing literature for learning MFC. We will show this practically relevant fact up to an approximation. The general PG for stationary $\hat{\pi}^\theta$ [221, 259] is $\nabla_\theta J(\hat{\pi}^\theta) = (1 - \gamma)^{-1}\mathbb{E}_{\mu\sim d_{\hat{\pi}^\theta}, \xi\sim\hat{\pi}^\theta(\mu)}\left[Q^\theta(\mu, \xi)\nabla_\theta \log \hat{\pi}^\theta(\xi \mid \mu)\right]$ with $Q^\theta(\hat{\mu}, \xi) = \mathbb{E}[\sum_{t=0}^\infty \gamma^t r(\hat{\mu}_t) \mid \hat{\mu}_0 = \mu, \xi_0 =$

$\xi]$ under parametrized actions ξ_t in Eq. (4.3.36), and using sums $d_{\hat{\pi}^\theta} = (1-\gamma)\sum_{t\in\mathcal{T}} \gamma^t \mathcal{L}_{\hat{\pi}^\theta}(\hat{\mu}_t)$ of laws of $\hat{\mu}_t$ under $\hat{\pi}^\theta$. Our approximation motivates MFC for MARL by showing that the *underlying background Dec-MFC MDP* is approximately solved under Lipschitz parametrizations, e.g. we normalize parameters ξ to finite action probabilities, or use bounded diagonal Gaussian parameters.

Assumption 4.3.3. *The policy $\hat{\pi}^\theta(\xi \mid \mu)$ and its log-gradient $\nabla_\theta \log \hat{\pi}^\theta(\xi \mid \mu)$ are $L_{\hat{\Pi}}$, $L_{\nabla\hat{\Pi}}$-Lipschitz in μ and ξ (or alternatively in μ for any ξ, and uniformly bounded). The parameter-to-distribution map is $\Lambda(\xi)(u \mid y) := \sum_b \kappa(y_b, y)\lambda_b(u \mid \xi) / \sum_b \kappa(y_b, y)$, with kernels κ and L_λ-Lipschitz $\lambda_b \colon \Xi \to \mathcal{P}(\mathcal{U})$.*

Theorem 4.3.3. *Centralized training on system Eq.* (4.3.37) *approximates the true gradient of the underlying Dec-MFC MDP, i.e. under RBF kernels κ as in Proposition 4.3.3, Assumptions 4.3.1a to 4.3.1b and Assumption 4.3.3, as $N \to \infty$,*

$$\left\| (1-\gamma)^{-1} \mathbb{E}_{\mu \sim d^N_{\hat{\pi}^\theta}, \xi \sim \hat{\pi}^\theta(\mu)} \left[\tilde{Q}^\theta(\mu, \xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \mu) \right] - \nabla_\theta J(\hat{\pi}^\theta) \right\| \to 0$$

with $d^N_{\hat{\pi}^\theta} = (1-\gamma)\sum_{t\in\mathcal{T}} \gamma^t \mathcal{L}_{\hat{\pi}^\theta}(\tilde{\mu}^N_t)$ and $\tilde{Q}^\theta(\mu, \xi) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(\tilde{\mu}^N_t) \mid \mu_0 = \mu, \xi_0 = \xi\right]$.

The value function $\tilde{Q}^\theta$ in the finite system is then substituted in actor-critic manner by on-policy and critic estimates. The Lipschitz conditions of $\hat{\pi}^\theta$ in Assumption 4.3.3 are fulfilled by Lipschitz NNs [105, 116, 220] and our parametrizations. The approximation is novel, building a foundation for MARL via MFC directly on a finite MARL problem. Our results also apply to fully observable MFC by $y_t = x_t$. Though gradient estimates allow convergence guarantees in finite MDPs (e.g. [92, Theorem 5]), Dec-MFC MDP state-actions are always non-finite. In practice, we use empirically more efficient PPO [73, 88] to obtain Decentralized Partially-Observable Mean Field PPO (Dec-POMFPPO) in Algorithm 3.

Algorithm 3 Dec-POMFPPO (during centralized training)

for iteration $n = 1, 2, \ldots$ **do**
  for time $t = 0, \ldots, B_{\text{len}} - 1$ **do**
    Sample central Dec-MFC MDP action $\check{\pi}_t = \Lambda(\xi_t)$, $\xi_t \sim \hat{\pi}^\theta(\tilde{\mu}^N_t)$.
    for agent $i = 1, \ldots, N$ **do**
      Sample per-agent action $\tilde{u}^i_t \sim \check{\pi}_t(\tilde{u}^i_t \mid \tilde{y}^i_t)$ for observation $\tilde{y}^i_t$.
    Perform actions, observe reward $r(\tilde{\mu}^N_t)$, next MF $\tilde{\mu}^N_{t+1}$, termination flag $d_{t+1} \in \{0, 1\}$.
  for updates $i = 1, \ldots, N_{\text{PPO}}$ **do**
    Sample mini-batch b, $|b| = b_{\text{len}}$ from data $B := ((\tilde{\mu}^N_t, \xi_t, r^N_t, d_{t+1}, \tilde{\mu}^N_{t+1}))_{t\geq 0}$.
    Update policy $\hat{\pi}^\theta$ via PPO loss $\nabla_\theta L_\theta$ on b, using GAE [260].
    Update critic $V^{\theta'}$ via critic L_2-loss $\nabla_{\theta'} L_{\theta'}$ on b.

During training, the algorithm (i) assumes to observe the MF, and (ii) samples only one centralized h_t. Knowledge of the MF during training aligns our framework with the popular CTDE paradigm. By Theorem 4.3.3, we may learn directly on the MFC-type Dec-POMDP system Eq. (4.3.33). During execution, decentralized policies suffice for near-optimality by Corollary 4.3.3 without agents knowing the MF or coordinating centrally. Decentralized training can also be achieved, if the MF is observable and all agents use the same seed to correlate their actions.

4.3.3 *Experiments*

In this section, we empirically evaluate our algorithm, comparing against IPPO and MAPPO with state-of-the-art performance [42, 88]. For comparison, we share hyperparameters and architectures between algorithms, see Appendices F.17 to F.19.

PROBLEMS. In the **Aggregation** problem we consider a typical continuous single integrator model, commonly used in the study of swarm robotics [261, 262]. Agents observe their own position noisily and should aggregate. The classical **Kuramoto** model is used to study synchronization of coupled oscillators, finding application not only in physics, including quantum computation and laser arrays [263], but also in diverse biological systems, such as neuroscience and pattern formation in self-organizing systems [237, 264]. Here, via partial observability, we consider a version where each oscillator can see the distribution of relative phases of its neighbors. Finally, we implement the Kuramoto model on a random geometric graph (e.g. [265]) via omitting movement in its independent generalization, the **Vicsek** model [231, 266]. Agents j have two-dimensional position p_t^j and current headings ϕ_t^j, to be controlled by their actions. The key metric of interest for both Kuramoto and Vicsek is polarization via the *polar order parameter* $R = |\frac{1}{N}\sum_j \exp(i\phi_t^j)|$. Here, R ranges from 0 – fully unsynchronized – to 1 – perfect alignment of agents. Experimentally, we consider various environments, such as the torus, Möbius strip, projective plane and Klein bottle. Importantly, agents only observe relative headings of others.

TRAINING RESULTS. In Figure 4.16 it is evident that the training process of MFC for many agents is relatively stable by guidance via MF and reduction to single-agent RL. In Appendix F.17, we also see similar results with significantly fewer agents and comparable to the results obtained with a larger number of agents. This observation highlights that the training procedure yields satisfactory outcomes, even in scenarios where the MF approximation may not yet be perfectly exact. These findings underscore the generality of the proposed framework and its ability to adapt across regimes.

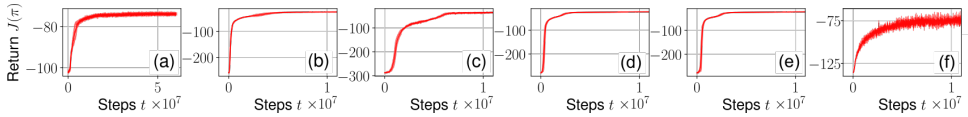


FIGURE 4.16: Dec-POMFPPO training curves (episode return) with shaded standard deviation over 3 seeds for $N = 200$ in (a) Aggregation; Vicsek on a (b): torus; (c): Möbius strip; (d): projective plane; (e): Klein bottle; and (f) Kuramoto on a torus.

On the same note, we see by comparison with Figure 4.17, that our method is usually on par with state-of-the-art IPPO and MAPPO for many agents, e.g. here $N = 200$, though with worse sample complexity.

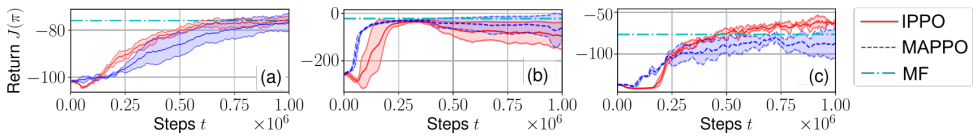


FIGURE 4.17: Training curves (episode return) with shaded standard deviation over 3 seeds and $N = 200$, in (a) Aggregation (box), (b) Vicsek (torus), (c) Kuramoto (torus). For comparison, we also plot the best return averaged over 3 seeds for Dec-POMFPPO in Figure 4.16 (MF).

VERIFICATION OF THEORY. In Figure 4.18, as the number of agents rises, the performance quickly tends to its limit, i.e. the objective converges, supporting Theorem 4.3.1 and Corollary 4.3.1, as well as applicability to arbitrarily many agents.

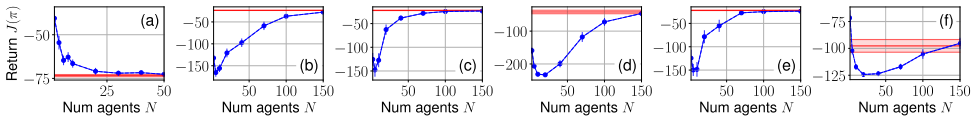


FIGURE 4.18: The performance of the best of 3 final MFC policies transferred to N-agent systems (in blue), with error bars for 95% confidence interval, averaged over 50 episodes, and compared against the performance in the training system (in red). Problems (a)-(f) and training are as in Figure 4.16.

Analogously, conducting open-loop experiments on our closed-loop trained system in Figure 4.19 demonstrates the robust generality of learned collective behavior with respect to the randomly sampled initial agent states, supporting Corollary 4.3.3 and Corollary 4.3.2.

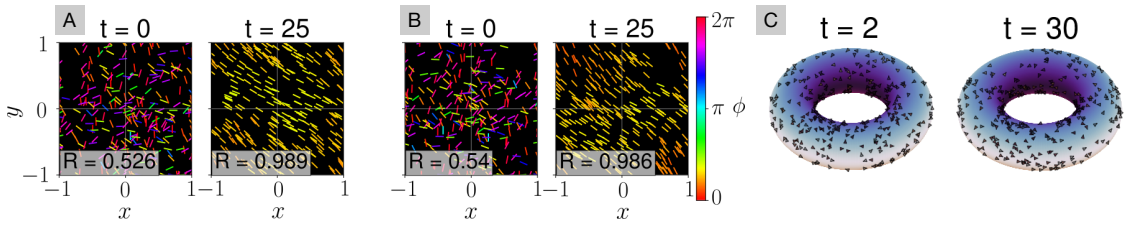


FIGURE 4.19: Qualitative behavior of Dec-POMFC in the Vicsek problem on the torus. A, B: For the Vicsek (torus) problem with forward velocity control, the open-loop behavior (B) shows little difference in performance of agents (rods, color indicating heading) over the closed-loop behavior (A). C: Visualization of agents (triangles) under the Vicsek model on the torus.

QUALITATIVE ANALYSIS. In the Vicsek model, as seen exemplarily in Figure 4.19 and Appendix F.17, the algorithm learns to align in various topological spaces. In all considered topologies, the polar order parameter surpasses 0.9, with the torus system even reaching a value close to 0.99. As for the angles at different iterations of the training process, as depicted in Figure 4.20, the algorithm gradually learns to form a concentrated cluster of angles. Note that the cluster center angle is not fixed, but rather changes over time. This behavior can not be observed in the classical Vicsek model, though extensions using more sophisticated equations of motion for angles have reported similar results [237]. For more details, see Appendices F.17 to F.19.

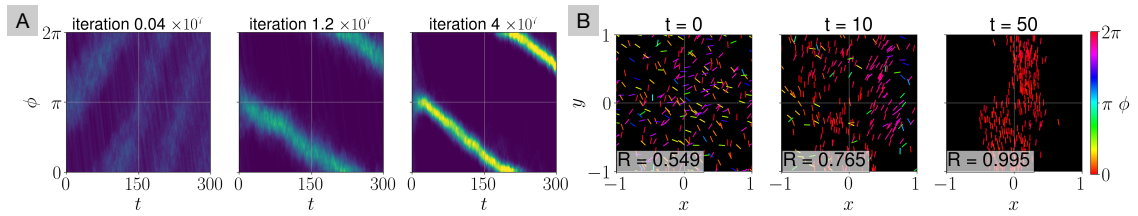


FIGURE 4.20: Qualitative behavior of Dec-POMFC in the Vicsek problem. A: Agent angle alignment in the Vicsek model on the torus, plotted as density over time; B: Alignment of agents in the Vicsek model on the projective plane, as in Figure 4.19.

Figure 4.20 and additional figures, with similar results for other topologies in Appendix F.17 illustrate the qualitative behavior observed across the different manifolds. Agents on the continuous torus demonstrate no preference for a specific direction across consecutive training runs. Conversely, agents trained on other manifolds exhibit a tendency to avoid the direction that leads to an angle flip

when crossing the corresponding boundary. Especially for the projective plane topology, the agents tend to aggregate more while aligning, even without adding another reward for aggregation.

For Aggregation in Figure 4.21, we also find successful aggregation of agents in the middle. In practice, one may define any objective of interest. For example, we can achieve misalignment in Figure 4.21, resulting in polar order parameters on the order of magnitude of 10^{-2}, and showing the generality of the framework.

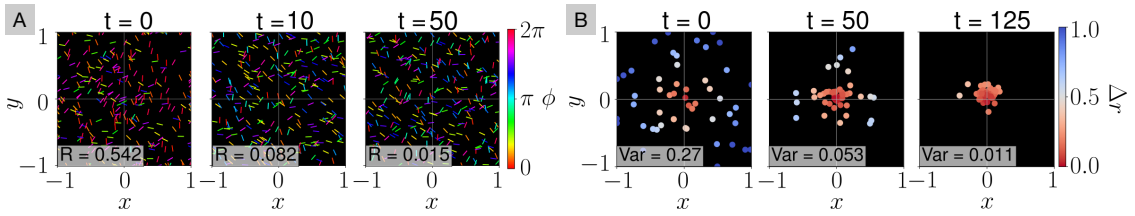


FIGURE 4.21: Qualitative behavior of Dec-POMFC in the Vicsek and Aggregation problem. A: Qualitative behavior for misalignment of agents in the Vicsek (torus) problem. B: The two-dimensional Aggregation problem, with agent distances to mean as colors.

ADDITIONAL EXPERIMENTS Some other experiments are discussed in Appendix F.17, including the generalization of our learned policies to different starting conditions, a comparison of the Vicsek model trained or transferred to different numbers of agents, additional interpretative visualizations, similar success for the Kuramoto model, and a favorable comparison between RBF and histograms for higher dimensions, showing the generality of the framework and supporting our claims.

4.3.4 *Summary*

Our framework provides a novel methodology for engineering artificial collective behavior in a rigorous and tractable manner, whereas existing scalable learning frameworks often focus on competitive or fully observable models [132, 267]. We hope our work opens up new applications of partially-observable swarm systems. Our method could be of interest due to (i) its theoretical optimality guarantees while covering a large class of problems, and (ii) its surprising simplicity in rigorously reducing complex Dec-POMDPs to MDPs, with same complexity as MDPs from fully observable MFC, thus allowing analysis of Dec-POMDPs via a tractable MDP.

The current theory remains limited to non-stochastic MFs, which in the future could be analyzed for stochastic MFs via common noise [1, 127, 268]. Further, sample efficiency could be analyzed [269], and parametrizations for history-dependent policies using more general NNs could be considered, e.g. via hypernetworks [255, 270]. Lastly, extending the framework to consider additional practical constraints and sparser interactions, such as physical collisions or via graphical decompositions, may be fruitful.

4.4 CONCLUSION OF CHAPTER 4

In this chapter, we began our exposition of MFC by considering a simplified, quasi-static MFC problem with external dynamic environment states. There, we have shown that MFC solutions with identical policies suffice for approximate optimality in MF systems, and we have performed the reduction to the MFC MDP together with RL for MFC-based MARL, together with a basic DPP. The primary motivating example was a power-of-d load balancing system, which we have also studied in external collaborations not presented in this thesis [15, 21].

We then moved on to extend MFC towards more general settings. First, the weak interaction of agents was ameliorated by addressing strong interaction in the presence of many minor agents and a major player, through the framework of M3FC. There, we also analyzed RL for MFC-based MARL on the finite MARL system, showing the near-exactness of PGs in large systems with many agents under Lipschitz conditions on the policy. Propagation of chaos are also shown to hold under Lipschitz conditions on the model. Overall, the algorithms were verified and compared against the common PG MARL methods of IPPO and MAPPO, and showed that MFC-based MARL can outperform standard MARL techniques.

Finally, we have extended the applicability of MFC by considering partial observability and decentralization through the Dec-POMFC framework, which is particularly important in real large-scale systems, where centralization cannot typically be assumed under the presence of many agents. In the case of limiting deterministic MF, the otherwise difficult partially-observable MFC is reduced to a standard MFC during training using the CTDE paradigm. During execution, given an initial MF, the policies can then be executed decentrally. Theoretical results were extended by giving a novel DPP under restriction to equi-Lipschitz policies. The framework was compared against IPPO and MAPPO and achieved comparable results. Quantitative support was shown for theoretical results. In Kuramoto and Vicsek models, we also showed the generalization to unknown initial states and model variations.

In a nutshell, we have addressed our RQs I and II of learning and model generality, in order to synthesize scalable MARL algorithms through MFC in the cooperative case. While the subject of MFC in discrete time for MARL remains an active area of research, we hope that our contributions have shown how to obtain scalable MARL algorithms under the usage of MF approximations in more realistic, non-standard MFC scenarios. In the next and penultimate chapter, we will discuss applications of MFGs and MFC or large-scale MARL in general, including both a list of potential applications and particular ones in more detail. Answering RQ III, we hope to affirm the usefulness of our developed frameworks in applications beyond the ones already considered in preceding experiments.

5

PRACTICAL APPLICATIONS IN MANY-AGENT SYSTEMS

In this chapter, we study the applicability of MFGs and MFC in practical large-scale controlled systems. We begin by listing a number of potential applications for large-population, many-agent MARL beyond the ones already presented in preceding chapters. In particular, we list some possible applications in distributed computing, cyber-physical systems, autonomous mobility and natural or social sciences. We also refer the reader to further existing surveys on applications of MFGs and MFC with less focus on the learning aspect. We then move on to consider two exemplary applications of embodied drone swarms, where collisions between physical agents must be avoided, as well as edge computing scenarios for the balancing of computational offloading resources. We have also studied load balancing and network applications in external collaborations [10, 15, 21], which are however not presented in this thesis.

5.1 POTENTIAL APPLICATIONS OF LARGE-POPULATION MARL

Today, RL already finds application in various application areas such as LLMs [39], robotics [27], autonomous driving [28] or navigation of stratospheric balloons [29] as a method to realize effective sequential decision-making in complex problems. Similarly for MARL as a generalization of RL, potential applications for MARL are manifold and include e.g. teams of unmanned aerial vehicles [30, 271] or video games [35, 75]. While the domain of MARL is somewhat empirically successful, large-scale MARL still remains subject of active research. Here, the classes of competitive MFGs and cooperative MFC problems naturally contain a large number of real world scenarios with many agents, and can find varied applications, e.g., in analysis of power network resilience [272], smart heating [53], edge computing [273] or flocking [129].

Many more applications still remain to be considered. For example, intelligent and self-organizing robotic swarms provide engineering applications such as Internet of Things or precision farming, for which a general design framework remains elusive [232, 233] and requires scalable handling of partial information. Other domains include group decision-making and opinion dynamics [234], biomolecular self-assembly [235] and active matter [236, 237] such as self-propelled nano-particles [238], microswimmers [239], etc. [231]. Furthermore, in the recent years, there has been a surge of interest in large-scale multi-agent dynamical systems on higher-order networks due to their great generality and practical importance e.g. in epidemiology [274], opinion dynamics [275, 276], network synchronization [277, 278], neuroscience [279], and more. We refer the interested readers to the excellent review articles [280–282]. For all of the above, the application of MFG and MFC frameworks developed in this thesis may be fruitful.

In this section, we give an overview over selected areas of applications that could highly profit from further research in scalable MARL methods – and vice versa as inspirations for future algorithms. Systems with large numbers of agents are somewhat ubiquitous, and in the following we will give some general areas of application for which this is true. We note that applications are not only the motivation of developing general, scalable algorithms. Instead, in addition to motivating algorithms by applications, one can also find inspiration from specialized approaches developed for specific applications, in existing areas of research. Therefore, applications of MARL may also allow for further insight into how to develop general MARL algorithms and MFGs or MFC models. Apart from the applications listed here, there are many more applications of large-scale MARL also in other topics such as economics and finance, etc., for which we also refer to a variety of surveys [37, 38, 64, 283]. In particular, we point out the surveys on applications of MFGs and MFC specifically, in engineering [65] and finance [64]. The material presented in this section is based on [20].

5.1.1 *Distributed Computing*

An important area of application with high accessibility in terms of simulated training data, is given by networked computers and computing applications, including for example also video games, where the advantages of MARL have been prominently and repeatedly demonstrated in scenarios with up to, e.g., 10 agents [35, 36, 75, 284]. On the other hand, MARL in scenarios with significantly larger population sizes such as in Neural MMO [285] has not yet seen similar levels of success, and the above benchmark was only recently proposed. Here, future work towards a better understanding and successful design of large-scale multi-agent interaction is ongoing and could find application in making real games more interactive. Similarly, one could consider also other computing applications such as peer-to-peer systems [286] and decentralized finance [287], where automated game-theoretic analysis of user behavior by MFGs could help, e.g., in system design.

Apart from very specific computing applications, another important use case of large-scale MARL may be the optimization of distributed computing itself. For example, RL has long been used to find adaptive load balancing algorithms [288]. Even to this day, the study of load balancing remains an open problem, for example in the presence of partial observability and delayed information [289], where studies on systems with large populations of servers and clients remain of importance and – in consideration of today's increasingly large-scale computing infrastructure – continue to be investigated using e.g. MF analysis [230] and learning MFC, see also our external collaborations not discussed in this thesis [10, 15, 21]. Some related areas of research include throughput optimization [56], cloud resource sharing [109, 290] and edge computing [273] in systems with many devices, where MF approximations are often already used [57, 109]. However, such formulations are typically used to analytically derive results, which has the disadvantages of requiring extensive manual efforts and considering only special cases. Here, scalable and automated MARL algorithms could enable solutions for more complex or real systems.

5.1.2 *Cyber-Physical Systems*

Cyber-physical systems constitute another highly important and emerging subject area. Apart from the many applications of single-agent learning in robotics [27, 291], more scalable MARL methods could find further applications for swarms of embodied agents. The market for UAVs is developing rapidly, and swarms of drones could reach large-scale deployment in the near future [292], owing to their great number of potential applications. In general, swarms of terrestrial, marine and especially aerial drones could thus be a key technology for tasks such as establishing communication networks for disaster management [293], performing efficient search-and-rescue missions [294] or delivering packages [295]. However, deploying drone swarms in the real world is associated with a variety of coordination challenges [296], not seldom stemming from the complexity of real-world environments. Here, automated and effective decision-making for large swarms of drones remains yet an active area of research with few general design methods [233].

Similar applications are not restricted to embodied, interactive agents and can long be found in other sectors of industry such as energy [297], heating [298], and water distribution [299]. These wide-scale critical infrastructure sectors may similarly profit from developments in large-scale system control via deep RL, see e.g. recent works on power networks [272, 300, 301] or smart heating [53], many of which have also found solutions in formulations based on MF limits [65]. In the future, critical infrastructure could profit also from achieving the effective design of more decentralized control solutions also in terms of resilience, since secure, reliable electric power and water supply remain of paramount importance to society. Here, scalable learning methods on networks could enable key technologies such as smart grids, which may well be key to preparing and hardening the power grid against future natural and man-made disasters [302].

5.1.3 *Autonomous Mobility*

Somewhat related to cyber-physical systems, autonomous mobility and traffic control is among the most challenging application areas of MARL and has received well-deserved attention from both academia and the industry in the last decades. In both of these applications, large numbers of agents participate in real-life scenarios in which the coordination is both highly safety-critical and usually non-cooperative. The algorithms for such applications are required to deal with many challenges MARL problems pose simultaneously, in addition to fulfilling safety and other regulatory

concerns. Different heterogeneous vehicles of different capabilities, as well as other entities of traffic, such as traffic lights and vehicles, must be considered as part of the population. Since none of the agents would have a global view, the algorithms should not only tackle safety, but also partial-observability.

Many real-world scenarios associated with autonomous mobility involve a large-population of agents and therefore require scalable learning methods. Especially following the increasing connectivity of the vehicles and considering the highly dynamic and unpredictable nature of mobile systems. Important challenges of the area includes safety constraints, standardized or compatible algorithms for different vehicles and integration with existing systems, e.g. manually controlled vehicles and human interaction. Here, recent work for the lower-level control vehicles of employs e.g. graph-based models [303, 304] and distributed RL-based methods [305] for scalability. For a more detailed view of the latest works and open challenges in autonomous mobility, we refer the reader to the recent surveys [28, 306, 307].

On the higher level, traffic control with congestion as well as route planning problems such as vehicle routing problems [308] are further related application area that attract attention, as traffic congestion becomes more and more problematic with the increasing population and proliferation of private cars. In particular, this higher level comes with the requirement of dealing a large number of entities such as traffic lights, vehicles and pedestrians, and therefore is an excellent potential application for MF models. One way to benefit from MARL in traffic control is adaptive traffic signal control, which have been considered via our developed major-minor framework [61]. We also refer the reader to further recent works that propose MF-based approaches to this end [54, 55, 108].

5.1.4 *Natural and Social Sciences*

Lastly, foregoing control for a moment, the study of behavior of large-scale dynamical systems is quite classical: The MF theory originates in statistical physics for the description of magnetic materials [47], which has since also been used as a benchmark in large-population MARL [120]. Analogous approaches are also often found in social sciences through opinion dynamics on networks of people [51, 309], or in particular through analyzing general interacting particle systems on complex networks [310]. Oftentimes, each agent in such models can be endowed with decision-making capabilities, leading for example to applications such as the analysis of crowd dynamics in the case of building evacuations [60, 110, 111]. As a result, a potential natural application of MFG and MFC-based MARL can be found in natural and social sciences.

One example of particular relevance is the study of spread and control of epidemics [49], which is not restricted to biological epidemics but includes also e.g. malware spread on computer networks [50, 311, 312], and is of recent interest due to the COVID-19 pandemic. Such systems can be seen as multi-agent systems connected via complex and adaptive networks, see for example [313–315] for work in this direction. Recently, many works using RL for finding optimal decisions in epidemic situations have emerged. The works in this direction include but are not limited to [316–319]. Here, we and other existing works have also used graph-based approaches to represent heterogeneous interactions between players together with learning [7, 153].

5.2 COLLISION-FREE MEAN FIELD CONTROL FOR EMBODIED DRONE SWARMS

MARL remains challenging both in its theoretical analysis and empirical design of algorithms, especially for large swarms of embodied robotic agents where a definitive toolchain remains part of active research. We use emerging state-of-the-art MFC techniques in order to convert many-agent swarm control into more classical single-agent control of distributions. This allows profiting from advances in single-agent RL at the cost of assuming weak interaction between agents. However, the MF model is violated by the nature of real systems with embodied, physically colliding agents. Thus, we combine collision avoidance and learning of MFC into a unified framework for tractably designing intelligent robotic swarm behavior. On the theoretical side, we provide novel approximation guarantees for general MFC both in continuous spaces and with collision avoidance. On the practical side, we show that our approach outperforms MARL and allows for decentralized open-loop application while avoiding collisions, both in simulation and real UAV swarms. Overall, we propose a framework for the design of swarm behavior that is both mathematically well-founded and practically useful, enabling the solution of otherwise intractable swarm problems. The material presented in this section is based upon our work [5].

Over the past decades, the field of swarm robotics [227, 320, 321] has received considerable attention [322]. Various areas of potential applications include for example industrial inspection tasks [323], such as for turbines, cooperative object transport [324–326], agriculture [327], aerial combat [328], and cooperative search [329]. A recent promising approach for engineering many-agent systems such as intelligent robot swarms is MARL [40], which has found success in diverse complex problems such as strategic video games [75], communication networks [330] or traffic control [331]. However, MARL algorithms suffer from issues such as credit assignment, non-stationarity and scalability to many agents [40]. Meanwhile, robotic swarms such as fleets of UAVs usually consist of many interacting UAVs and remain of considerable interest due to their inherent robustness, scalability to large-scale deployment and decentralization, which can be considered the ultimate goal of the study of swarm intelligence and robotics [227, 251]. Here, scalable control approaches and highly general toolchains for swarm robotics remain to be established [320].

A classical approach to formulate systems with large numbers of agents with low complexity is via MF models, describing swarms of drones by their distribution, see also [332] and [124] for reviews on MF swarm robotics and MFC. However, most prior literature is based on analytic derivations and continuous-time models, which are less conducive to advances in RL. For example, stabilizing control of swarms to distributions are designed in [333–335]. Other works such as [336, 337] consider population density estimates via collisions for task allocation problems, while [338] study robots for stick-pulling. Lastly, a variety of approaches use PDE-based formulations, e.g. [339, 340] for density control, or [341, 342] for general analytic frameworks, though they are significantly more difficult to treat both rigorously and from a learning perspective. Especially MF-based learning algorithms often remain restricted to competitive settings such as MFGs [22, 23] by learning e.g. Nash [95, 115, 117, 120, 131], regularized [9, 142] or correlated equilibria [343, 344]. For instance, works such as [345] or [208] investigate trajectory control of selfish UAV agents, while [346] considers formation flight in dense environments. Although selfish control problems are interesting for many applications, aligning selfish or local cost functions with a certain cooperative, global behavior can be difficult [347]. Solutions for cooperative joint objectives without necessity of manual cost function tuning are therefore of practical interest for artificially engineering swarm behaviors.

In this work, we propose a discrete-time MFC-based swarm robotics framework that is conducive to powerful deep RL techniques. Only very recently were MFC [104, 105, 107] and related histogram

observations for MARL [348] proposed as a potential solution to cooperative scalable MARL, which could enable both the solution of otherwise intractable tasks as well as model-free application to swarms, adapting to environments and tasks. However, an eminent issue of MFC for robotic systems is violation of the MFC model due to physical collisions between robots. To solve this issue, we combine MFC with deep RL and collision avoidance algorithms. Here, collision avoidance algorithms could range from classical rule-based [349] over planning-based [350] to learning-based approaches [16, 32], and similarly for RL, see e.g. [351]. Importantly, our approach (i) is able to utilize advances in RL, circumventing MARL and solving otherwise difficult swarm problems without extensive manual and analytical design of algorithms, and (ii) closes the gap between MF models and reality, as collisions between agents violate the weak interaction principle of MF models and are usually to be avoided, e.g. in UAVs. As a result, our approach is highly practical, with the advantage of automatic design of swarm algorithms for swarm problems.

Our contribution can be summarized as follows: (i) We combine RL with MFC and collision avoidance algorithms for general task-driven control of robotic swarms; (ii) We give novel theoretical approximation guarantees of MFC in finite swarms as well as in the presence of additional collision avoidance maneuvers; (iii) We demonstrate in a variety of tasks that MFC outperforms state-of-the-art MARL, can be applied in a decentralized open-loop manner and avoids collisions, both in simulation and real UAV swarms. Overall, we provide a general framework for tractable swarm control that could be applied directly to swarms of UAVs.

5.2.1 *A Model of Embodied Swarms*

In order to tractably describe a plethora of swarm tasks, we formulate a MF model where all agents are anonymous and it is sufficient to consider their distribution.

5.2.1.1 *Finite Swarm Model*

Formally, we consider compact state and action spaces $\mathcal{X}, \mathcal{U} \subseteq \mathbb{R}^2$ (though our results are easily extended to $\mathbb{R}^3$) representing possible locations and movement choices of an agent. For any $N \in \mathbb{N}$, at each time $t = 0, 1, \ldots$, the states and actions of agent $i = 1, \ldots, N$ are denoted by $x_t^{i,N}$ and $u_t^{i,N}$. We denote by $\mathcal{P}(\mathcal{X})$ the space of probability measures on $\mathcal{X}$, equipped with the topology of weak convergence. Define the empirical state distribution $\mu_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{x_t^{i,N}} \in \mathcal{P}(\mathcal{X})$, which represents all agents anonymously by their states. We consider policies $\pi = \{\pi_t\}_{t \geq 0} \in \Pi$ from a space of policies Π with shared Lipschitz constant, such that agents act on their location and the distribution of all agents, $\pi_t \colon \mathcal{X} \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{U})$. The assumption of Lipschitz continuity is standard in the literature, includes e.g. neural networks [105, 116, 352], and may allow approximation of less regular policies.

Under a policy $\pi \in \Pi$, the finite swarm system shall evolve by sampling an initial state $x_0^{i,N} \sim \mu_0$ from an initial distribution μ_0 of agents, and subsequently taking movement actions $u_t^{i,N} \sim \pi_t(x_t^{i,N}, \mu_t^N)$, resulting in new states $x_{t+1}^{i,N} = x_t^{i,N} + u_t^{i,N} + \epsilon_t^i$ for all agents i with optional i.i.d. Gaussian noise $\epsilon_t^i \sim \mathcal{N}(0, \Sigma)$ and diagonal covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$. In other words, each drone can move a distance limited to $\mathcal{U}$, up to some smoothing or inaccuracy ϵ_t^i. In simulation, we further clip agent positions to stay inside $\mathcal{X}$. The objective is then given by an arbitrary function

$r\colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ of the spatial distribution of agents, giving rise to the infinite-horizon discounted objective

$$J^N(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(\mu_t^N)\right]. \tag{5.2.1}$$

Since MARL can be difficult in the presence of many agents (see e.g. combinatorial nature in [40]), we will formulate and verify the limiting infinite-agent system.

5.2.1.2 *Mean Field Swarm Model*

In the limit as $N \to \infty$, single agents become indiscernible and we need only model their distribution (MF) $\mu_t \in \mathcal{P}(\mathcal{X})$. Starting at μ_0, under policy $\pi \in \Pi$, deterministically

$$\mu_{t+1} = T^{\pi_t(\mu_t)}(\mu_t) := \iint \mathcal{N}(x+u, \sigma^2)\, \pi_t(\mathrm{d}u \mid x, \mu_t)\, \mu_t(\mathrm{d}x) \tag{5.2.2}$$

with the deterministic MF transition operator $T^{\pi_t(\mu_t)}(\mu_t)$ as a function of $\pi_t(\mu_t) \in \mathcal{P}(\mathcal{U})^{\mathcal{X}}$ and the current MF μ_t, giving way to the MFC problem with objective function

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(\mu_t)\right]. \tag{5.2.3}$$

Remark 5.2.1. *A dependence of r on joint state-action distributions in $\mathcal{P}(\mathcal{X} \times \mathcal{U})$ can be modelled by splitting time steps into two and using the new state space $\mathcal{X} \cup (\mathcal{X} \times \mathcal{U})$.*

For simplicity of analysis, we assume absence of common noise, leading to a deterministic MF limit, though in our experiments we also allow reactions to a random external environment. Under a mild continuity assumption, weaker than the common Lipschitz assumption in existing literature [105, 352], we obtain rigorous approximation guarantees.

Assumption 5.2.1. *The reward function r is continuous.*

By compactness of $\mathcal{P}(\mathcal{X})$, r is bounded. As long as r is continuous, i.e. small changes in the agent distribution lead to small changes in reward, the MFC model is a good approximation for large swarms and its solution solves the finite agent system approximately optimally. As existing approximation properties still remain limited to finite $\mathcal{X}, \mathcal{U}$ [104, 105], we give a brief, novel proof for compact spaces.

Theorem 5.2.1. *Under Assumption 5.2.1, at all times $t \in \mathcal{T}$, the empirical reward $r(\mu_t^N)$ converges weakly and uniformly to the limiting reward $r(\mu_t)$ as $N \to \infty$, i.e.*

$$\sup_{\pi \in \Pi} \mathbb{E}\left[\left|r(\mu_t^N) - r(\mu_t)\right|\right] \to 0. \tag{5.2.4}$$

Proof. We can metrize $\mathcal{P}(\mathcal{X})$ via the metric $d(\mu,\nu) \coloneqq \sum_{m=1}^{\infty} 2^{-m}|\mu(f_m) - \nu(f_m)|$ for a sequence of continuous and bounded $f_m \colon \mathcal{X} \to \mathbb{R}$, $|f_m| \leq 1$ (cf. [216, Theorem 6.6]).

Consider any (uniformly) equicontinuous set $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{P}(\mathcal{X})}$ of functions, i.e. there exists an increasing (concave, cf. [353, p. 41]) $\omega_{\mathcal{F}} \colon [0,\infty) \to [0,\infty)$ (modulus of continuity) such that $\omega_{\mathcal{F}}(x) \to 0$ when $x \to 0$ and $|f(\mu) - f(\nu)| \leq \omega_{\mathcal{F}}(d(\mu,\nu))$ for all $f \in \mathcal{F}$. We show inductively for $t \geq 0$ that

$$\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_t^N) - f(\mu_t)\right|\right] \to 0, \tag{5.2.5}$$

which implies the desired property, since r is uniformly continuous by compactness of $\mathcal{P}(\mathcal{X})$ and Assumption 5.2.1.

At time $t = 0$, the proof follows from the weak LLN argument (see Eq. (5.2.7) and below). For the induction step,

$$\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_{t+1}^N) - f(\mu_{t+1})\right|\right] && (5.2.6)\\
&\quad\leq \sup_{\pi\in\Pi} \mathbb{E}\left[\omega_{\mathcal{F}}(d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N)))\right] && (5.2.7)\\
&\qquad + \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(T^{\pi_t}(\mu_t^N)) - f(\mu_{t+1})\right|\right] && (5.2.8)
\end{aligned}$$

where for the first term Eq. (5.2.7), by Jensen's inequality we obtain

$$\mathbb{E}\left[\omega_{\mathcal{F}}(d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N)))\right] \leq \omega_{\mathcal{F}}\left(\mathbb{E}\left[d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N)\right]\right)$$

for concave $\omega_{\mathcal{F}}$. Abbreviating $x_t^N \equiv \{x_t^{i,N}\}_{i\in[N]}$, we have

$$\begin{aligned}
&\mathbb{E}\left[d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N)\right]\\
&\quad= \sum_{m=1}^{\infty} 2^{-m}\mathbb{E}\left[\left|\mu_{t+1}^N(f_m) - T^{\pi_t}(\mu_t^N)(f_m)\right|\right]\\
&\quad\leq \sup_{m\geq 1} \mathbb{E}\left[\mathbb{E}\left[\left|\mu_{t+1}^N(f_m) - T^{\pi_t}(\mu_t^N)(f_m)\right| \;\middle|\; x_t^N\right]\right],
\end{aligned}$$

where by the weak LLN argument, the squared term

$$\begin{aligned}
&\mathbb{E}\left[\left|\mu_{t+1}^N(f_m) - T^{\pi_t}(\mu_t^N)(f_m)\right| \;\middle|\; x_t^N\right]^2\\
&\quad\leq \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(f_m(x_{t+1}^{i,N}) - \mathbb{E}\left[f_m(x_{t+1}^{i,N}) \;\middle|\; x_t^N\right]\right)\right|^2 \;\middle|\; x_t^N\right]\\
&\quad= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left(f_m(x_{t+1}^{i,N}) - \mathbb{E}\left[f_m(x_{t+1}^{i,N}) \;\middle|\; x_t^N\right]\right)^2 \;\middle|\; x_t^N\right]\\
&\quad\leq \frac{4}{N} \to 0
\end{aligned}$$

since for any f_m, the cross-terms are zero and $|f_m| \leq 1$.

For the second term Eq. (5.2.8), by induction assumption we have

$$\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(T^{\pi_t}(\mu_t^N)) - f(\mu_{t+1})\right|\right]\\
&\quad\leq \sup_{\pi\in\Pi}\sup_{g\in\mathcal{G}} \mathbb{E}\left[\left|g(\mu_t^N) - g(\mu_t)\right|\right] \to 0
\end{aligned}$$

using $g = f \circ T^{\pi_t}$ and the corresponding class $\mathcal{G}$ of functions with modulus of continuity $\omega_{\mathcal{G}} \coloneqq \omega_{\mathcal{F}} \circ \omega_T$, where ω_T denotes the uniform modulus of continuity of T^{π_t} by uniform Lipschitz continuity of $\pi \in \Pi$. □

As a result, the MFC approach is a theoretically rigorous approach to approximately optimally solving large-scale swarm problems with complexity independent of N.

Corollary 5.2.1. *Under Assumption 5.2.1, an optimal solution $\pi^* \in \Pi$ to the MFC problem constitutes an ε-optimal solution to the finite swarm problem, where $\varepsilon \to 0$ as $N \to \infty$.*

Proof. For any $\pi \in \Pi$ and $\varepsilon > 0$, we can choose T such that $\sum_{t=T+1}^{\infty} \gamma^t \mathbb{E}\left[\left|r(\mu_t^N) - r(\mu_t)\right|\right] \leq 2^{-T} \max_\mu 2|r(\mu)| < \frac{\varepsilon}{4}$, and for sufficiently large N $\sum_{t=0}^{T} \gamma^t \mathbb{E}\left[\left|r(\mu_t^N) - r(\mu_t)\right|\right] < \frac{\varepsilon}{4}$ by Theorem 5.2.1. Therefore, we have $J^N(\pi^*) - \max_{\pi\in\Pi} J^N(\pi) = \min_{\pi\in\Pi}(J^N(\pi^*) - J^N(\pi)) \geq \min_{\pi\in\Pi}(J^N(\pi^*) - J(\pi^*)) + \min_{\pi\in\Pi}(J(\pi^*) - J(\pi)) + \min_{\pi\in\Pi}(J(\pi) - J^N(\pi)) \geq -\frac{\varepsilon}{2} + 0 - \frac{\varepsilon}{2} = -\varepsilon$ by the prequel and optimality of π^* in the MFC problem. □

5.2.2 *MFC with Collision Avoidance*

In order to remove the two remaining obstacles of (i) solving the MFC problem, and (ii) resolving the real-world gap of MFC for embodied agents, we combine MFC with arbitrary powerful RL and collision avoidance techniques. The overall hierarchical structure is found in Figure 5.1. The MFC solution is learned via RL and gives high-level directions, which are realized by each agent while avoiding collisions.

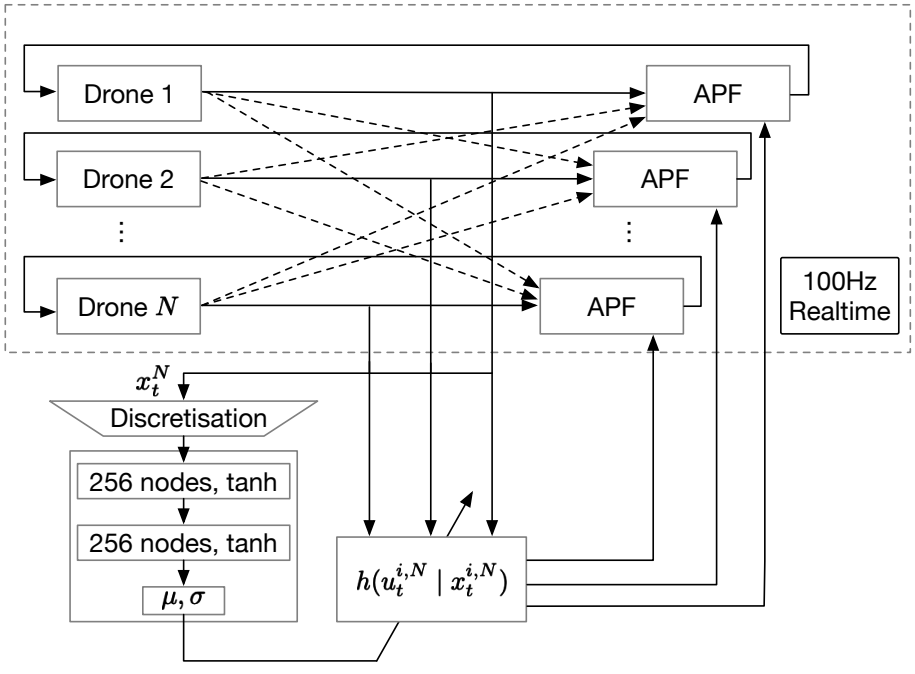


FIGURE 5.1: A hierarchical overview of our approach. The learned high-level MFC policy sends movement instructions to the UAV swarm, while each agent uses a real-time collision avoidance algorithm – here Artificial Potential Field (APF) – to avoid collisions with others.

5.2.2.1 *Reinforcement Learning*

For the MFC problem, it is known that there exists an optimal stationary solution [107, Theorem 19], which may be found by solving the MFC MDP, a single-agent but infinite-dimensional RL problem with $\mathcal{P}(\mathcal{X})$-valued states μ_t and $\mathcal{P}(\mathcal{U})^{\mathcal{X}}$-valued actions h_t evolving according to $\mu_{t+1} = T^h(\mu_t)$. To deal with the infinite dimensionality of $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{U})^{\mathcal{X}}$, we discretize $\mathcal{X}$ and use a binned histogram of $\mathcal{P}(\mathcal{X})$ as in [107] with $M = 6^2 = 36$ bins by trading off between tractability (good training, low M) and performance (high M), while $\mathcal{P}(\mathcal{U})$ is parametrized by Gaussians with means $\theta \in \mathcal{U}$ and diagonal covariances $\sigma_1, \sigma_2 \in (0, 0.25]$, of which the samples $u_t^i \sim h(u_t^i \mid x_t^i) = \mathcal{N}(\theta, \mathrm{diag}(\sigma_1, \sigma_2))$ are clipped to $\mathcal{U}$. As exact computation of μ_t is difficult, we use the finite system with $N = 300$ agents (though less works fine) and their empirical distribution analogous to particle filtering, which can be understood as directly learning on a large finite swarm.

We use the RLlib 1.13.0 implementation [76] of PPO RL [73] and a diagonal Gaussian neural network policy with two hidden $\tanh$-layers of 256 nodes, sampling clipped values in $[-1, 1]$ affinely transformed to $\theta, \sigma_1, \sigma_2$. Hyperparameters are printed in Table 5.1, of which sufficiently high minibatch sizes appeared most important.

TABLE 5.1: Hyperparameter configurations for PPO.

Symbol	Name	Value
γ	Discount factor	0.99
λ	GAE lambda	1
β	KL coefficient	0.03
ϵ	Clip parameter	0.2
l_r	Learning rate	0.00005
B_b	Training batch size	4000
B_m	Minibatch size	1000
T_b	Updates per training batch	5

5.2.2.2 *Collision Avoidance Subroutine*

A solution of the MF system does not directly translate into applicable real-world behavior, since the MF solution ignores physical constraints. While e.g. UAVs could fly at different heights, a general swarm algorithm should explicitly avoid collisions in order to guarantee suitability of the weakly-interacting MFC model. This is done by separating concerns, decomposing the issue into MFC plus sequences of collision-avoiding navigation subproblems between decision epochs. For example, we could choose $\mathcal{U}$ slightly smaller than the maximum speed range to allow for additional avoidance maneuvers. Then, assuming the time Δt between two MFC decisions t and $t + 1$ is sufficiently long, and that agents have finer, direct control over their positions, a collision-avoiding navigation subroutine could approximately achieve the desired positions up to an error that becomes arbitrarily small with agent radius r.

For N drones and agent radius r we hence assume existence of such a subroutine F which mildly perturbs all positions and their distribution μ_t^N at each time step and thereby achieves a collision-free MF, which we write as $F(\mu_t^N)$, such that $\|x_t^i - x_t^j\|_2 > 2r$ for all i, j. We further assume that F is near-optimal, i.e. each drone's position is perturbed at most by a distance of $4Nr$. Indeed, this is possible for sufficiently small r, e.g. if $\mathcal{X} = [-m, m]^2$ for $m > 0$: At any $x \in \mathcal{X}$, on an arbitrary line of length greater $2m$ passing through x, we can always choose a position that is at most $4Nr$

away from x, as in the worst case all other $N-1$ drones are located on the line along which F moves the drone and have a distance of slightly less than $4r$ between each other. Under F, we can show that a collision-avoiding finite swarm of sufficiently many small agents is solved well by our approach.

Theorem 5.2.2. *Let $\pi \in \Pi$ be an optimal solution to the MFC problem, and let F be the near-optimal collision avoidance subroutine as defined above. Then for each $\varepsilon > 0$ there exists an N' such that for all $N \geq N'$ and agent radii $r_{N,\varepsilon}$, the solution π gives an ε-optimal solution to the finite swarm problem with collision avoidance.*

Proof. The definition of F allows us to define new model dynamics with new random MF variables denoted by $\mu_t'^N$, where we leave out the definition of each agent variable for brevity. For the new dynamics, at each time step t we apply function F to the current MF $\mu_t'^N$ and underlying positions. Subsequently, the MF $\mu_{t+1}'^N$ is obtained by applying the usual transition dynamics.

Now, we show via induction over t that for all t,

$$\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_t^N) - f(F(\mu_t'^N))\right|\right] \to 0. \tag{5.2.9}$$

Analogous to the proof of Theorem 5.2.1, the induction start follows from a weak LLN argument. For the induction step,

$$\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_{t+1}^N) - f(F(\mu_{t+1}'^N))\right|\right] \\
&\quad\leq \sup_{\pi\in\Pi} \mathbb{E}\left[\omega_{\mathcal{F}}(d(\mu_{t+1}^N, T^{\pi_t}(\mu_t^N))\right] \\
&\quad+ \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(T^{\pi_t}(\mu_t^N)) - f(T^{\pi_t}(F(\mu_t'^N)))\right|\right] \\
&\quad+ \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(T^{\pi_t}(F(\mu_t'^N))) - f(\mu_{t+1}'^N)\right|\right] \\
&\quad+ \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_{t+1}'^N) - f(F(\mu_{t+1}'^N))\right|\right]
\end{aligned} \tag{5.2.10}$$

where the first two summands converge to zero by arguments as in the proof of Theorem 5.2.1. The third term converges to zero by a weak LLN argument while the forth summand is bounded by $\omega_{\mathcal{F}}(4Nr_{N,\varepsilon})$, see the explanation above. By choosing $r_{N,\varepsilon} = o(1/N)$, the last summand in Eq. (5.2.10) converges to 0. This concludes the induction.

For ε-optimality, we proceed as in Corollary 5.2.1 and obtain

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t \left|r(\mu_t) - r(F(\mu_t'^N))\right|\right] < \frac{\varepsilon}{2}$$

for N large enough by applying statement Eq. (5.2.9). The terms beyond $T-1$ can be bounded by $\varepsilon/2$ as in Corollary 5.2.1. □

Hence, for a given allowed sub-optimality specification ε, we can find a number N and size r of drones such that solving the MFC problem is ε-optimal in the finite swarm system. In practice, this means that if we can use sufficiently many sufficiently small drones, MFC provides good solutions.

In this work, for simplicity we use APF as in [354] with attractive velocity $F_d = 1.5(\hat{x}_t^i - x_t^i)$ in simulation, where $\hat{x}_t^i$ denotes the MFC-based target position, and similarly repulsive velocity from

agent j on agent i, $F_{ji} = 1.5c_{\text{rep}} \cdot (\frac{1}{\|x_t^i - x_t^j\|_2} - 1) \cdot \frac{x_t^i - x_t^j}{\|x_t^i - x_t^j\|_2^3}$ whenever $\|x_t^i - x_t^j\|_2 \leq 1$ and zero otherwise, where $c_{\text{rep}} > 0$ is a variable repulsion coefficient. However, we stress that other more advanced collision avoidance algorithms could be used.

5.2.3 *Experiments*

In this section, we verify the usefulness of MFC-based robotic swarm control experimentally.

5.2.3.1 *Problems*

We consider three problems of increasing complexity to demonstrate our approach. In the following, we consider uniform initial state distributions $\mu_0 = \text{Unif}(\mathcal{X})$ and let $\mathcal{X} = [-2, 2]^2$, allowing circular-constrained, noise-free movement, i.e. circular $\mathcal{U}$ such that $\|u_t^i\|_2 \leq 0.2$ with $\epsilon_t^i \equiv 0$.

AGGREGATION In the simple Aggregation or Rendezvous [348] problem, the goal of agents is to aggregate into a point while minimizing movement. Hence, we choose rewards $r(\nu_t) = \iint -\|x - \int x\, \nu_t(\mathrm{d}x, \mathrm{d}u)\|_2 - 0.3\|u\|_2\, \nu_t(\mathrm{d}x, \mathrm{d}u)$ for joint state-actions $\nu_t = \mu_t \otimes h_t \in \mathcal{P}(\mathcal{X} \times \mathcal{U})$ (see Remark 5.2.1).

FORMATION In the Formation problem, the goal is to achieve an anonymous formation flight of large swarms, i.e. matching the distribution of agent positions with a given distribution – e.g. for providing coverage for surveillance or communication. The rewards are given by the Wasserstein distance $r(\mu_t) = \inf_{X,Y:\, \mathcal{L}(X)=\mu_t, \mathcal{L}(Y)=\mu^*} \mathbb{E}\left[\|X - Y\|_2\right]$ [214] between agent distribution μ_t and e.g. a Gaussian mixture $\mu^* = \frac{1}{2}\mathcal{N}(e_1, \text{diag}(0.05, 0.05)) + \frac{1}{2}\mathcal{N}(-e_1, \text{diag}(0.05, 0.05))$ with unit vector e_1, computed via the empirical Wasserstein distance between agents and 300 samples of μ^*. In principle, it is also possible to add movement costs as in Aggregation.

TASK ALLOCATION Lastly, we formulate a problem with stochasticity even in the limit. Consider randomly generated, spatially localized tasks such as providing a UAV-based communication uplink, or emergency operations for clearing rubble and firefighting. We add spatially localized tasks to the model which are observed via an additional histogram of task locations. Here, in each time step, $N_t = \text{Pois}(0.4)$ tasks l arrive at uniformly random points $x^l \in \mathcal{X}$, up to a maximum of 5 total tasks. Each task l begins with length $L_t = 10$ and at each time step is processed abstractly by proximity of nearby agents according to $L_{t+1}^l = L_t^l - \Delta L^l(\mu_t)$, $\Delta L^l(\mu_t) := \min(1, \int (1 - 2\|x - x^l\|_2) \mathbf{1}_{\|x - x^l\|_2 \leq 0.5}\, \mu_t(\mathrm{d}x)$, until it is fully processed and disappears. The reward is defined by the processed task lengths $r(\mu_t) = \sum_l \Delta L^l(\mu_t)$.

5.2.3.2 *Experimental Results*

In the following, we show results demonstrating the power of our MFC framework for task-driven swarm control, namely their theoretical and numerical advantage over standard MARL, the potential for decentralized open-loop control, and the influence of collision avoidance on optimality, both in simulation and in the real world.

TRAINING RESULTS In our implementation, each training episode consists of 50, 100 and 200 time steps for the Aggregation, Formation and Task Allocation problems respectively, of which the average sum of rewards will constitute the return values shown in the following figures. As can be seen in Figure 5.2, the learning curve of PPO in the MFC problem is smoothly increasing as expected, since the MFC MDP leads to a single-agent problem solvable via standard RL with better understood theory than MARL, e.g. [74].

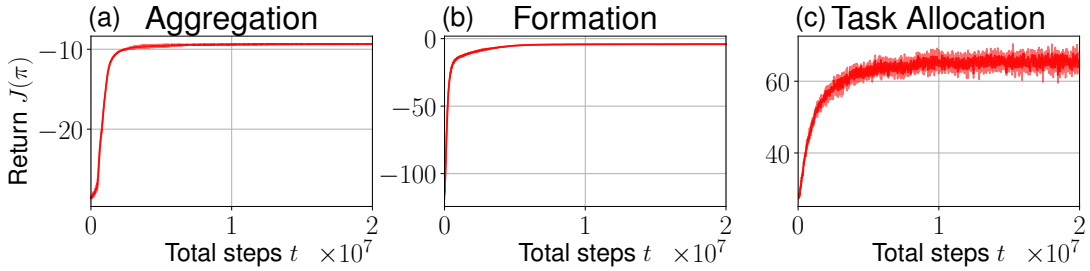


FIGURE 5.2: Training curves of the MFC algorithm trained on $N = 300$, plotting the average achieved objective over time steps taken, with its standard deviation over 3 seeds. The MFC approach leads to stable learning results for all of our considered problems. (a): Aggregation; (b): Formation; (c): Task Allocation.

In contrast, state-of-the-art MARL techniques miss theoretical guarantees. We compare to PPO with parameter-sharing [90] and independent learning [86], which has repeatedly achieved state-of-the-art performance in benchmarks [42, 87–89] and remains applicable to arbitrary numbers of homogeneous agents. For comparability, we use the same architecture and implementation as in our MFC experiments, outputting parameters of a Gaussian over actions. Each agent simply observes the same information plus the agent's own position.

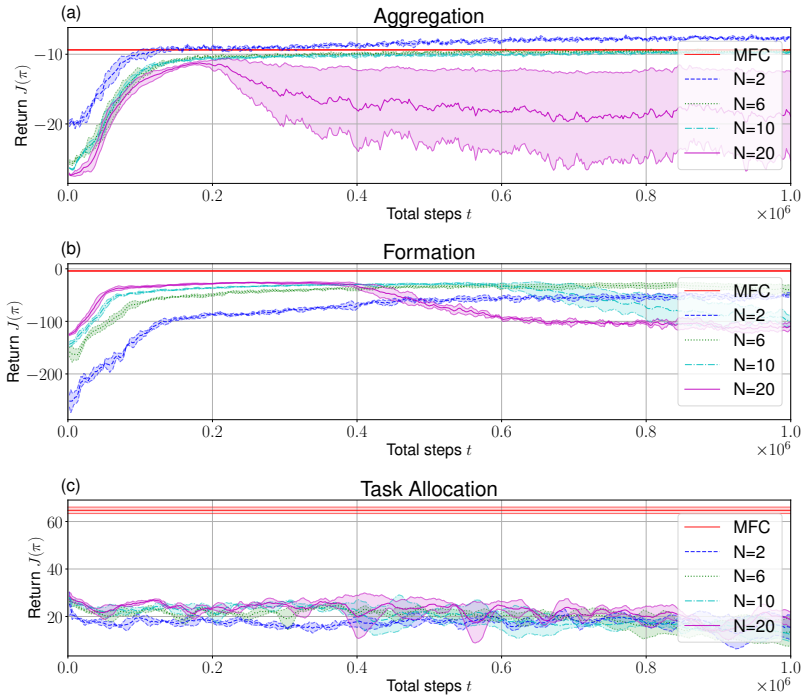


FIGURE 5.3: Training curves of N-agent IPPO, plotting the average achieved objective over time steps taken, together with its standard deviation over 3 seeds and compared to final MFC performance (red, $N = 300$). (a): In the simple Aggregation task, MARL and MFC are comparable for few agents, but MARL fails for many agents. (b-c): In more complex scenarios, MFC converges to a better solution than common MARL.

As seen in Figure 5.3, MARL works well in the very simple Aggregation task, but becomes increasingly unstable for many agents, especially in the more complex Formation scenario, finally failing entirely in Task Allocation due to non-stationarity of learning [40]. Although MARL could work for other hyperparameter configurations, it shows that standard MARL can suffer from worse stability than single-agent RL in even high-dimensional single-agent MFC MDPs, congruent with the outstanding issue of theoretical MARL convergence guarantees [40].

As seen exemplarily for the Formation problem in Figure 5.4 and a variation of the problem with real drones (later) in Figure 5.8, the MFC solution successfully achieves the desired mixture of Gaussian formation of agents.

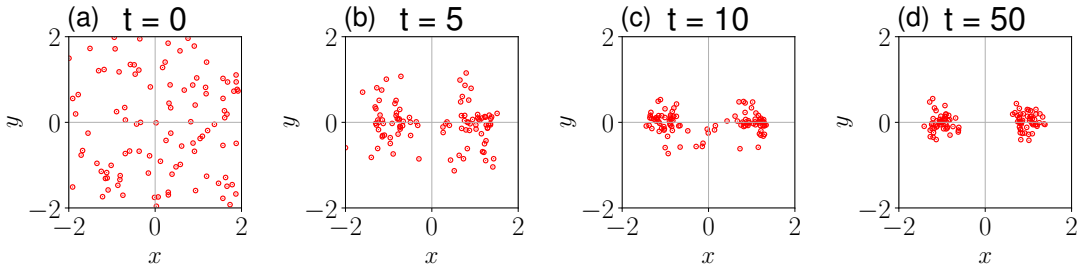


FIGURE 5.4: One sample run of the MFC solution to the Formation problem, applied to a system with $N = 100$ agents and plotted at times $t \in \{0, 5, 10, 50\}$. Agents successfully form a mixture of two Gaussians.

In Figure 5.5, it can be seen that (i) the MFC solution outperforms MARL, and (ii) the MFC solution quickly converges to the limiting deterministic objective in Figure 5.2 as N grows large, verifying the MFC approximation properties in Theorem 5.2.1.

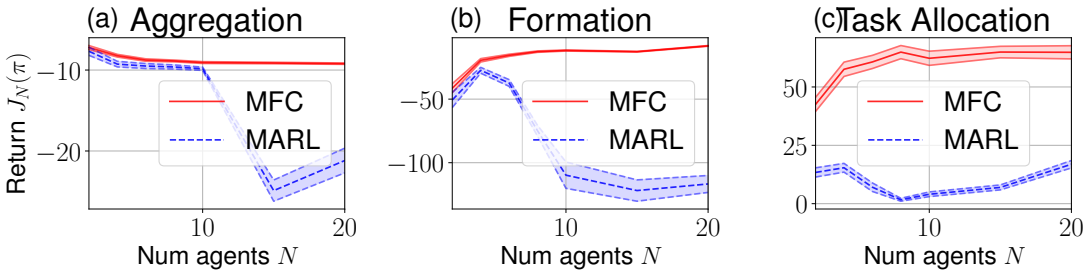


FIGURE 5.5: Comparison of achieved objectives in the finite swarm of MFC and MARL solutions over 100 sample episodes, with 95% confidence interval (shaded). The MFC algorithm quickly converges to the deterministic, limiting MF objective as N becomes large. In simple scenarios such as Aggregation (a), MARL outperforms MFC in the finite system, while in more complex scenarios (b-c), MFC outperforms MARL (at end of training).

DECENTRALIZED OPEN-LOOP CONTROL In the absence of global information, it makes sense for large swarms to let agents act stochastically and independently, especially since agents are interchangeable and anonymous. For this purpose, as long as the limiting MFC is deterministic (e.g. Aggregation and Formation), we can compute an optimal open-loop control sequence $h_0, h_1, \dots$ of MFC actions $h_t \in \mathcal{P}(\mathcal{U})^{\mathcal{X}}$ for a given starting μ_0, and apply h_t to each agent. This results in both open-loop and decentralized control, as each agent moves depending on its own local position only. As expected by determinism of MFC, in Figure 5.6 we observe that the open-loop performance becomes practically indistinguishable from the closed-loop performance in Aggregation, as well as approaches it in Formation for sufficiently large swarms. We note that at least for finite spaces, very recently a similar decentralization result was also rigorously shown [352].

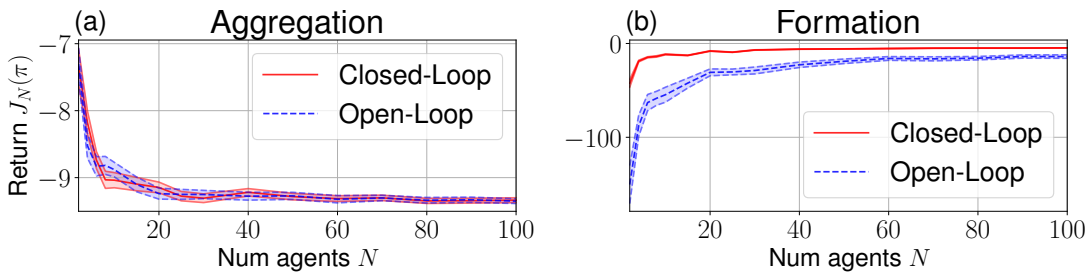


FIGURE 5.6: Comparison of mean objectives in the finite swarm system of closed-loop and open-loop MFC over 100 sample episodes, with 95% confidence interval (shaded). In Aggregation (a), little difference can be seen between the closed-loop and open-loop performance. In Formation (b), the open-loop policy is unable to react to stochastic initialization effects of finite swarm size, only approaching optimality in the large swarm limit.

INFLUENCE OF COLLISION AVOIDANCE For MFC with collision avoidance, we simulate $\Delta t = 2$ and 100 explicit Euler steps of length 0.02 between each decision epoch t. Furthermore, to avoid bad initializations, we resample initial states until the minimal inter-agent distance is above 0.1. As seen in Figure 5.7, the minimal inter-agent distance is easily tuned via c_{rep}, rising up to the initialization distance 0.1. We find that for strong collision avoidance, the performance deteriorates in the presence of many agents, whereas for smaller collision avoidance coefficients the performance approaches the MF limit, verifying Theorem 5.2.2.

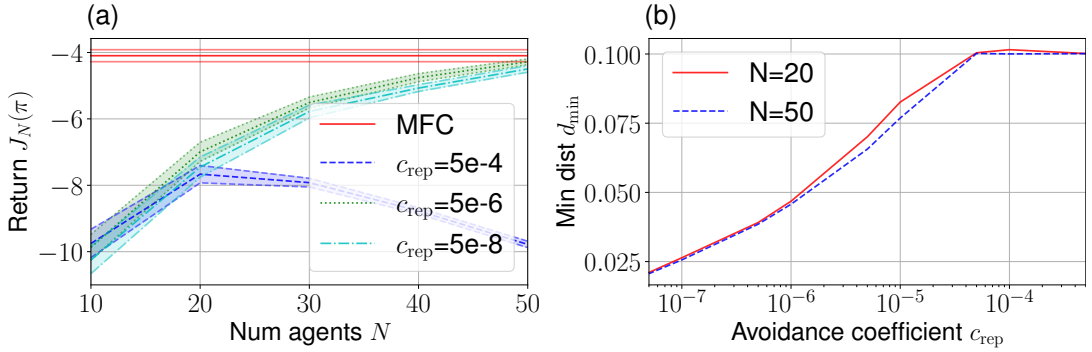


FIGURE 5.7: Comparison of results in finite swarms of MFC solution with collision avoidance for various collision avoidance coefficients c_{rep} in the Formation problem. (a): Mean objectives averaged over 100 sample episodes, with 95% confidence interval (shaded). (b): Minimum occurring inter-agent distance over 100 sample episodes.

ILLUSTRATIVE REAL-LIFE EXPERIMENT Lastly, we show the results of applying a variant of the Formation task – tracking a single time-variant Gaussian moving on a circle – to a real fleet of Crazyflie quadcopters [355], each peer-to-peer-broadcasting only their local Lighthouse-based state estimates [356]. Here, we use the aforementioned decentralized open-loop control and APF-based collision avoidance. Although our experiments remain small-scale due to space constraints and downwash effects, we nonetheless show that our approach works in practice and can be applied to even small swarm sizes. In the future, we imagine similar approaches to be scaled up to larger fleets. As can be seen in Figure 5.8, the agents successfully track the formation without colliding.

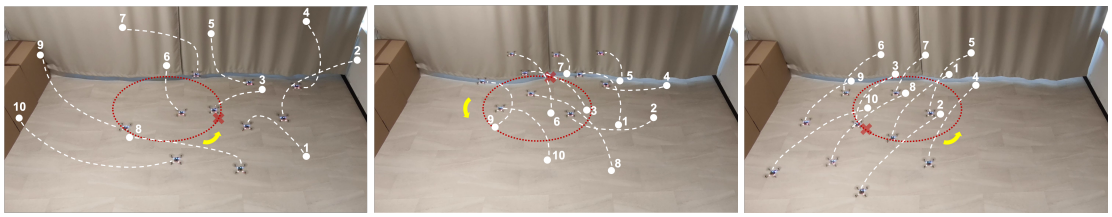


FIGURE 5.8: Real world coverage experiment with a swarm of 10 Crazyflie nano-quadcopters and a variant of the Formation problem where agents track a single time-varying Gaussian distribution (current center of Gaussian shown as red cross), moving counter-clockwise on a circle (red dotted line). The drones successfully track the time-varying Gaussian distribution using the open-loop control policy without collision. Time progresses from left to right.

5.2.4 *Summary*

In this work, we have proposed a scalable task-driven approach to robotic swarm control that allows for model-free solution of swarm tasks while remaining applicable in practice by using deep RL, MFC and collision avoidance. Our approach is hierarchical, in principle allowing to profit from any state-of-the-art RL and collision avoidance techniques. Our work is a step towards general toolchains for robotic swarm control, which yet remain part of active research [320]. We have solved part of the limitations of MF theory for embodied agents by integrating collision avoidance into the toolchain, but more work on more sparsely interacting MF models may be necessary, e.g. for UAV-based communication with strongly neighbor-dependent interaction, by incorporating graph structure [7, 139, 357]. Extensions to non-linear dynamics and dynamical constraints may be fruitful. Lastly, while our Gaussian parametrization of $\mathcal{P}(\mathcal{U})$ is efficient, the state discretization still suffers from a curse of dimensionality, as the number of bins rises quickly with fineness of discretization, which was state of the art [104, 107] and could be supplemented e.g. by convolutional techniques [118] or kernel methods as discussed in Section 4.3.

5.3 EDGE COMPUTING AND SERVER LOAD BALANCING

The optimal offloading of tasks in heterogeneous edge-computing scenarios is of great practical interest, both in the selfish and fully cooperative setting. In practice, such systems are typically very large, rendering exact solutions in terms of cooperative optima or Nash equilibria intractable. For this purpose, we adopt a general MF formulation in order to solve the competitive and cooperative offloading problems in the limit of infinitely large systems. We give theoretical guarantees for the approximation properties of the limiting solution and solve the resulting MF problems numerically. Furthermore, we verify our solutions numerically and find that our approximations are accurate for systems with dozens of edge devices. As a result, we obtain a tractable approach to the design of offloading strategies in large edge-computing scenarios with many users. The material presented in this section is based upon our work [6].

In recent years, a rapid growth of data generated from the network edge is witnessed, especially, the Cisco Annual Internet Report 2020 forecasts a rapid deployment of billions of Machine-To-Machine (M2M) devices until 2023 [358]. Multi-access Edge Computing (MEC) is a key technology to compensate strictly limited M2M devices in their processing by enabling computation offloading to cloudlet servers with computation resources in their vicinity. Additionally, the number of User Edge devices (UEs) like smartphones, tablets and laptops also have increased tremendously due to the ease of their availability and low costs. These devices can also gain from offloading their tasks that demand intensive computations and low latencies, e.g., virtual reality, real-time face recognition, LLM. A MEC system is a multi-agent system where each UE is an agent who decides, for each arriving task, whether to offload it to the MEC server or not. There has been great interest in finding the optimal policy for these UEs, to either offload or process locally, depending on factors such as their own available resources, network conditions and offloading computation costs. Even though several computation offloading strategies between UEs and MEC servers have been proposed in the literature, finding scalable solutions for MEC multiagent systems remains an important problem considering the continuously increasing number of agents.

There are two main categories in which agents can work in a multi-agent system, either cooperatively to maximize a global goal or competitively to maximize their own reward, or a combination of both. MFGs provide a way to analyze and solve large-scale competitive problems in a tractable manner. On the other hand, MFC may be used to tractably model cooperative settings in many-agent systems. We will formulate offloading in edge-computation as both a one-shot problem with theoretical guarantees and alternatively a time-stationary problem, allowing for competitive and cooperative solutions in a unified, tractable manner.

Various prior works have used MF approximations for similar offloading problems. In [290], the authors model a shared MEC competitive offloading problem as an MFG in continuous time and solve the resulting partial differential equations using FPI. However, their model considers only the non-cooperative case and results in a continuous time model, while in our work we also consider a cooperative setting and obtain a model for the time-stationary case. Similarly, in [273, 359], the authors consider both cooperative and non-cooperative computational offloading problems, though through the special case of a linear-quadratic model, while we solve a non-linear problem. MF approximations have also been used to model large-scale MEC systems with D2D collaborations [360] particularly on graphs as a deterministic ODE system, though the model does not consider entirely selfish nodes. Finally, authors in [109] model and solve a large-scale resource-sharing problem using MF theory and both cooperative and non-cooperative strategies, i.e. a more centralized setting without local computational capabilities. In contrast to our work, their models focus on graph-based job forwarding and continuous-action resource-sharing problems, whereas our model

focuses on offloading decisions. For a variety of other works on applications of the MF approach in communication systems, see also e.g. [361–364]. Apart from the discussed differences, all of the models in prior works and our work are diverse and apply to a variety of differing scenarios. Here, in contrast to previous work, our model will consider cooperative and competitive optimization of binary offloading decisions in edge-computing, where users may choose to offload or compute locally. In this work, we will formulate a unified MF framework for both the competitive and cooperative setting of offloading decisions in edge-computing. In particular, our tractable solution considers both a one-shot and time-stationary scenario that approximates the finite user system.

We begin our analysis with a computationally expensive-to-solve one-shot game, a knapsack problem, where many edge devices must independently decide whether to offload or not for a given distribution of task configurations. Our contribution can be summarized as follows: (i) We pass to the infinite-user limit in order to obtain a tractable problem with a complexity independent of the number of users; (ii) The model is theoretically motivated by showing novel existence and approximate optimality properties of solutions in large finite systems, both in terms of Nash equilibria and Pareto optima; (iii) Since, in practice, a one-shot game may not be sufficiently realistic, we extend our model to a new time-stationary model with a Poisson task arrival process and find analogous competitive and cooperative solutions in the limit of large systems; and finally (iv), the proposed models are verified in simulation and solved or learned with complexity independent of the number of users, while concurrently giving a good solution to large finite systems. As a result, we tractably solve an otherwise intractable many-agent offloading problem.

5.3.1 *A MFG and MFC Model*

In the considered scenario, a multi-cell ultra-dense network includes M MEC servers and N UEs associated with MEC servers. It is assumed that M MEC servers are connected through a fiber loop to share their computational resources, pooled into a single centralized but distributed MEC pool, where we also assume the bandwidth across the MEC servers is large enough for connecting all UEs and the delay in resource sharing through fiber loop is neglected, similar to the framework presented in [290]. The scenario is depicted in Figure 5.9 with M MEC servers and N UEs.

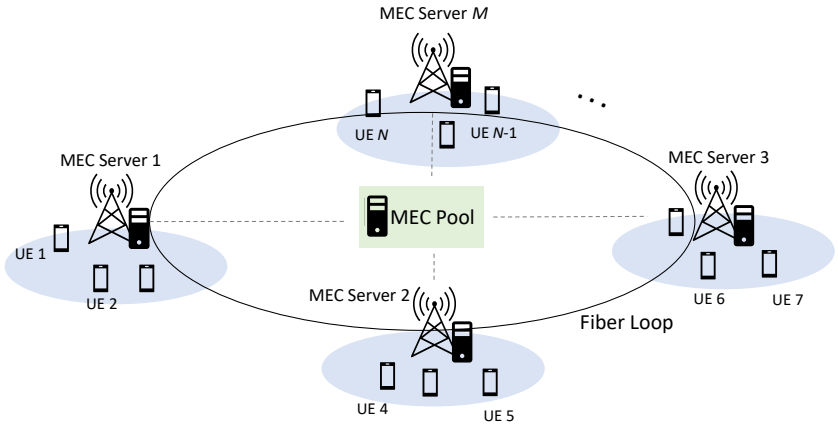


FIGURE 5.9: MEC scenario with N UEs offloading their tasks to M MEC servers, where computation resources of MEC servers are shared through a fiber loop connecting them, acting as a single processing pool.

Each UE, modeled by $i = 1, \dots, N$, is given a task configuration

$$C^i \coloneqq (W^i, L^i, f^i, R^i) \in \mathbb{K} \subseteq \mathbb{R}^4_{\geq 0} \tag{5.3.11}$$

with transmission length W^i in bits, processing complexity L^i in CPU cycles to be computed (either by offloading to a MEC server or by computing locally), transmission rate R^i between user and MEC server measured in bits per second, and a local processing rate f^i in CPU cycles per second. Each user i may make a decision $u^i \in \{0, 1\}$ of whether to offload their task.

For a given overall MEC server processing rate $f_{\text{pool}} \in \mathbb{R}$, we assume each offloading UE is allocated a proportional amount of processing power from the MEC processing pool to their offloaded task's complexity. The time T_i^{tx} to transmit the i-th users' task to a MEC server and the time T_i^{off} to compute at the MEC server are given by

$$T_i^{\text{tx}} = \frac{W^i}{R^i}, \quad T_i^{\text{off}} = \frac{L^i}{f_{\text{pool}} \cdot \frac{1}{\sum_j u^j}} \tag{5.3.12}$$

assuming that each offloading task is assigned equal processing power. Alternatively, one could also easily consider completion of all offloaded tasks at once, i.e. $T_i^{\text{off}} = \frac{u^j \sum_j L^j}{f_{\text{pool}}}$. The time to compute a task locally is given by

$$T_i^{\text{loc}} = \frac{L^i}{f^i}. \tag{5.3.13}$$

The result of the computed tasks are assumed to be negligible in size compared to original task size W^i, therefore, the time needed for the reception of results is not considered.

5.3.1.1 *Competitive Game Setting*

In the competitive, selfish setting, each user i independently decides on whether to offload or not via the random variable $u^i \in \{0, 1\}$ so as to minimize only their own expected computation time, i.e. either time to compute locally or offload

$$\mathbb{P}(u^i = 1) \cdot (T_i^{\text{tx}} + T_i^{\text{off}}) + \mathbb{P}(u^i = 0) \cdot T_i^{\text{loc}} = \mathbb{E}\left[u^i(T_i^{\text{tx}} + T_i^{\text{off}}) + (1 - u^i)T_i^{\text{loc}}\right] \tag{5.3.14}$$

Under full information, we obtain a standard static game with the classical solution concept of mixed Nash equilibria: Each of the users chooses whether to offload according to a policy π_i which gives the conditional probability of offloading

$$\mathbb{P}(u^i = 1 \mid C^1, \dots, C^N) \equiv \pi_i(C^1, \dots, C^N) \tag{5.3.15}$$

which results in the minimization objective of each user i,

$$J_i^N(\pi_1, \dots, \pi_N) = \mathbb{E}\left[u^i(T_i^{\text{tx}} + T_i^{\text{off}}) + (1 - u^i)T_i^{\text{loc}}\right]. \tag{5.3.16}$$

An approximate ε-Nash equilibrium is now defined as a tuple of policies $(\pi_1, \dots, \pi_N)$ such that no user can gain by unilaterally changing their policy, i.e. for any $i = 1, \dots, N$,

$$J_i^N(\pi_i, \pi_{-i}) \leq \max_{\pi \in \Pi} J_i^N(\pi, \pi_{-i}) + \varepsilon \tag{5.3.17}$$

where π_{-i} denotes all policies other than the i-th policy. Here, the minimal such ε is also referred to as the exploitability of policies. An exact Nash equilibrium for $\varepsilon = 0$ is indeed guaranteed to exist as long as $\mathbb{K}$ is compact (e.g. [77]).

Unfortunately, it is known that the computation of Nash equilibria is hard, see [45]. Instead, we shall consider the many-agent case through MF analysis to obtain a tractable solution for large systems. At the same time, the solution will consist of decentralized policies. To this end, we shall now assume that there exists an underlying distribution $\mu_0 \in \mathcal{P}(\mathbb{K})$ of user specifications, i.e. for $i = 1, \dots, N$ we have random variables $C^i = (W^i, L^i, f^i, R^i) \sim \mu_0$. To obtain a reasonable solution, we must also assume that the MEC pool computing power scales suitably with the number of users, i.e. $f_{\text{pool}} = N \cdot f_{\text{per}}$ for some $f_{\text{per}} \in \mathbb{R}$, since otherwise in the limit of many agents, offloading will become pointless. In practice, for fixed f_{pool} and N in given finite N-agent systems, this may be realized by defining $f_{\text{per}} := \frac{f_{\text{pool}}}{N}$.

We now consider a decentralized control setting by allowing each agent to decide whether to offload depending only on their own configuration C_i. For motivation, note that since all other agents are exchangeable from the perspective of a single agent, only the own state and overall distribution of behaviors of other agents matters. Furthermore, in the limit of $N \to \infty$, the other users' distribution is uninformative, since under a common offloading strategy, the distribution converges to some fixed MF by the LLN. Additionally, decentralized control policies may be motivated in practice by limited agent information. Since the computation of Nash equilibria in this setting nonetheless remains hard, this motivates the MF formulation.

For tractability, we formulate a MFG as $N \to \infty$, as popularized by [23] and [22] for stochastic differential games. Here, we propose a MF model with near-Nash properties as N grows large, as we will also verify theoretically. Consider a policy π shared by all users. The policy induces a joint distribution $\mu = \mu_0 \otimes \pi$ over user states and offloading decisions. Under this fixed distribution μ, the objective of a single, representative user becomes

$$J^\mu(\pi) = \mathbb{E}\left[u(\tilde{T}^{\text{tx}} + \tilde{T}^{\text{off}}) + (1-u)\tilde{T}^{\text{loc}}\right] \tag{5.3.18}$$

where we have expectations of random variables of the representative agent $(W, L, f, R, u) \sim \mu \otimes \pi$, and random transfer or processing times of the MF system W, L, f, R

$$\tilde{T}^{\text{tx}} = \frac{W}{R}, \quad \tilde{T}^{\text{off}} = \frac{L \int u \, \mathrm{d}\mu}{f_{\text{per}}}, \quad \tilde{T}^{\text{loc}} = \frac{L}{f}. \tag{5.3.19}$$

The $N \to \infty$ analogue of Nash equilibria is the MF equilibrium, defined as a tuple (π^*, μ^*) of policy and MF, such that the policy is optimal under the MF generated by itself, i.e. defined through the fixed point equation

$$\pi^* = \arg\min_\pi J^{\mu^*}(\pi), \tag{5.3.20a}$$

$$\mu^* = \mu_0 \otimes \pi^*. \tag{5.3.20b}$$

Analytically, for any fixed MF μ, we could find such a best response policy $\text{BR}(\mu)$ by defining

$$\pi^*(W, L, f, R) = \mathbf{1}_{\tilde{T}^{\text{tx}}(W,L,f,R) + \tilde{T}^{\text{off}}(W,L,f,R) < \tilde{T}^{\text{loc}}(W,L,f,R)}. \tag{5.3.21}$$

However, simply iterating the two fixed point equations is generally not guaranteed to converge to an equilibrium. Thus, we will learn equilibria through FP [96].

5.3.1.2 *Cooperative Control Setting*

In contrast to the selfish, competitive setting, in a cooperative setting it may be of interest to minimize the average processing time of all users. One may formulate a centralized optimization problem as

$$\min_{u^1,\dots,u^N} \quad \frac{1}{N}\sum_{i=1}^{N} u^i(T_i^{\text{tx}} + T_i^{\text{off}}) + (1-u^i)T_i^{\text{loc}} \tag{5.3.22a}$$

$$\text{s.t.} \quad u^i \in \{0,1\} \qquad \forall i \in \{1,\dots,N\} \tag{5.3.22b}$$

under full information. However, again this problem is known to be difficult to solve exactly for large N, as it is a knapsack problem [365]. Furthermore, we may again be interested in a decentralized solution, where each agent uses an independent policy, eliminating the need for centralized knowledge and only requiring knowledge of the local configuration C^i. Reformulating as optimization over decentralized policies π_i and optimizing over the expected cost, we have

$$\min_{\pi^1,\dots,\pi^N} \quad \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} u^i(T_i^{\text{tx}} + T_i^{\text{off}}) + (1-u^i)T_i^{\text{loc}}\right] \tag{5.3.23a}$$

$$\text{s.t.} \quad \pi_i \colon \mathbb{K} \to [0,1] \qquad \forall i \in \{1,\dots,N\} \tag{5.3.23b}$$

where each offloading decision $u^i \sim \text{Bernoulli}(\pi_i(C_i))$ follows from the policy π_i.

As $N \to \infty$, under the policy π for all agents, we can obtain the corresponding MFC problem, which is more tractable than directly solving the N-user system, given as

$$\min_{\pi} \quad \int u(\tilde{T}^{\text{tx}} + \tilde{T}^{\text{off}}) + (1-u)\tilde{T}^{\text{loc}} \,\mathrm{d}(\mu_0 \otimes \pi) \tag{5.3.24}$$

again with the previous definitions. Note that although we impose a shared, common policy π, sharing a policy across all agents will indeed be sufficient for optimality [8].

Since the problem is now reduced to the choice of $\pi \colon \mathbb{K} \to [0,1]$, the combinatorial optimization problem has been reduced to optimization over a bounded function π with complexity independent of N. If we further assume that μ_0 has finite support, i.e. $K := |\mathbb{K}| < \infty$ and

$$\mu_0 = \sum_{j=1}^{K} p_j \delta_{(W_j,L_j,f_j,R_j)}, \quad \sum_{j=1}^{K} p_j = 1, \tag{5.3.25}$$

for some $p_j \geq 0$, $(W_j, L_j, f_j, R_j) \in \mathbb{K}$, then we obtain

$$\begin{aligned}
&\int u(\tilde{T}^{\text{tx}} + \tilde{T}^{\text{off}}) + (1-u)\tilde{T}^{\text{loc}} \,\mathrm{d}(\mu_0 \otimes \pi) \\
&= \sum_{j=1}^{K} p_j \pi_j \left(\frac{W_j}{R_j} + \frac{L_j \sum_{k=1}^{K} p_k \pi_k}{f_{\text{per}}}\right) + p_j(1-\pi_j)\frac{L_j}{f_j} \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K} \frac{p_j p_k L_j}{f_{\text{per}}} \pi_j \pi_k + \sum_{j=1}^{K}\left(\frac{p_j W_j}{R_j} - \frac{p_j L_j}{f_j}\right)\pi_j + \sum_{j=1}^{K}\frac{p_j L_j}{f_j} \\
&= \boldsymbol{\pi}^T \boldsymbol{Q} \boldsymbol{\pi} + \boldsymbol{c}^T \boldsymbol{\pi} + \text{const.}
\end{aligned}$$

for $\boldsymbol{\pi} \equiv (\pi_1, \dots, \pi_K)^T$, $\pi_j = \pi(W_j, L_j, f_j, R_j)$ and appropriate $\boldsymbol{Q}$, $\boldsymbol{c}$. Therefore, we obtain a non-convex quadratic program

$$\min_{\pi_1,\dots,\pi_K} \quad \boldsymbol{\pi}^T \boldsymbol{Q} \boldsymbol{\pi} + \boldsymbol{c}^T \boldsymbol{\pi} \tag{5.3.26a}$$

$$\text{s.t.} \quad \pi_j \in [0,1] \qquad \forall j \in \{1, \dots, K\} \tag{5.3.26b}$$

with box constraints, which though NP-hard [366] in the cardinality of the support of μ_0, can be solved numerically. Most importantly, the complexity remains independent of N, giving us a tractable solution for sufficiently small K. To handle more general densities μ_0 with non-finite but compact support $\mathbb{K}$, we may discretize distributions and solve the resulting finite-support problem. As a result, we have obtained a tractable solution to the otherwise intractable offloading problem for many devices, as we will verify in the sequel.

5.3.2 *Time-Stationary Equilibrium Behavior*

While the previous model assumes an instantaneous problem where we let all users play a one-shot game, another important and interesting setting is to assume a continuous flow of tasks arriving over time. While a theoretically rigorous analysis of this setting is beyond the scope of our work, we nonetheless consider this setting at its time-stationary equilibrium and solve it numerically.

At all times, let the arrival process of tasks be given by a Poisson process with constant rate λN, which is equivalent to Poisson arrival rates λ for each of N users. At equilibrium, in the limit there must be a time-stationary bandwidth per user f_{alloc} allocated to a user choosing to offload their task. This bandwidth is given by dividing the total processing power $f_{\text{pool}} = N f_{\text{per}}$ by the number of jobs in the system. Since the processing time for any offloaded job arriving at equilibrium is given by $T^{\text{tx}} = \frac{W}{R}$ and $T^{\text{off}} = \frac{L}{f_{\text{alloc}}}$, the expected number of jobs in the system as $N \to \infty$ is given by

$$\mathbb{E}\left[N_{\text{tot}}\right] = \lambda N \mathbb{E}\left[u(T^{\text{tx}} + T^{\text{off}})\right] \tag{5.3.27}$$

and is given by a sum of N Poisson variables N_{tot}^i, the numbers of jobs in the system from each user i. Therefore, by the central limit theorem, the fluctuations of N_{tot} are on the order of $O(\sqrt{N})$, resulting in the allocated processing rate per user

$$f_{\text{alloc}} = \frac{f_{\text{pool}}}{\mathbb{E}\left[N_{\text{tot}}\right] + O(\sqrt{N})} = \frac{f_{\text{per}}}{\lambda \mathbb{E}\left[u(T^{\text{tx}} + T^{\text{off}})\right] + O(\frac{1}{\sqrt{N}})} \to \frac{f_{\text{per}}}{\lambda \mathbb{E}\left[\frac{uW}{R}\right] + \frac{\lambda \mathbb{E}[uL]}{f_{\text{alloc}}}} \tag{5.3.28}$$

as $N \to \infty$, which for $f_{\text{alloc}} \neq 0$ gives

$$f_{\text{alloc}} = \frac{f_{\text{per}} - \lambda \mathbb{E}\left[uL\right]}{\lambda \mathbb{E}\left[\frac{uW}{R}\right]} \tag{5.3.29}$$

and the natural constraint

$$f_{\text{per}} - \lambda \int uL \, \mathrm{d}(\mu_0 \otimes \pi) > 0. \tag{5.3.30}$$

Intuitively, this constraint formalizes the notion of sufficient MEC resources, i.e. the rate of assigned jobs times their complexity must not exceed the possible compute assigned per node, as otherwise

the MEC servers will be unable to catch up with assigned tasks, resulting in no time-stationary solution. Note that this constraint is trivially fulfilled if

$$f_{\text{per}} > \lambda \mathbb{E}\left[L\right]. \tag{5.3.31}$$

Optimizing the average waiting times of all agents in the cooperative case gives the MFC problem

$$\min_{\pi \in [0,1]^{\mathbb{K}}} \mathbb{E}\left[u(T^{\text{tx}} + T^{\text{off}}) + (1-u)\tilde{T}^{\text{loc}}\right] \tag{5.3.32}$$

where for finite $\mathbb{K}$ we have

$$\begin{aligned}
&\mathbb{E}\left[u(T^{\text{tx}} + T^{\text{off}}) + (1-u)\tilde{T}^{\text{loc}}\right] \\
&\quad = \mathbb{E}\left[\frac{uW}{R} + \frac{uL\lambda\mathbb{E}\left[\frac{uW}{R}\right]}{f_{\text{per}} - \lambda\mathbb{E}\left[uL\right]} + (1-u)\frac{L}{f}\right] \\
&\quad = \sum_{j=1}^{K} p_j \pi_j \left(\frac{W_j}{R_j} + \frac{L_j \lambda \sum_{k=1}^{K} \frac{p_k \pi_k W_k}{R_k}}{f_{\text{per}} - \lambda \sum_{k=1}^{K} p_k \pi_k L_k}\right) + p_j(1-\pi_j)\frac{L_j}{f_j}.
\end{aligned}$$

For the competitive MFG, we can analogously define an equilibrium as any fixed point policy π^* such that

$$\pi^* \in \arg\max_{\pi} \mathbb{E}_{\pi}\left[\frac{uW}{R} + \frac{uL\lambda\mathbb{E}_{\pi^*}\left[\frac{uW}{R}\right]}{f_{\text{per}} - \lambda\mathbb{E}_{\pi^*}\left[uL\right]} + (1-u)\frac{L}{f}\right]. \tag{5.3.33}$$

5.3.3 *Theoretical Guarantees*

In this section, we state a number of theoretical guarantees for the one-shot MF problems. Extensions to the time-stationary case are deferred to future work. The results follow from formulating the problem as certain standard MFG and MFC problems and applying existing results. In particular, note that our systems can be reformulated as standard MFGs with action space $\{0,1\}$ and state space $\mathbb{K} \cup (\mathbb{K} \times \mathbb{U})$, see also [8, 24]. For the competitive setting, as $N \to \infty$, the MFG equilibrium exists and will constitute an approximate Nash equilibrium.

Theorem 5.3.1. *A solution (π^*, μ^*) of the MFG problem Eq. (5.3.20) exists, and π^* constitutes an ϵ_N-Nash equilibrium of the finite N-user system with $\epsilon_N \to 0$ as $N \to \infty$.*

Proof. See [24, Theorem 4.1]. □

Furthermore, it is known that the FP algorithm will converge in terms of exploitability, giving us the desired approximate Nash equilibrium.

Theorem 5.3.2. *The exploitability of the solution of the FP algorithm converges to zero.*

Proof. The system fulfills [96, Assumption 1] and in particular the monotonicity property, since the offloading cost only increases when more agents offload. Therefore, by [96, Corollary 8.2], we have convergence of the FP algorithm to the unique MF equilibrium. □

Similarly, the cooperative MFC solution has an optimal solution, which will constitute an approximate Pareto optimum in the finite user system.

Theorem 5.3.3. *For distributions μ_0 with finite support, an optimizer π^* of Eq. (5.3.26) exists, and π^* constitutes an ϵ_N-Pareto optimum of the finite N-user system with $\epsilon_N \to 0$ as $N \to \infty$.*

Proof. Existence is trivially guaranteed by the extreme value theorem, since the objective is a continuous function of $\pi \in [0,1]^K$, and $[0,1]^K$ is compact. For approximate Pareto-optimality, see [8, Corollary 1]. □

5.3.4 *Experiments*

In this section, we present numerical simulations for the systems established in the prequel. For the quadratic program MFC, we could apply convex quadratic program solvers if the problem is indeed convex. However, in general the MFC problem may be non-convex and thus results in a NP-hard problem [366]. Still, we again stress that the complexity scales only with the size of $\mathbb{K}$ and remains independent of the number of users N. Therefore, our formulation will be of lower complexity than solving the finite user model for large systems. We do not compare run times, since finite model solvers will trivially exceed the run time of our solution for sufficiently large systems. We may follow any global optimization algorithm, and for simplicity we apply a simple grid search, though more sophisticated algorithms such as Bayesian optimization can easily be substituted.

As can be observed in Figure 5.10, for the competitive MFG problem Eq. (5.3.20), the FP algorithm

$$\pi_{n+1} \equiv \frac{1}{n+1}\left(n\pi_{1:n} + \arg\min_{\pi} J^{\mu_0 \otimes \pi_{1:n}}(\pi)\right) \tag{5.3.34}$$

with the past average policy $\pi_{1:n} \coloneqq \frac{1}{n}\sum_{m=1}^{n} \pi_m$ quickly converges in terms of the exploitability

$$\Delta J(\pi) \coloneqq \max_{\pi^*} J^{\mu_0 \otimes \pi}(\pi^*) - J^{\mu_0 \otimes \pi}(\pi) \tag{5.3.35}$$

which must be equal to zero for an exact equilibrium.

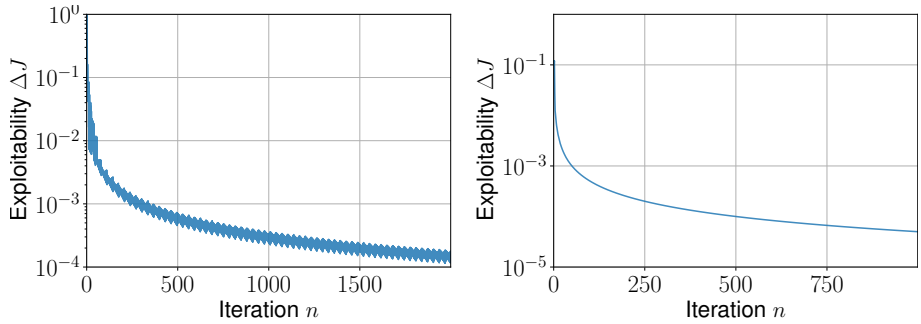


FIGURE 5.10: Learning curve for the exploitability ΔJ in the competitive MFG problem Eq. (5.3.20) (left) over 5000 iterations n using FP. The FP algorithm quickly converges to the equilibrium $\pi^* \approx (1, 0.65, 0)$. Here, we used $(p_1, p_2, p_3) = (0.2, 0.4, 0.4)$, $f_{\text{per}} = 0.5$ and $(C_i)_{i=1,2,3} = ((1,1,1,20),(3,2,1,20),(5,3,1,20))$. Similar results are achieved in the time-stationary MFG problem Eq. (5.3.33), converging to the equilibrium $\pi^* \approx (1, 0.5, 0)$ for $(p_1, p_2, p_3) = (0.2, 0.4, 0.4)$, $f_{\text{per}} = 0.5$, $\lambda \approx 0.225$ and $(C_i)_{i=1,2,3} = ((1,1,5,10),(3,2,5,10),(5,3,5,10))$.

For the parameters given in Figure 5.10, the resulting equilibrium $\pi^* \approx (0, 0.81, 1)$ is intuitive: Only the second configuration splits between offloading and local computation at a ratio that equilibrates offloading and local computation time, since the second configuration has longer offloading times than the first, and longer local computation time than the third. The result are offloading decisions where each UE gains little by deviating.

In Figure 5.11, we can observe the cost function for an illustrative case where $K = 2$. As can be seen in the example, the optimum computational offloading policy is reached at around $\pi^* \approx (0.52, 0)$. Similarly, a solution can be reached for the time-stationary problem at around $\pi^* \approx (0.24, 1)$. Here, we solve the problem for an illustrative 3D example in a few seconds, though similar results can easily be obtained for larger problems. Thus, we obtained nearly optimal offloading decisions, minimizing the average computation times.

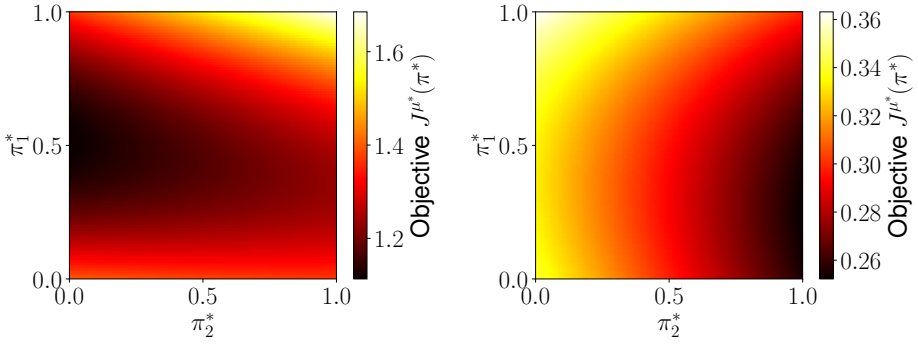


FIGURE 5.11: Exemplary 2D-case for the quadratic MFC problem Eq. (5.3.26) (left) and $K = 2$, reaching the optimal objective of around 1.26 at around $\pi^* \approx (0.52, 0)$. Here, we used $(p_1, p_2) = (0.8, 0.2)$, $f_{\text{per}} = 3$ and $(C_i)_{i=1,2} = ((3, 5, 3, 10), (1.5, 1.5, 5, 25))$. Similar results are achieved for the time-stationary MFC problem Eq. (5.3.32) (right), where we achieve the optimal value 0.25 at $\pi^* \approx (0.24, 1)$ for $(p_1, p_2) = (0.8, 0.2)$, $f_{\text{per}} = 3$, $\lambda = 0.6$ and $(C_i)_{i=1,2} = ((3, 1.5, 5, 12), (1.5, 1, 2, 20))$.

In Figure 5.12, we can observe that the time-homogeneous problem empirically shows a number of jobs in the system that converges to the MF description when rescaled by N, leading us to the conclusion that the MF model we proposed is a good approximation to the finite user system as long as the system is sufficiently large.

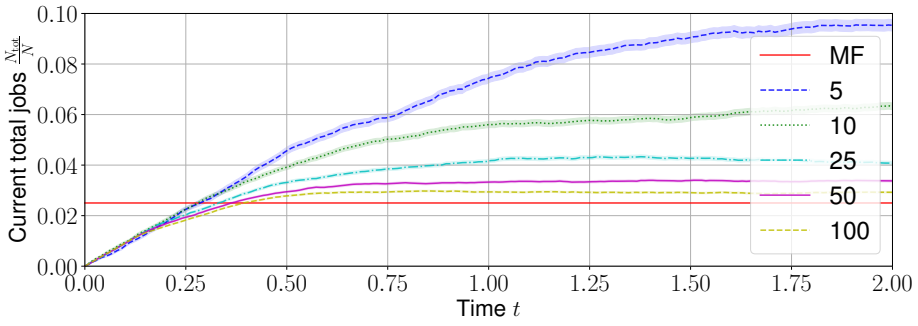


FIGURE 5.12: The evolution of the expected total number of jobs in the queue divided by N with 68% confidence interval, plotted for the configuration from Figure 5.10 and various N from 5 to 100, compared against the stationary MF solution (MF). We average over 5000 sample trajectories. As we consider increasingly large systems, the expected rescaled number of jobs in the system converges to the limiting MF description, letting us conclude that the limiting MF system is a good approximation for the finite user system.

Finally, in Figure 5.13, we can observe (i) the exploitability in the competitive finite user system, i.e. the expected maximum gain by deviating to any other policy in Eq. (5.3.14), and (ii) the deviation

of the objective Eq. (5.3.23) from the computed MF objective Eq. (5.3.26) in the cooperative setting. Here we estimated the exploitability for each value of N by taking the maximum over all pure policies $\pi \in \{0,1\}^{\mathbb{K}}$ over 100 000 samples. Similarly, we estimated the deviation between Eq. (5.3.23) and Eq. (5.3.26) over 20 000 samples of the finite user system. We observe that the exploitability and deviation of objectives tends to zero as the number of agents increases, showing that the MF solution solves the finite system well.

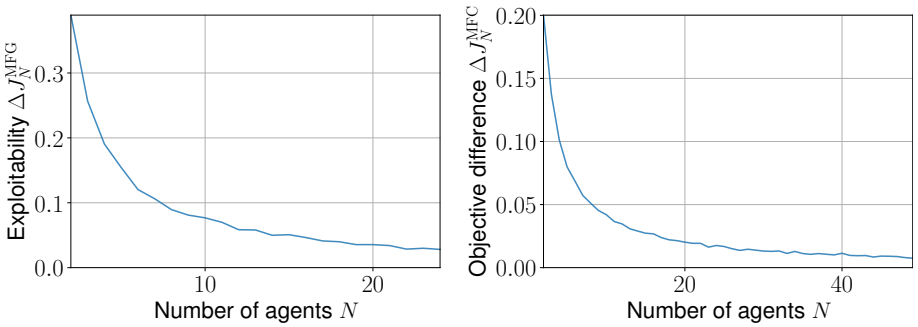


FIGURE 5.13: Comparison of N-agent exploitability in the MFG Eq. (5.3.20) (left), and N-agent objective deviation from the limiting objective in the MFC Eq. (5.3.26) (right) for the configurations used in Figures 5.10 and 5.11. The exploitability quickly decreases to zero, and similarly the cooperative problem is quickly well-approximated by the MFC.

5.3.5 *Summary*

In this work, we have shown the general applicability of rigorous MF frameworks for both competitive and cooperative scenarios in offloading for edge-computing. In particular, we have shown that the MF approximation quickly becomes a good approximation and can reliably be solved with a complexity independent of the number of agents. As a result, we have obtained good and tractable solutions for large-scale, decentralized edge-computing systems. In future work, one could extend rigorous theoretical analysis to the time-stationary case. Other interesting directions could be an extension to Markov-modulated task arrival rates and thereby a non-time-stationary case, or an even more distributed setting with multiple separate limited-access MEC pools. Finally, an application to real systems may be of interest.

5.4 CONCLUSION OF CHAPTER 5

In this chapter, we started by giving an overview on potential applications of large-scale MARL and MF models such as MFGs and MFC. There, we have first given some literature that shows the applicability of MFGs and MFC, e.g., in practical engineering and financial applications. We have also discussed possible future applications in the fields of distributed computing, cyber-physical systems, autonomous mobility and natural or social sciences. With this, we hope that the reader is convinced of the usefulness of MFGs and MFC models.

With the general applicability of MFGs and MFC out of the way, we then focussed on two particular applications of MFGs and MFC. The first exemplary application considered embodied aerial drone swarms, which due to safety constraints must avoid collisions between each other at all cost. There, the difficulty of applying MF models becomes apparent due to the strong interaction of agents in collisions, which violates the typical MF assumptions of weak interaction. The problem is solved by considering the MFC system without collisions and analyzing its error with respect to a practical subroutine that avoids collisions in real swarms, allowing for the application of MFC-based MARL onto real UAV swarms. The applicability was then verified on a real swarm of nano-quadcopters.

As the second exemplary application, we have considered an edge computing scenario where many edge devices must balance the usage of computational offloading resources. There, we have used a unified MF model and compared the cooperative MFC solution and the competitive MFG solution. The system was analzed in its time-stationary equilibrium state, and verified experimentally through numerical simulations. In addition, we have also studied load balancing and network applications in external collaborations [10, 15, 21], which extend the load balancing system in Section 4.1 and the Dec-POMFC model in Section 4.3, but were not presented in this thesis.

We have thus addressed RQ III for applicability of MFGs and MFC and large-scale MARL generally, by surveying a list of potential applications, and specifically, by looking into UAV swarms as well as edge-computing offloading scenarios. In addition to Chapter 5, in Chapters 3 and 4 we have given various applications of our extended models in engineering collective behavior, controlling formations and analyzing epidemics control problems. Concluding this chapter, we have addressed all three of our primary RQs in our thesis. In the following final chapter, we will conclude by giving a summary and discussion of the achieved results, as well as providing an outlook.

6

CONCLUSION AND DISCUSSION

In this thesis, we have addressed the topic of whether general MARL can be scaled through MFG and MFC in a realistic and applicable manner. To begin, in the introduction and Chapter 2 we have briefly given an overview of general MARL models as well as the basic setting of MFGs and MFC. The first two RQs I and II we have asked ourselves were *how to learn in MFGs and MFC,* and *whether the strict limitations of MF models may be relaxed*. The last RQ III we asked ourselves was *how MFGs and MFC can be applied in practice*. Overall, in the previous chapters we have answered our RQs by presenting various methods for learning, extending and applying MF models as follows.

6.1 SUMMARY OF CONTRIBUTIONS

In Chapter 3, we have answered the first two RQs I and II for the competitive MFG case. For RQ I, a regularization-based MFG learning approach for general evolutive MFGs was presented, together with results on the difficulty of commonly assumed FPI contractivity. While regularization trades off optimality of agent policies for convergence, this may be desirable still, as there is still no general but tractable method for solving MFGs without assumed conditions. We then moved on to address RQ II through extension towards GMFGs and M3FGs, which allowed graph-based agent interaction and major agents respectively. The resulting models were then analyzed and solved by extending results for standard MFGs, including propagation of chaos for approximate optimality, and learning algorithms such as the aforementioned regularization or FP. The applicability of MFGs in RQ III was also verified in experiments motivated by practical scenarios such as epidemics control or regulating duopolies.

Moving on to the cooperative MFC case, in Chapter 4 we provided new "hierarchical" MARL algorithms based on MFC, together with approximation guarantees for its PGs, addressing RQ I. The motivation for the MFC approach was given by showing sufficiency for optimality of identical policies in a simplified quasi-static MFC problem with external environment states. We have also analyzed RL for MFC-based MARL on the finite MARL problem, where we showed the approximation of PGs in many-agent systems. For RQ II, similar to the MFG case, we allowed an additional, complex major agent that does not fit into the MF approximation. Going one step further, we also considered the realistic case of partial information, where each agent may only observe limited information in the system. The approximation properties were generalized from basic MFC towards the extended settings, and algorithms were successfully compared against state-of-the-art MARL techniques.

Lastly, in Chapter 5 we briefly explored potential application areas for large-scale MARL as well as MFGs and MFC. We also explicitly considered the scenarios of UAV swarm control and load balancing in edge computing to give an example usage of MFGs and MFC. In the former, we demonstrated how

to apply MFC despite model violations, and in the latter we demonstrated how to apply both MFGs and MFC in practice. The results close the gap between theory of MFGs and MFC and application thereof.

Overall, we have addressed RQs I and II by providing scalable MARL algorithms through MFC in the cooperative case, and equilibrium learning algorithms for MFGs in the competitive case. While general algorithms without additional assumptions for MFGs and more scalable MARL algorithms through MFC still continue to be developed, we hope that our contributions have extended the theory and algorithms of MFGs and MFC to make them more useful. Our last RQ III is not only directly addressed by surveying and exemplarily demonstrating applications in Chapter 5, but also indirectly addressed by Chapter 3 and Chapter 4, by extending the classes of problems that may be solved through MF approximations and algorithms. We hope that this thesis shows how to learn and generalize MFGs and MFC-based techniques for particular settings, enabling their future application in practice and further generalizations.

6.2 OUTLOOK

So far, we find MFGs and MFC already provide the potential to solve a variety of problems, under quite general assumptions of large systems with many agents that interact through their distribution, the MF. As seen throughout the thesis, the MF approximation is quite flexible and can be generalized to particular settings of interest. MF models address most classical challenges of MARL [40], e.g., most notably the combinatorial nature of MARL, by avoiding the exponential blow-up in joint state-actions via reducing large systems to a complexity independent of the number of agents by the MF. Another challenge that is addressed by MFC-based MARL is credit assignment, as credit is assigned to MFC actions, which control all agents at once and receive aggregated credit. This avoids the problem where actions of too many other agents become noise for estimating the goodness of any single agent's taken action. And the classical MARL challenge of non-stationarity is solved in both cooperative MFC and competitive MFG by reducing hard MARL problems into single-agent RL and fixed-point problems respectively, taking into account all the other agents directly. Nonetheless, we believe that many challenges remain, not only for the learning frameworks of MFGs and MFC, but also for the theoretical generality of MF-based large-scale MARL and its applications.

ADDRESSING ALGORITHMIC LIMITATIONS So far, in MFC-based RL the formulation of the MFC MDP is the standard way of obtaining an algorithm for multi-agent control [107, 201]. However, so far most works (except e.g. [367] for MFGs) appear to solve the limiting problem instead of formulating an algorithm for a finite MARL problem without accessing the underlying model. Further, this is usually done by discretization. In Section 4.3, we proposed some first analysis for algorithms learning on the finite MARL problem together with kernel-based parametrizations. However, the analysis remains limited to PG approximations and should be extended towards convergence properties, while the choice of kernel-based parametrizations is still susceptible to manual parameter tuning and was not scaled beyond 5 or 6 dimensions.

On the other hand, in MFGs for competitive games, the MARL problem is classically reduced to a fixed point problem of computing a MF generated by a policy, which needs to be optimal under the MF [23]. As discussed in the prequel, the resulting problem has its own difficulties in solving the general case, and so far can mostly be solved tractably under certain conditions such as monotonicity or contractivity, with uniqueness of the equilibrium. Here, a next step could be considering new methods for the solution of arbitrary discrete-time MFGs in a tractable manner. In the presence of multiple equilibria, the computation of a particular or all equilibria may be of interest. Furthermore,

existing algorithms should also be generalized towards more general MF models discussed in the following, which can be difficult. For instance, the simultaneous presence of stochastic MFs and partial information structure significantly complicates finding optimal MFC policies or equilibria in MFGs, which is not yet addressed by our the approaches in Sections 3.4, 4.2 and 4.3. Lastly, an important recent venue of research considers how to directly design algorithms or learning behaviors in model-*free* MARL manner, that are able to converge to MFE on unknown systems [367]. Such research could be implemented and scaled up, or bridged with other MFG algorithms.

ADDRESSING GENERALITY AND THEORY OF MF APPROXIMATIONS While we have addressed the generality of MF approximations through RQ II to a certain degree, we could not consider all possible generalizations of MF models in this thesis. In particular we have addressed some challenges of dealing with strongly-interacting agents that cannot be summarized by a MF, with agents that are connected on networks or graphs, and with agents that have only partial information of the system. Some important generalizations remain an open question: For example, our partially observable systems remain limited to MFC and with deterministic MFs and could be extended, while the graph-based studies similarly remain limited to MFGs. Further, one could consider various combinations of cooperation and competition between teams of agents such as in [128], applying some of our GMFGs extensions [13, 14, 19] also to MFC, or working on even sparser but more realistic networks [368]. Moreover, so far we have a priori assumed the correct form of MF models for our experiments. In reality, a data-driven approach to choosing or identifying an appropriate MF model may be of relevance.

The theory of MFGs and MFC also remains to be studied more. For example, quantifying the convergence rate of MF models under weaker assumptions could be of interest to improve guarantees of approximate optimality when using MFGs and MFC. Some optimality properties remain to be analyzed more closely, such as the sufficiency of MF-independent policies in standard MFGs, or the open questions for MFC discussed in Appendix E.17. Bridging the gap between the large amount of existing continuous-time probabilistic literature [67] and recent machine learning frameworks in discrete time [25] may be an additional source of algorithmic advances. First results on sample efficiency and complexity classes of MFGs have appeared as another future avenue of research [269]. For MFGs, other equilibria types than Nash also remain to be considered, see e.g. some first works on correlated equilibria [343, 344, 369] where agents obtain advice from a mediator to align their actions, or Stackelberg equilibria with a leader-follower structure [190–192, 194]. As for heterogeneity of agents, it may be interesting to consider limiting MF models under more heterogeneous policies and policy classes for different agents. For example, in MFC and its two-team generalization the heterogeneity of agent policies has been shown to have negligible impact in standard MFC models [128], which could be pursued further.

Moreover, the aforementioned generalizations of MF models may sometimes be applied orthogonally. This means that one could attempt to formulate a maximally general framework such as the partially-observable SG in game theory for MFGs and MFC, similar to the Rainbow DQN algorithm. The usefulness of such a model for MF-based MARL must however be investigated, as one may run into a tradeoff between model generality and tractability. In practice, it could hence be of interest to automatically choose and tune MF-based MARL algorithms, similar to AutoML techniques for supervised / unsupervised machine learning [370].

ADDRESSING IMPORTANT APPLICATIONS VIA MF APPROXIMATIONS Finally, in future work, one could perform extended analyses for any of the scenarios listed in Chapter 5. For example, the exemplary application to edge computing could be supplemented by rigorous theoretical results for

the time-stationary setting, more realistic scenarios without time-stationarity, or an application to real systems. In UAV swarms, related methods could also be applied to realize various intelligent collective behaviors as listed in [332]. Moreover, since their publication, our developed frameworks have already seen application in scenarios such as traffic control [54, 61] or networking [15, 21], which may be further pursued. Many other applications in engineering and finance have also been considered through MFGs and MFC before, but the usage of recently developed learning algorithms may further boost their future applicability in real systems. Lastly, we note that MF theory has also been used in learning theory for analyzing neural networks themselves [371, 372], such that it seems natural to look for applications in the overlap of MF-based analysis and MF-based MARL.

As one particularly relevant and fruitful direction, we also point out MF-based MARL for optimization problems, which is related to the aforementioned MF-based analysis of neural networks and learning. Many problems are well-known to be difficult to solve in full generality. For example, graph coloring problems [373], vehicle routing problems [308] or facility location problems and clustering [374] are known to be NP-hard. Oftentimes, scalability of existing algorithmic solutions is therefore limited. Certain special cases amenable to MF approximation may hence be of interest, whenever applicable. And more generally, it may also be of interest to apply MFC to general optimization, which has great potential for applications. For example, particle swarm optimization algorithms themselves could be optimized through similar partially-observable approaches as the Dec-POMFC model in Section 4.3, for which first results using MF models have been obtained [375].

APPENDICES

A

APPENDIX A: SUPPLEMENTARY DETAILS ON SECTION 3.1

A.1 COMPLETENESS OF MEAN FIELD AND POLICY SPACES

Lemma A.1.1. *The metric spaces* (Π, d_Π) *and* $(\mathcal{M}, d_\mathcal{M})$ *are complete metric spaces.*

Proof. The metric space $(\mathcal{M}, d_\mathcal{M})$ is a complete metric space. Let $(\mu^n)_{n\in\mathbb{N}} \in \mathcal{M}^\mathbb{N}$ be a Cauchy sequence of MFs. Then by definition, for any $\varepsilon > 0$ there exists integer $N > 0$ such that for any $m, n > N$ we have

$$\begin{aligned} & d_\mathcal{M}(\mu^n, \mu^m) < 0.5\varepsilon \\ \implies \forall t \in \mathcal{T} : d_{TV}(\mu_t^n, \mu_t^m) &= \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu_t^n(x) - \mu_t^m(x)| < 0.5\varepsilon \\ \implies \forall t \in \mathcal{T}, x \in \mathcal{X} &: |\mu_t^n(x) - \mu_t^m(x)| < \varepsilon\,. \end{aligned}$$

By completeness of $\mathbb{R}$ there exists the limit of $(\mu_t^n(x))_{n\in\mathbb{N}}$ for all $t \in \mathcal{T}, x \in \mathcal{X}$, suggestively denoted by $\mu_t(x)$. The MF $\mu = \{\mu_t\}_{t\in\mathcal{T}}$ with the probabilities defined by the aforementioned limits fulfills $\mu^n \to \mu$ and is in $\mathcal{M}$, showing completeness of $\mathcal{M}$.

We do this analogously for (Π, d_Π). Thus, (Π, d_Π) and $(\mathcal{M}, d_\mathcal{M})$ are complete metric spaces. □

A.2 LIPSCHITZ CONTINUITY

Lemma A.2.1. *Assume bounded and Lipschitz functions $f : X \to \mathbb{R}$ and $g : X \to \mathbb{R}$ mapping from a metric space (X, d_X) into $\mathbb{R}$ with Lipschitz constants C_f, C_g and bounds $|f(x)| \leq M_f$, $|g(x)| \leq M_g$. The sum of both functions $f + g$, the product of both functions $f \cdot g$ and the maximum of both functions $\max(f, g)$ are all Lipschitz and bounded with Lipschitz constants $C_f + C_g$, $(M_f C_g + M_g C_f)$, $\max(C_f, C_g)$ and bounds $M_f + M_g$, $M_f M_g$, $\max(M_f, M_g)$.*

Proof. Let $x, y \in X$ be arbitrary. By the triangle inequality, we obtain

$$|f(x) + g(x) - (f(y) + g(y))| \leq |f(x) - f(y)| + |g(x) - g(y)| \leq (C_f + C_g) d_X(x, y)\,.$$

Analogously, we obtain

$$\begin{aligned} |f(x)g(x) - f(y)g(y)| &\leq |f(x)g(x) - f(x)g(y)| + |f(x)g(y) - f(y)g(y)| \\ &\leq (M_f C_g + M_g C_f) d_X(x, y)\,. \end{aligned}$$

For the maximum of both functions, consider case by case. If $f(x) \geq g(x)$ and $f(y) \geq g(y)$ we obtain

$$|\max(f(x), g(x)) - \max(f(y), g(y))| = |f(x) - f(y)| \leq C_f d_X(x, y)$$

and analogously for $g(x) \geq f(x)$ and $g(y) \geq f(y)$

$$|\max(f(x), g(x)) - \max(f(y), g(y))| = |g(x) - g(y)| \leq C_g d_X(x, y)\,.$$

On the other hand, if $g(x) < f(x)$ and $g(y) \geq f(y)$, we have either $g(y) \geq f(x)$ and thus

$$\begin{aligned} |\max(f(x), g(x)) - \max(f(y), g(y))| &= |f(x) - g(y)| = g(y) - f(x) < g(y) - g(x) \\ &\leq C_g d_X(x, y) \end{aligned}$$

or $g(y) < f(x)$ and thus

$$\begin{aligned} |\max(f(x), g(x)) - \max(f(y), g(y))| &= |f(x) - g(y)| = f(x) - g(y) \leq f(x) - f(y) \\ &\leq C_f d_X(x, y)\,. \end{aligned}$$

The case for $f(x) < g(x)$ and $f(y) \geq g(y)$ as well as boundedness is analogous. □

A.3 PROOF OF PROPOSITION 3.1.1

Proof. Since we work with finite $\mathcal{T}, \mathcal{X}, \mathcal{U}$, we identify the space of MFs $\mathcal{M}$ with the $|\mathcal{T}|(|\mathcal{X}| - 1)$-dimensional simplex $S_{|\mathcal{T}|(|\mathcal{X}|-1)} \subseteq \mathbb{R}^{|\mathcal{T}|(|\mathcal{X}|-1)}$ via the values of the probability mass functions at all times and states. Analogously the space of policies Π is identified with $S_{|\mathcal{T}||\mathcal{X}|(|\mathcal{U}|-1)} \subseteq \mathbb{R}^{|\mathcal{T}||\mathcal{X}|(|\mathcal{U}|-1)}$.

Define the set-valued map $\hat{\Gamma} : S_{|\mathcal{T}||\mathcal{X}|(|\mathcal{U}|-1)} \to 2^{S_{|\mathcal{T}||\mathcal{X}|(|\mathcal{U}|-1)}}$ mapping from a policy π represented by the input vector, to the set of vector representations of optimal policies in the MDP induced by $\Psi(\pi)$.

A policy π is optimal in the MDP induced by $\mu \in \mathcal{M}$ if and only if its value function defined by

$$V^\pi(\mu, t, s) = \sum_{u \in \mathcal{U}} \pi_t(u \mid x) \left(r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) V^\pi(\mu, t+1, x') \right),$$

is equal to the optimal action-value function defined by

$$V^*(\mu, t, s) = \max_{u \in \mathcal{U}} \left(r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) V^*(\mu, t+1, x') \right)$$

for every $t \in \mathcal{T}, x \in \mathcal{X}$, with terminal conditions $V^*(\mu, T, s) \equiv V^\pi(\mu, T, s) \equiv 0$. Moreover, an optimal policy always exists. For more details, see e.g. [68]. Define the optimal action-value function for every $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ via

$$Q^*(\mu, t, x, u) = r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) V^*(\mu, t+1, x')$$

with terminal condition $Q^*(\mu, T, x, u) \equiv 0$. Then, the following lemma characterizes optimality of policies.

Lemma A.3.1. *A policy π fulfills $\pi \in \hat{\Gamma}(\hat{\pi})$ if and only if*

$$\pi_t(u \mid x) > 0 \implies a \in \arg\max_{u' \in \mathcal{U}} Q^*(\Psi(\hat{\pi}), t, x, u')$$

for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.

Proof. To see the implication, consider $\pi \in \hat{\Gamma}(\hat{\pi})$. Then, if the right-hand side was false, there exists a maximal $t \in \mathcal{T}$ and $x \in \mathcal{X}, u \in \mathcal{U}$ such that $\pi_t(u \mid x) > 0$ but $a \notin \arg\max_{u' \in \mathcal{U}} Q^*(\Psi(\hat{\pi}), t, x, u')$. Since for any $t' > t$ we have optimality, $V^\pi(\mu, t+1, x') = V^*(\mu, t+1, x')$ by induction. However, $V^\pi(\mu, t, s) < V^*(\mu, t, s)$ since the suboptimal action is assigned positive probability, contradicting optimality of π. On the other hand, if the right-hand side is true, then $V^\pi(\mu, t, s) = V^*(\mu, t, s)$ by induction, which implies that π is optimal. ∎

We will now check that the requirements of Kakutani's fixed point theorem hold for $\hat{\Gamma}$. The finite-dimensional simplices are convex, closed and bounded, hence compact. $\hat{\Gamma}$ maps to a non-empty set, as the induced MF is uniquely defined and any finite MDP (induced by this MF) has an optimal policy.

For any π, $\hat{\Gamma}(\pi)$ is convex, since the set of optimal policies is convex as shown in the following. Consider a convex combination $\tilde{\pi} = \lambda\pi + (1-\lambda)\pi'$ of optimal policies π, π' for $\lambda \in [0, 1]$. Then, the resulting policy will be optimal, since we have

$$\tilde{\pi}_t(u \mid x) > 0 \implies \pi_t(u \mid x) > 0 \vee \pi'_t(u \mid x) > 0 \implies a \in \arg\max_{u \in \mathcal{U}} Q^*(\Psi(\hat{\pi}), t, x, u)$$

for any $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ and thus optimality by Lemma A.3.1.

Finally, we show that $\hat{\Gamma}$ has a closed graph. Consider arbitrary sequences $(\pi_n, \pi'_n) \to (\pi, \pi')$ with $\pi'_n \in \hat{\Gamma}(\pi_n)$. It is then sufficient to show that $\pi' \in \hat{\Gamma}(\pi)$. By the standing assumption, we have continuity of Ψ and $\mu \to Q^*(\mu, t, x, u)$ for any $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$, as sums, products and compositions of continuous functions remain continuous. Therefore, the composition $\pi \to Q^*(\Psi(\pi), t, x, u)$ is continuous. To show that $\pi' \in \hat{\Gamma}(\pi)$, assume that $\pi' \notin \hat{\Gamma}(\pi)$. By Lemma A.3.1 there exists $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ such that $\pi'_t(u \mid x) > 0$ and further there

exists $u' \in \mathcal{U}$ such that $Q^*(\Psi(\pi), t, x, u') > Q^*(\Psi(\pi), t, x, u)$. Fix such an $u' \in \mathcal{U}$. Let $\delta \equiv Q^*(\Psi(\pi), t, x, u') - Q^*(\Psi(\pi), t, x, u)$, then by continuity there exists $\varepsilon > 0$ such that for all $\hat{\pi} \in \Pi$ we have

$$d_\Pi(\hat{\pi}, \pi) < \varepsilon \implies |Q^*(\Psi(\hat{\pi}), t, x, u) - Q^*(\Psi(\pi), t, x, u)| < \frac{\delta}{2}.$$

By convergence, there is an integer $N \in \mathbb{N}$ such that for all $n > N$ we have $d_\Pi(\pi_n, \pi) < \varepsilon$ and therefore

$$Q^*(\Psi(\pi_n), t, x, u') > Q^*(\Psi(\pi), t, x, u') - \frac{\delta}{2} = Q^*(\Psi(\pi), t, x, u) + \frac{\delta}{2} > Q^*(\Psi(\pi_n), t, x, u).$$

Since $(\pi'_n)_t(u \mid x) \to \pi'_t(u \mid x) > 0$, there also exists $M \in \mathbb{N}$ such that for all $m > M$,

$$\left|(\pi'_m)_t(u \mid x) - \pi'_t(u \mid x)\right| < \pi'_t(u \mid x).$$

Let $n > \max(N, M)$, then it follows that $(\pi'_n)_t(u \mid x) > 0$ but $a \notin \arg\max_{u' \in \mathcal{U}} Q^*(\Psi(\pi), t, x, u')$ since we have $Q^*(\Psi(\pi_n), t, x, u') > Q^*(\Psi(\pi_n), t, x, u)$, contradicting $\pi'_n \in \hat{\Gamma}(\pi_n)$ by Lemma A.3.1. Hence, $\hat{\Gamma}$ must have a closed graph.

By Kakutani's fixed point theorem, there exists a fixed point π^* that generates some MF $\Psi(\pi^*)$. The associated pair $(\pi^*, \Psi(\pi^*))$ is an MFE by definition. □

A.4 PROOF OF PROPOSITION 3.1.3

Proof. The space of MFs $(\mathcal{M}, d_\mathcal{M})$ is equivalent to convex and compact finite-dimensional simplices. In this representation, each coordinate of the operators $\tilde{\Gamma}_\eta(\mu)$ and $\Gamma_\eta(\mu)$ consists of compositions, sums and products of continuous functions, since the functions $r(x, u, \mu_t)$ and $p(x' \mid x, u, \mu_t)$ are assumed to be continuous. Existence of a fixed point follows immediately by Brouwer's fixed point theorem. □

A.5 PROOF OF THEOREM 3.1.1

Proof. The proof is a slightly simplified version of the one found in [24]. Note that we require the results later, so for convenience we give the full details.

The empirical measure $\mu[\mathbf{x}_t]$ is a random variable on $\mathcal{P}(\mathcal{X})$, i.e. its law $\mathcal{L}(\mu[\mathbf{x}_t]) \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ is a distribution over probability measures. Since we want to show convergence of the empirical measure to the MF, let us pick a metric on $\mathcal{P}(\mathcal{P}(\mathcal{X}))$. Remember that we metrized $\mathcal{P}(\mathcal{X})$ with the total variation distance. We metrize $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ with the 1-Wasserstein metric defined for any $\Phi, \Psi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ by the infimum over couplings

$$W_1(\Phi, \Psi) \equiv \inf_{\mathcal{L}(X_1)=\Phi, \mathcal{L}(X_2)=\Psi} \mathbb{E}\left[d_{TV}(X_1, X_2)\right].$$

Lemma A.5.1. *Let $\{\Phi_n\}_{n\in\mathbb{N}}$ be a sequence of measures with $\Phi_n \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ for all $n \in \mathbb{N}$. Further, let $\mu \in \mathcal{P}(\mathcal{X})$ arbitrary. Then, the following are equivalent.*

(a) $W_1(\Phi_n, \delta_\mu) \to 0$ *as* $n \to \infty$

(b) $\mathbb{E}\left[|F(X_n) - F(X)|\right] \to 0$ *as* $n \to \infty$ *for any continuous, bounded* $F : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$*, any sequence* $\{X_n\}_{n\in\mathbb{N}}$ *of* $\mathcal{P}(\mathcal{X})$*-valued random variables and any* $\mathcal{P}(\mathcal{X})$*-valued random variable* X *with* $\mathcal{L}(X_n) = \Phi_n$ *and* $\mathcal{L}(X) = \delta_\mu$*.*

(c) $\mathbb{E}\left[|X_n(f) - X(f)|\right] \to 0$ *as* $n \to \infty$ *for any* $f : \mathcal{X} \to \mathbb{R}$*, any sequence* $\{X_n\}_{n\in\mathbb{N}}$ *of* $\mathcal{P}(\mathcal{X})$*-valued random variables and any* $\mathcal{P}(\mathcal{X})$*-valued random variable* X *with* $\mathcal{L}(X_n) = \Phi_n$ *and* $\mathcal{L}(X) = \delta_\mu$*.*

Proof. Define the only possible coupling $\Delta_n \equiv \Phi_n \times \delta_\mu$.

(b), (c) $\implies$ (a):

Define $F_s(x) \equiv x(x)$ and $f_s(x') \equiv \mathbf{1}_{\{s\}}(x')$ for all $x \in \mathcal{X}$, where F_s is continuous. By assumption,

$$\begin{aligned} W_1(\Phi_n, \delta_\mu) &= \inf_{\mathcal{L}(X_n)=\Phi_n, \mathcal{L}(X)=\delta_\mu} \mathbb{E}\left[d_{TV}(X_n, X)\right] \\ &= \frac{1}{2}\int_{\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X})} \sum_{x\in\mathcal{X}} |X_n(x) - X(x)|\, d\Delta_n \\ &= \frac{1}{2}\sum_{x\in\mathcal{X}} \mathbb{E}\left[|X_n(x) - X(x)|\right] \to 0 \end{aligned}$$

since for any $x \in \mathcal{X}$, we have

$$\mathbb{E}\left[|X_n(x) - X(x)|\right] = \mathbb{E}\left[|F_s(X_n) - F_s(X)|\right] = \mathbb{E}\left[|X_n(f_s) - X(f_s)|\right] .$$

(a) $\implies$ (b), (c):

We have

$$\begin{aligned} \mathbb{E}\left[|F(X_n) - F(X)|\right] &= \int_{\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{X})} |F(\nu) - F(\nu')|\, \Delta_n(d\nu, d\nu') \\ &= \int_{\mathcal{P}(\mathcal{X})} |F(\nu) - F(\mu)|\, \Phi_n(d\nu) \\ &\to \int_{\mathcal{P}(\mathcal{X})} |F(\nu) - F(\mu)|\, \delta_\mu(d\nu) = 0 \end{aligned}$$

by continuity and boundedness of $|F(\nu) - F(\mu)|$, and convergence in W_1 implying weak convergence. Analogously,

$$\mathbb{E}\left[|X_n(f) - X(f)|\right] = \int_{\mathcal{P}(\mathcal{X})} |\nu(f) - \mu(f)|\, \Phi_n(d\nu) \to \int_{\mathcal{P}(\mathcal{X})} |\nu(f) - \mu(f)|\, \delta_\mu(d\nu) = 0$$

since f and thus $|\nu(f) - \mu(f)|$ is automatically bounded from finiteness of $\mathcal{X}$, and $\nu(f) = \sum_{x\in\mathcal{X}} \nu(x)f(x) \to \sum_{x\in\mathcal{X}} \mu(x)f(x)$ as $\nu \to \mu$ in total variation distance implies continuity of $|\nu(f) - \mu(f)|$. ∎

First, it is shown that when all other agents follow the same policy π, then the empirical distribution is essentially the deterministic MF as $N \to \infty$, i.e. $\mathcal{L}(\mu[\mathbf{x}_t]) \to \mathcal{L}(\mu_t) \equiv \delta_{\mu_t}$ with $\mu = \Psi(\pi)$

Lemma A.5.2. *Consider a set of policies* $(\tilde{\pi}, \pi, \ldots, \pi) \in \Pi^N$ *for all agents. Under this set of policies, the law of the empirical distribution* $\mathcal{L}(\mu[\mathbf{x}_t]) \in \mathcal{P}(\mathcal{M})$ *converges to* δ_{μ_t} *where* $\mu = \Psi(\pi)$ *as* $N \to \infty$ *in 1-Wasserstein distance.*

Proof. Define the Markov kernel $P^{\pi}_{t,\nu}$ such that its probability mass function fulfills

$$P^{\pi}_{t,\nu}(x' \mid x) \equiv \sum_{u \in \mathcal{U}} \pi_t(u \mid x) p(x' \mid x, u, \nu)$$

for any $t \in \mathcal{T}, x \in \mathcal{X}, \nu \in \mathcal{P}(\mathcal{X}), \pi \in \Pi$ and analogously

$$\tilde{\nu} P^{\pi}_{t,\nu}(x') \equiv \sum_{x \in \mathcal{X}} \tilde{\nu}(x) \sum_{u \in \mathcal{U}} \pi_t(u \mid x) p(x' \mid x, u, \nu)$$

for any $\tilde{\nu} \in \mathcal{P}(\mathcal{X})$. Note that $\mu_{t+1} = \mu_t P^{\pi}_{t,\mu_t}(g)$ for MF $\mu = \Psi(\pi)$ induced by π.

We show that $\mathbb{E}\left[|\mu[\mathbf{x}_t](f) - \mu_t(f)|\right] \to 0$ as $N \to \infty$ for any function $f : \mathcal{X} \to \mathbb{R}$ and any time $t \in \mathcal{T}$. From this, the desired result follows by Lemma A.5.1. Since $\mu[\mathbf{x}_t] \equiv \frac{1}{N}\sum_{i=1}^{N} \delta_{x_t^i}$ and $x_0^i \sim \mu_0$ we have at time $t = 0$ that

$$\lim_{N\to\infty} \mathbb{E}\left[|\mu[\mathbf{x}_0](f) - \mu_0(f)|\right] = \lim_{N\to\infty} \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} f(x_0^i) - \mathbb{E}\left[f(x_0^i)\right]\right|\right] = 0$$

by the strong LLN and the dominated convergence theorem.

Assuming this holds for t, then for $t+1$ we have

$$\begin{aligned}
\mathbb{E}\left[|\mu[\mathbf{x}_{t+1}](f) - \mu_{t+1}(f)|\right] \leq\ & \mathbb{E}\left[|\mu[\mathbf{x}_{t+1}](f) - \check{\mu}[\mathbf{x}_{t+1}](f)|\right] \\
& + \mathbb{E}\left[\left|\check{\mu}[\mathbf{x}_{t+1}](f) - \check{\mu}[\mathbf{x}_t] P^{\pi}_{t,\mu[\mathbf{x}_t]}(f)\right|\right] \\
& + \mathbb{E}\left[\left|\check{\mu}[\mathbf{x}_t] P^{\pi}_{t,\mu[\mathbf{x}_t]}(f) - \mu[\mathbf{x}_t] P^{\pi}_{t,\mu[\mathbf{x}_t]}(f)\right|\right] \\
& + \mathbb{E}\left[\left|\mu[\mathbf{x}_t] P^{\pi}_{t,\mu[\mathbf{x}_t]}(f) - \mu_t P^{\pi}_{t,\mu_t}(f)\right|\right]
\end{aligned}$$

where we defined $\check{\mu}[\mathbf{x}_t] \equiv \frac{1}{N-1}\sum_{i=2}^{N} \delta_{x_t^i}$.

For the first term, we have as $N \to \infty$

$$\begin{aligned}
\mathbb{E}\left[|\mu[\mathbf{x}_{t+1}](f) - \check{\mu}[\mathbf{x}_{t+1}](f)|\right] &= \mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} f(x_{t+1}^i) - \frac{1}{N-1}\sum_{i=2}^{N} f(x_{t+1}^i)\right|\right] \\
&\leq \frac{1}{N}\mathbb{E}\left[\left|f(x_{t+1}^1)\right|\right] + \left|\frac{1}{N} - \frac{1}{N-1}\right| \sum_{i=2}^{N} \mathbb{E}\left[\left|f(x_{t+1}^i)\right|\right] \\
&\leq \left(\frac{1}{N} + \frac{N-1}{N(N-1)}\right) \max_{x\in\mathcal{X}} |f(x)| \to 0\,.
\end{aligned}$$

For the second term, as $N \to \infty$ we have by Jensen's inequality and bounds $|f| \leq M_f$ (by finiteness of $\mathcal{X}$)

$$\begin{aligned}
&\mathbb{E}\left[\left|\check{\mu}[\mathbf{x}_{t+1}](f) - \check{\mu}[\mathbf{x}_t] P^{\pi}_{t,\check{\mu}[\mathbf{x}_t]}(f)\right|\right]^2 \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\left|\check{\mu}[\mathbf{x}_{t+1}](f) - \check{\mu}[\mathbf{x}_t] P^{\pi}_{t,\check{\mu}[\mathbf{x}_t]}(f)\right| \mid x_t\right]\right]^2 \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\left|\frac{1}{N-1}\sum_{i=2}^{N} \left(f(x_{t+1}^i) - \mathbb{E}\left[f(x_{t+1}^i) \mid x_t\right]\right)\right| \mid x_t\right]\right]^2
\end{aligned}$$

$$\leq \frac{1}{(N-1)^2}\sum_{i=2}^{N}\mathbb{E}\left[\mathbb{E}\left[\left(f(x_{t+1}^i)-\mathbb{E}\left[f(x_{t+1}^i)\mid x_t\right]\right)^2\mid x_t\right]\right]$$
$$\leq \frac{1}{N-1}\cdot 4M_f^2 \to 0\,.$$

For the third term, we again have as $N\to\infty$

$$\begin{aligned}
&\mathbb{E}\left[\left|\check{\mu}[\mathbf{x}_t]P^{\pi}_{t,\mu[\mathbf{x}_t]}(f)-\mu[\mathbf{x}_t]P^{\pi}_{t,\mu[\mathbf{x}_t]}(f)\right|\right]\\
&\quad=\mathbb{E}\left[\left|\sum_{x\in\mathcal{X}}(\check{\mu}[\mathbf{x}_t](x)-\mu[\mathbf{x}_t](x))\sum_{u\in\mathcal{U}}\pi_t(u\mid x)\sum_{x'\in\mathcal{X}}p(x'\mid x,u,\mu[\mathbf{x}_t])f(x')\right|\right]\\
&\quad\leq\mathbb{E}\left[\left|\left(\frac{1}{N-1}-\frac{1}{N}\right)\sum_{i=2}^{N}\sum_{u\in\mathcal{U}}\pi_t(u\mid x_t^i)\sum_{x'\in\mathcal{X}}p(x'\mid x_t^i,u,\mu[\mathbf{x}_t])f(x')\right|\right]\\
&\qquad+\mathbb{E}\left[\left|\frac{1}{N}\sum_{u\in\mathcal{U}}\pi_t(u\mid x_t^1)\sum_{x'\in\mathcal{X}}p(x'\mid x_t^1,u,\mu[\mathbf{x}_t])f(x')\right|\right]\\
&\quad\leq\left(\frac{N-1}{N(N-1)}+\frac{1}{N}\right)\max_{x\in\mathcal{X}}|f(x)|\to 0\,.
\end{aligned}$$

For the fourth term, define $F:\mathcal{P}(\mathcal{X})\to\mathbb{R}$, $F(\nu)=\nu P^{\pi}_{t,\nu}(f)$ and observe that F is continuous, since $\nu\to\nu'$ if and only if $\nu(x)\to\nu'(x)$ for all $x\in\mathcal{X}$, and therefore (as p is assumed continuous by Assumption 1)

$$F(\nu)=\nu P^{\pi}_{t,\nu}(f)=\sum_{x\in\mathcal{X}}\nu(x)\sum_{u\in\mathcal{U}}\pi_t(u\mid x)\sum_{x'\in\mathcal{X}}p(x'\mid x,u,\nu)f(x')$$

is continuous. By Lemma A.5.1, we have from the induction hypothesis $\mu_t^N=\mu[\mathbf{x}_t]\to\mu_t$ that

$$\mathbb{E}\left[\left|\mu[\mathbf{x}_t]P^{\pi}_{t,\mu[\mathbf{x}_t]}(f)-\mu_t P^{\pi}_{t,\mu_t}(f)\right|\right]\to 0\,.$$

Therefore, $\mathbb{E}\left[|\mu[\mathbf{x}_{t+1}](f)-\mu_{t+1}(f)|\right]\to 0$ which implies the desired result by induction. ■

Consider the case where all agents follow a set of policies $(\pi^N,\pi,\ldots,\pi)\in\Pi^N$ for each $N\in\mathbb{N}$. Define new single-agent random variables x_t^{μ} and u_t^{μ} with $x_0^{\mu}\sim\mu_0$ and

$$\mathbb{P}(u_t^{\mu}=u\mid x_t^{\mu}=x)=\pi_t^N(u\mid x),$$
$$\mathbb{P}(x_{t+1}^{\mu}=x'\mid x_t^{\mu}=x,u_t^{\mu}=u)=p(x'\mid x,u,\mu_t)\,,$$

where the deterministic MF μ is used instead of the empirical distribution.

Lemma A.5.3. *Consider an equicontinuous, uniformly bounded family of functions $\mathcal{F}$ on $\mathcal{P}(\mathcal{X})$ and define*

$$F_t(\nu)\equiv\sup_{f\in\mathcal{F}}|f(\nu)-f(\mu_t)|$$

for any $t\in\mathcal{T}$. Then, F_t is continuous and bounded and by Lemma A.5.1 we have

$$\lim_{N\to\infty}\mathbb{E}\left[\sup_{f\in\mathcal{F}}|f(\mu[\mathbf{x}_t])-f(\mu)|\right]=0$$

Proof. F_t is continuous, since for $\nu_n \to \nu$

$$|F_t(\nu_n) - F_t(\nu)| = \left| \sup_{f\in\mathcal{F}} |f(\nu_n) - f(\mu_t)| - \sup_{f\in\mathcal{F}} |f(\nu) - f(\mu_t)| \right| \leq \sup_{f\in\mathcal{F}} |f(\nu_n) - f(\nu)| \to 0$$

by equicontinuity. Further, F_t is bounded since $|F_t(\nu)| \leq \sup_{f\in\mathcal{F}} |f(\nu)| + |f(\mu_t)|$ is uniformly bounded. By Lemma A.5.2, we have $W_1(\mu[\mathbf{x}_t], \delta_{\mu_t}) \to 0$ as $N \to \infty$, therefore Lemma A.5.1 applies. ■

Lemma A.5.4. *Suppose that at some time $t \in \mathcal{T}$, it holds that*

$$\lim_{N\to\infty} \left| \mathcal{L}(x_t^1)(g_N) - \mathcal{L}(x_t^\mu)(g_N) \right| = 0$$

for any sequence of functions $\{g_N\}_{N\in\mathbb{N}}$ from $\mathcal{X}$ to $\mathbb{R}$ that is uniformly bounded. Then, we have

$$\lim_{N\to\infty} \left| \mathcal{L}(x_t^1, \mu[\mathbf{x}_t])(T_N) - \mathcal{L}(x_t^\mu, \mu_t)(T_N) \right| = 0$$

for any sequence of functions $\{T_N\}_{N\in\mathbb{N}}$ from $\mathcal{X} \times \mathcal{P}(\mathcal{X})$ to $\mathbb{R}$ that is equicontinuous and uniformly bounded.

Proof. We have

$$\begin{aligned}
&\left| \mathcal{L}(x_t^1, \mu[\mathbf{x}_t])(T_N) - \mathcal{L}(x_t^\mu, \mu_t)(T_N) \right| \\
&\quad \leq \left| \mathcal{L}(x_t^1, \mu[\mathbf{x}_t])(T_N) - \mathcal{L}(x_t^1, \mu_t)(T_N) \right| + \left| \mathcal{L}(x_t^1, \mu_t)(T_N) - \mathcal{L}(x_t^\mu, \mu_t)(T_N) \right|
\end{aligned}$$

The first term becomes

$$\begin{aligned}
&\left| \mathcal{L}(x_t^1, \mu[\mathbf{x}_t])(T_N) - \mathcal{L}(x_t^1, \mu_t)(T_N) \right| \\
&\quad = \left| \int T_N(x,\nu) \mathcal{L}(x_t^1, \mu[\mathbf{x}_t])(\mathrm{d}x, \mathrm{d}\nu) - \int T_N(x,\nu) \mathcal{L}(x_t^1, \mu_t)(\mathrm{d}x, \mathrm{d}\nu) \right| \\
&\quad \leq \mathbb{E}\left[\mathbb{E}\left[\left| T_N(x_t^1, G_{x_t}^N) - T_N(x_t^1, \mu_t) \right| x_t^1 \right] \right] \\
&\quad \leq \mathbb{E}\left[\sup_{f\in\{T_N(x,\cdot)\}_{x\in\mathcal{X}, N\in\mathbb{N}}} \left| f(G_{x_t}^N) - f(\mu_t) \right| \right] \to 0
\end{aligned}$$

by Lemma A.5.3, since $\{T_N\}_{N\in\mathbb{N}}$ is equicontinuous and uniformly bounded. Similarly for the second term,

$$\left| \mathcal{L}(x_t^1, \mu_t)(T_N) - \mathcal{L}(x_t^\mu, \mu_t)(T_N) \right| = \left| \mathbb{E}\left[T_N(x_t^1, \mu_t) - T_N(x_t^\mu, \mu_t) \right] \right| \to 0$$

by the assumption, since T_N fulfills the condition of being uniformly bounded. ■

Lemma A.5.5. *For any sequence $\{g_N\}_{N\in\mathbb{N}}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ that is uniformly bounded, we have*

$$\lim_{N\to\infty} \left| \mathcal{L}(x_t^1)(g_N) - \mathcal{L}(x_t^\mu)(g_N) \right| = 0$$

for all times $t \in \mathcal{T}$.

Proof. Define $l_{N,t}$ as

$$l_{N,t}(x,\nu) \equiv \sum_{u\in\mathcal{U}} \pi_t^N(u \mid x) \sum_{x'\in\mathcal{X}} p(x' \mid x,u,\nu) g_N(x')\,.$$

$\{l_{N,t}(x,\cdot)\}_{x\in\mathcal{X},N\in\mathbb{N}}$ is equicontinuous, since for any $\nu,\nu' \in \mathcal{M}$ with $d_{TV}(\nu,\nu') \to 0$,

$$\begin{aligned}
&\sup_{x\in\mathcal{X},N\in\mathbb{N}} \left|l_{N,t}(x,\nu) - l_{N,t}(x,\nu')\right| \\
&\leq M_g \sup_{x\in\mathcal{X},N\in\mathbb{N}} \left|\sum_{u\in\mathcal{U}} \pi_t^N(u \mid x) \sum_{x'\in\mathcal{X}} \big(p(x' \mid x,u,\nu) - p(x' \mid x,u,\nu')\big)\right| \\
&\leq M_g|\mathcal{X}| \max_{x\in\mathcal{X}} \max_{u\in\mathcal{U}} \max_{x'\in\mathcal{X}} |p(x' \mid x,u,\nu) - p(x' \mid x,u,\nu')| \to 0
\end{aligned}$$

since $|g_N| < M_g$ is uniformly bounded and p is continuous by assumption. Furthermore, $l_{N,t}(x,\nu)$ is always uniformly bounded by M_g. Now the result can be shown by induction.

For $t=0$, $\mathcal{L}(x_0^\mu) = \mathcal{L}(x_0^1)$ fulfills the hypothesis. Assume this holds for t, then

$$\left|\mathcal{L}(x_{t+1}^1)(g_N) - \mathcal{L}(x_{t+1}^\mu)(g_N)\right| = \left|\mathcal{L}(x_t^1,\mu[\mathbf{x}_t])(l_{N,t}) - \mathcal{L}(x_t^\mu,\mu_t)(l_{N,t})\right| \to 0$$

as $N\to\infty$ by Lemma A.5.4. ■

Thus, for any sequence of policies $\{\pi^N\}_{N\in\mathbb{N}}$ with $\pi^N \in \Pi$ for all $N\in\mathbb{N}$, the achieved return of the N-agent game converges to the return of the MFG under the MF generated by the other agent's policy π as $N\to\infty$.

Lemma A.5.6. *Let $\{\pi^N\}_{N\in\mathbb{N}}$ with $\pi^N \in \Pi$ for all $N\in\mathbb{N}$ be an arbitrary sequence of policies and $\pi\in\Pi$ an arbitrary policy. Further, let the MF $\mu = \Psi(\pi)$ be generated by π. Then, under the joint policy $(\pi^N,\pi,\dots,\pi)$, we have as $N\to\infty$ that*

$$\left|J_1^N(\pi^N,\pi,\dots,\pi) - J^\mu(\pi^N)\right| \to 0\,.$$

Proof. Define for any $t\in\mathcal{T}$, $N\in\mathbb{N}$

$$r_{\pi_t^N}(x,\nu) \equiv \sum_{u\in\mathcal{U}} r(x,u,\nu)\pi_t^N(u \mid x)$$

such that the family $\{r_{\pi_t^N}(x,\cdot)\}_{x\in\mathcal{X},N\in\mathbb{N}}$ is equicontinuous, since for any $\nu_n,\nu' \in \mathcal{M}$ as $d_{\mathcal{M}}(\nu_n,\nu') \to 0$,

$$\max_{x\in\mathcal{X},N\in\mathbb{N}} \left|r_{\pi_t^N}(x,\nu_n) - r_{\pi_t^N}(x,\nu')\right| \leq \max_{x\in\mathcal{X},u\in\mathcal{U}} \left|r(x,u,\nu_n) - r(x,u,\nu')\right| \to 0$$

by continuity of r. The function $r_{\pi_t^N}$ is uniformly bounded for all $N\in\mathbb{N}$ by assumption of uniformly bounded r. By Lemmas A.5.4 and A.5.5,

$$\begin{aligned}
&\lim_{N\to\infty} \left|\mathbb{E}\left[r(x_t^1,u_t^1,\mu[\mathbf{x}_t])\right] - \mathbb{E}\left[r(x_t^\mu,u_t^\mu,\mu_t)\right]\right| \\
&= \lim_{N\to\infty} \left|\mathbb{E}\left[r_{\pi_t^N}(x_t^1,\mu[\mathbf{x}_t])\right] - \mathbb{E}\left[r_{\pi_t^N}(x_t^\mu,\mu_t)\right]\right| = 0\,.
\end{aligned}$$

such that we have

$$\begin{aligned}
&\lim_{N\to\infty} \left|J_1^N(\pi^N,\pi,\dots,\pi) - J^\mu(\pi^N)\right| \\
&\leq \sum_{t\in\mathcal{T}} \lim_{N\to\infty} \left|\mathbb{E}\left[r(x_t^1,u_t^1,\mu[\mathbf{x}_t])\right] - \mathbb{E}\left[r(x_t^\mu,u_t^\mu,\mu_t)\right]\right| = 0\,.
\end{aligned}$$

which is the desired result. ■

From Lemma A.5.6, it follows that for any sequence of optimal exploiting policies $\{\pi^N\}_{N\in\mathbb{N}}$ with $\pi^N \in \Pi$ for all $N \in \mathbb{N}$ and

$$\pi^N \in \arg\max_{\pi\in\Pi} J_1^N(\pi, \pi^*, \ldots, \pi^*)$$

for all $N \in \mathbb{N}$, it holds that for any MFE $(\pi^*, \mu^*) \in \Pi \times \mathcal{M}$,

$$\begin{aligned}\lim_{N\to\infty} J_1^N(\pi^N, \pi^*, \ldots, \pi^*) &\leq \max_{\pi\in\Pi} J^{\mu^*}(\pi) \\ &= J^{\mu^*}(\pi^*) \\ &= \lim_{N\to\infty} J_1^N(\pi^*, \ldots, \pi^*)\end{aligned}$$

and by instantiating for arbitrary $\epsilon > 0$, for sufficiently large N we obtain

$$\begin{aligned}J_1^N(\pi^N, \pi^*, \ldots, \pi^*) - \epsilon &= \max_{\pi\in\Pi} J_1^N(\pi, \pi^*, \ldots, \pi^*) - \epsilon \\ &\leq \max_{\pi\in\Pi} J^{\mu^*}(\pi) - \frac{\epsilon}{2} \\ &= J^{\mu^*}(\pi^*) - \frac{\epsilon}{2} \\ &= J_1^N(\pi^*, \pi^*, \ldots, \pi^*)\end{aligned}$$

which is the desired approximate Nash property that applies to all agents by symmetry. □

A.6 PROOF OF THEOREM 3.1.2

Proof. If Φ or Ψ is constant, or if the restriction Ψ_{Π_Φ} of Ψ to Π_Φ is constant, then $\Gamma = \Psi \circ \Phi$ is constant. Assume that this is not the case.

Then there exist distinct $\pi, \pi' \in \Pi_\Phi$ such that $\Psi(\pi) \neq \Psi(\pi')$. By definition of Π_Φ there also exist distinct $\mu, \mu' \in \mathcal{M}$ such that $\Phi(\mu) = \pi$ and $\Phi(\mu') = \pi'$. Note that for any $\nu, \nu' \in \mathcal{M}$ with $\Gamma(\nu) \neq \Gamma(\nu')$,

$$d_{\mathcal{M}}(\Gamma(\nu), \Gamma(\nu')) \geq \min_{\pi,\pi'\in\Pi_\Phi, \Psi(\pi)\neq\Psi(\pi')} d_{\mathcal{M}}(\Psi(\pi), \Psi(\pi'))$$

where the right-hand side is greater zero by finiteness of Π_Φ. This holds for μ, μ'.

To show that Γ cannot be Lipschitz continuous, assume that Γ has a Lipschitz constant $C > 0$. We can find an integer N such that

$$d_{\mathcal{M}}(\mu^i, \mu^{i+1}) = \frac{d_{\mathcal{M}}(\mu, \mu')}{N-1} < \frac{\min_{\pi,\pi'\in\Pi_\Phi, \Psi(\pi)\neq\Psi(\pi')} d_{\mathcal{M}}(\Psi(\pi), \Psi(\pi'))}{C}$$

for all $i \in \{0, \ldots, N-1\}$ by defining

$$\mu^i = \frac{i}{N}\mu + \frac{N-i}{N}\mu'$$

for all $i \in \{0, \ldots, N\}$, and $\mu^i \in \mathcal{M}$ holds. By the triangle inequality

$$d_{\mathcal{M}}(\Gamma(\mu), \Gamma(\mu')) \leq d_{\mathcal{M}}(\Gamma(\mu^0), \Gamma(\mu^1)) + \ldots + d_{\mathcal{M}}(\Gamma(\mu^{N-1}), \Gamma(\mu^N))$$

there exists a pair (μ^i, μ^{i+1}) with $\Gamma(\mu^i) \neq \Gamma(\mu^{i+1})$. Therefore, for this pair, by the prequel

$$d_{\mathcal{M}}(\Gamma(\mu^i), \Gamma(\mu^{i+1})) \geq \min_{\pi,\pi' \in \Pi_\Phi, \Psi(\pi) \neq \Psi(\pi')} d_{\mathcal{M}}(\Psi(\pi), \Psi(\pi'))\,.$$

On the other hand, since Γ is Lipschitz with constant C, we have

$$d_{\mathcal{M}}(\Gamma(\mu^i), \Gamma(\mu^{i+1})) \leq C \cdot d_{\mathcal{M}}(\mu^i, \mu^{i+1}) < \min_{\pi,\pi' \in \Pi_\Phi, \Psi(\pi) \neq \Psi(\pi')} d_{\mathcal{M}}(\Psi(\pi), \Psi(\pi'))$$

which is a contradiction. Thus, Γ cannot be Lipschitz continuous and by extension cannot be contractive. □

A.7 PROOF OF THEOREM 3.1.3

Proof. For all $\eta > 0, \mu \in \mathcal{M}, t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$, the soft action-value function of the MDP induced by $\mu \in \mathcal{M}$ is given by

$$\begin{aligned}\tilde{Q}_\eta(\mu, t, x, u) &= r(x, u, \mu_t) \\ &+ \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \eta \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right)\end{aligned}$$

and terminal condition $\tilde{Q}_\eta(\mu, T-1, x, u) \equiv r(x, u, \mu_{T-1})$. Analogously, the action-value function of the MDP induced by $\mu \in \mathcal{M}$ is given by

$$Q^*(\mu, t, x, u) = r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \max_{u' \in \mathcal{U}} Q^*(\mu, t+1, x', u')$$

and the similarly defined policy action-value function for $\pi \in \Pi$ is given by

$$Q^\pi(\mu, t, x, u) = r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \sum_{u' \in \mathcal{U}} \pi_{t+1}(u' \mid x') Q^\pi(\mu, t+1, x', u')\,,$$

with terminal conditions $Q^*(\mu, T-1, x, u) \equiv Q^\pi(\mu, T-1, x, u) \equiv r(x, u, \mu_{T-1})$.

We will show that we can find a Lipschitz constant $K_{\tilde{Q}_\eta}$ of $\tilde{Q}_\eta$ that is independent of η if η is not arbitrarily small. To show this, we will explicitly compute such a Lipschitz constant. Note first that $\tilde{Q}_\eta$, Q^* and Q^π are all uniformly bounded by $M_Q \equiv |\mathcal{T}| M_r$ by assumption, where M_r is the uniform bound of r.

Lemma A.7.1. *The functions $\tilde{Q}_\eta(\mu, t, x, u)$, $Q^*(\mu, t, x, u)$ and $Q^\pi(\mu, t, x, u)$ are uniformly bounded for all $\eta > 0, \mu \in \mathcal{M}, t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ by*

$$\left|\tilde{Q}_\eta(\mu, t, x, u)\right| \leq (T-t) M_r \leq T M_r =: M_Q$$

where M_r is the uniform bound of $|r(x, u, \mu_t)| \leq M_r$, and $T = |\mathcal{T}|$.

Proof. Make the induction hypothesis for all $t \in \mathcal{T}$ that

$$\left|\tilde{Q}_\eta(\mu, t, x, u)\right| \leq (T-t) M_r$$

for all $\eta > 0, \mu \in \mathcal{M}, x \in \mathcal{X}, u \in \mathcal{U}$ and note that this holds for $t = T-1$, as by assumption

$$\left|\tilde{Q}_\eta(\mu, T-1, x, u)\right| = |r(x, u, \mu_t)| \leq M_r \,.$$

The induction step from $t+1$ to t holds by

$$\begin{aligned}
&\left|\tilde{Q}_\eta(\mu, t, x, u)\right| \\
&= \left| r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \eta \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right)\right| \\
&\leq |r(x, u, \mu_t)| + \eta \max_{x' \in \mathcal{X}} \left| \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right)\right| \\
&\leq M_r + \eta \left| \log\left(\exp\left(\frac{(T-t-1)M_r}{\eta}\right)\right)\right| \\
&= M_r + (T-t-1)M_r = (T-t)M_r \,.
\end{aligned}$$

By maximizing over all $t \in \mathcal{T}$, we obtain the uniform bound. The other cases are analogous. ■

Now we can find a Lipschitz constant of $\tilde{Q}_\eta(\mu, t, x, u)$ that is independent of η.

Lemma A.7.2. *Let C_r be a Lipschitz constant of $\mu \to r(x, u, \mu_t)$ and C_p a Lipschitz constant of $\mu \to p(x' \mid x, u, \mu_t)$. Further, let $\eta_{\min} > 0$. Then, for all $\eta > \eta_{\min}, t \in \mathcal{T}$, the map $\mu \mapsto \tilde{Q}_\eta(\mu, t, x, u)$ is Lipschitz for all $x \in \mathcal{X}, u \in \mathcal{U}$ with a Lipschitz constant $K^t_{\tilde{Q}_\eta}$ independent of η. Therefore, by picking $K_{\tilde{Q}_\eta} \equiv \max_{t \in \mathcal{T}} K^t_{\tilde{Q}_\eta}$, we have one single Lipschitz constant for all $\eta > \eta_{\min}, t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.*

Proof. We show by induction that for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$, we can find Lipschitz constants such that $\tilde{Q}_\eta(\mu, t, x, u)$ is Lipschitz in μ with a Lipschitz constant that does not depend on η.

To see this, note that this is true for $t = T-1$ and any $x \in \mathcal{X}, u \in \mathcal{U}$, as for any μ, μ' we have

$$\left|\tilde{Q}_\eta(\mu, T-1, x, u) - \tilde{Q}_\eta(\mu', T-1, x, u)\right| = \left|r(x, u, \mu_{T-1}) - r(x, u, \mu'_{T-1})\right| \leq C_r d_{\mathcal{M}}(\mu, \mu') \,.$$

The induction step from $t+1$ to t is

$$\begin{aligned}
&\left|\tilde{Q}_\eta(\mu, t, x, u) - \tilde{Q}_\eta(\mu, t, x, u)\right| \\
&\quad \leq \left|r(x, u, \mu_t) - r(x, u, \mu'_t)\right| \\
&\qquad + \sum_{x' \in \mathcal{X}} \left| p(x' \mid x, u, \mu_t) \eta \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right)\right. \\
&\qquad\quad \left. - p(x' \mid x, u, \mu'_t) \eta \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu', t+1, x', u')}{\eta}\right)\right| \\
&\quad \leq C_r d_{\mathcal{M}}(\mu, \mu') + \eta |\mathcal{X}| \max_{x' \in \mathcal{X}} 1 \cdot \left| \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right)\right. \\
&\qquad\quad \left. - \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu', t+1, x', u')}{\eta}\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&+ \eta|\mathcal{X}| \max_{x' \in \mathcal{X}} \frac{M_Q}{\eta} \cdot \left| p(x' \mid x, u, \mu_t) - p(x' \mid x, u, \mu'_t) \right| \\
&\leq C_r d_{\mathcal{M}}(\mu, \mu') + \eta|\mathcal{X}| \max_{x' \in \mathcal{X}} \sum_{u' \in \mathcal{U}} \left| \frac{\frac{1}{\eta} q_{t+1}(u' \mid x') \exp\left(\frac{\xi_{u'}}{\eta}\right)}{\sum_{u'' \in \mathcal{U}} q_{t+1}(u'' \mid x') \exp\left(\frac{\xi_{u''}}{\eta}\right)} \right| \\
&\quad \cdot \left| \tilde{Q}_\eta(\mu, t+1, x', u') - \tilde{Q}_\eta(\mu', t+1, x', u') \right| + |\mathcal{X}| M_Q \cdot C_p d_{\mathcal{M}}(\mu, \mu') \\
&\leq C_r d_{\mathcal{M}}(\mu, \mu') + \frac{|\mathcal{U}| q_{\max}}{|\mathcal{U}| q_{\min}} \exp\left(2 \cdot \frac{M_Q}{\eta}\right) K_{\tilde{Q}_\eta}^{t+1} d_{\mathcal{M}}(\mu, \mu') + |\mathcal{X}| M_Q C_p d_{\mathcal{M}}(\mu, \mu') \\
&< \left(C_r + \frac{q_{\max}}{q_{\min}} \exp\left(\frac{2M_Q}{\eta_{\min}}\right) K_{\tilde{Q}_\eta}^{t+1} + |\mathcal{X}| M_Q C_p \right) d_{\mathcal{M}}(\mu, \mu')
\end{aligned}
$$

where we use the mean value theorem to obtain some $\xi_u \in [-M_Q, M_Q]$ for all $u \in \mathcal{U}$ bounded by Lemma A.7.1, Lemma A.2.1 for the second inequality, and defined $q_{\max} = \max_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}} q_t(u \mid x)$, $q_{\min} = \min_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}} q_t(u \mid x)$. Since $x \in \mathcal{X}, u \in \mathcal{U}$ were arbitrary, this holds for all $x \in \mathcal{X}, u \in \mathcal{U}$.

Thus, we have the Lipschitz constant $K_{\tilde{Q}_\eta}^t \equiv \left(C_r + \frac{q_{\max}}{q_{\min}} \exp\left(\frac{2M_Q}{\eta_{\min}}\right) K_{\tilde{Q}_\eta}^{t+1} + |\mathcal{X}| M_Q C_p \right)$, as long as $\eta > \eta_{\min}$, since by induction assumption $K_{\tilde{Q}_\eta}^{t+1}$ is independent of η. ■

The optimal action-value function and the policy action-value function for any fixed policy are Lipschitz in μ.

Lemma A.7.3. *The functions $\mu \mapsto Q^*(\mu, t, x, u)$ and $\mu \mapsto Q^\pi(\mu, t, x, u)$ for any fixed $\pi \in \Pi, t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ are Lipschitz continuous. Therefore, for any fixed $\pi \in \Pi$ we can choose a Lipschitz constant K_Q for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ by taking the maximum over all Lipschitz constants.*

Proof. The action-value function is given by the recursion

$$
Q^*(\mu, t, x, u) = r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \max_{u' \in \mathcal{U}} Q^*(\mu, t+1, x', u')
$$

with terminal condition $Q^*(\mu, T-1, x, u) \equiv r(x, u, \mu_{T-1})$. The functions $r(x, u, \mu_t)$ and $p(x' \mid x, u, \mu_t)$ are Lipschitz continuous by Assumption 2. Note that for any $\mu, \mu' \in \mathcal{M}$ and any $t \in \mathcal{T}$, $d_{TV}(\mu_t, \mu'_t) \leq d_M(\mu, \mu')$. Therefore, the terminal condition and all terms in the above recursion are Lipschitz. Further, $Q^*(\mu, t, x, u)$ is uniformly bounded, since r is assumed uniformly bounded.

Since a finite maximum, product and sum of Lipschitz and bounded functions is again Lipschitz and bounded by Lemma A.2.1, we obtain Lipschitz constants $K_{Q,t,x,u}$ of the maps $\mu \to Q^*(\mu, t, x, u)$ for any $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ and define $K_Q \equiv \max_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}} K_{Q,t,x,u}$. The case for Q^π with fixed $\pi \in \Pi$ is analogous. ■

The same holds for $\Psi(\pi)$ mapping from policy π to its induced MF.

Lemma A.7.4. *The function $\Psi(\pi)$ is Lipschitz with some Lipschitz constant K_Ψ.*

Proof. Recall that $\Psi(\pi)$ maps to the MF μ starting with μ_0 and obtained by the recursion

$$\mu_{t+1}(x') = \sum_{x\in\mathcal{X}}\sum_{u\in\mathcal{U}} p(x' \mid x, u, \mu_t)\pi_t(u \mid x)\mu_t(x)\,.$$

We proceed analogously to Lemma A.7.3. μ is uniformly bounded by normalization. The constant function $\pi \mapsto \mu_0(x)$ is Lipschitz and bounded for any $x \in \mathcal{X}$. The functions $r(x, u, \mu_t)$ and $p(x' \mid x, u, \mu_t)$ are Lipschitz continuous by Assumption 2. Since a finite sum, product and composition of Lipschitz and bounded functions is again Lipschitz and bounded by Lemma A.2.1, we obtain Lipschitz constants $K_{\Psi,t,s}$ of the maps $\pi \to \mu_t(x)$ for any $t \in \mathcal{T}, x \in \mathcal{X}$ and define $K_\Psi \equiv \max_{t\in\mathcal{T},x\in\mathcal{X}} K_{\Psi,t,s}$, which is the desired Lipschitz constant of Ψ. ■

Finally, the map from an energy function to its associated Boltzmann distribution is Lipschitz for any $\eta > 0$ with a Lipschitz constant explicitly depending on η.

Lemma A.7.5. *Let $\eta > 0$ arbitrary and $f_u : \mathcal{M} \to \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant K_f for any $u \in \mathcal{U}$. Further, let $g : \mathcal{U} \to \mathbb{R}$ be bounded by $g_{\max} > g(u) > g_{\min} > 0$ for any $u \in \mathcal{U}$. The function*

$$\mu \mapsto \frac{g(u)\exp\left(\frac{f_u(\mu)}{\eta}\right)}{\sum_{u'\in\mathcal{U}} g(u')\exp\left(\frac{f_{u'}(\mu)}{\eta}\right)}$$

is Lipschitz with Lipschitz constant $K = \frac{(|\mathcal{U}|-1)K_f g_{\max}^2}{2\eta g_{\min}^2}$ for any $u \in \mathcal{U}$.

Proof. Let $\mu, \mu' \in \mathcal{M}$ be arbitrary and define

$$\Delta_u f_{u'}(\mu) \equiv f_{u'}(\mu) - f_u(\mu)$$

for any $u' \in \mathcal{U}$, which is Lipschitz with constant $2K_f$. Then, we have

$$\begin{aligned}
&\left|\frac{g(u)\exp\left(\frac{f_u(\mu)}{\eta}\right)}{\sum_{u'\in\mathcal{U}} g(u')\exp\left(\frac{f_{u'}(\mu)}{\eta}\right)} - \frac{g(u)\exp\left(\frac{f_u(\mu')}{\eta}\right)}{\sum_{u'\in\mathcal{U}} g(u')\exp\left(\frac{f_{u'}(\mu')}{\eta}\right)}\right| \\
&= \left|\frac{1}{1+\sum_{u'\neq u}\frac{g(u')}{g(u)}\exp\left(\frac{\Delta_u f_{u'}(\mu)}{\eta}\right)} - \frac{1}{1+\sum_{u'\neq u}\frac{g(u')}{g(u)}\exp\left(\frac{\Delta_u f_{u'}(\mu')}{\eta}\right)}\right| \\
&\leq \left|\sum_{u'\neq u}\frac{\frac{g(u')}{g(u)}\cdot\frac{1}{\eta}\exp\left(\frac{\xi_{u'}}{\eta}\right)}{\left(1+\sum_{u''\neq a}\frac{g(u'')}{g(u)}\exp\left(\frac{\xi_{u''}}{\eta}\right)\right)^2}\cdot\left(\Delta_u f_{u'}(\mu) - \Delta_u f_{u'}(\mu')\right)\right| \\
&\leq \sum_{u'\neq u}\left|\frac{\frac{g_{\max}}{g_{\min}}\cdot\frac{1}{\eta}\exp\left(\frac{\xi_{u'}}{\eta}\right)}{\left(1+\frac{g_{\min}}{g_{\max}}\exp\left(\frac{\xi_{u'}}{\eta}\right)\right)^2}\right|\cdot\left|\Delta_u f_{u'}(\mu) - \Delta_u f_{u'}(\mu')\right| \\
&\leq \frac{g_{\max}^2}{4\eta g_{\min}^2}\cdot\sum_{u'\neq u} 2K_f d_{\mathcal{M}}(\mu,\mu') = \frac{(|\mathcal{U}|-1)K_f g_{\max}^2}{2\eta g_{\min}^2}\cdot d_{\mathcal{M}}(\mu,\mu')
\end{aligned}$$

where we applied the mean value theorem to obtain some $\xi_{u'} \in \mathbb{R}$ for all $u' \in \mathcal{U}$ and used the maximum $\frac{1}{4c}$ of the function $\tilde{f}(x) = \frac{\exp(x/\eta)}{(1+c\cdot\exp(x/\eta))^2}$ at $x = 0$. ■

For RelEnt MFE, by Lemma A.7.2 we obtain a Lipschitz constant $K_{\tilde{Q}_\eta}$ of $\mu \to \tilde{Q}_\eta(\mu, t, x, u)$ as long as $\eta > \eta_{\min}$ for some $\eta_{\min} > 0$. Furthermore, note that for $\tilde{\pi}^{\mu,\eta} \equiv \tilde{\Phi}_\eta(\mu)$, we have

$$\begin{aligned}
&\left|\tilde{\pi}_t^{\mu,\eta}(u \mid x) - \tilde{\pi}_t^{\mu',\eta}(u \mid x))\right| \\
&= \left| \frac{q_t(u \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,x,u')}{\eta}\right)} - \frac{q_t(u \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu',t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu',t,x,u')}{\eta}\right)} \right|.
\end{aligned}$$

We obtain the Lipschitz constant of $\tilde{\Phi}_\eta$ by applying Lemma A.7.5 to each of the maps given by

$$\mu \mapsto \frac{q_t(u \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,x,u')}{\eta}\right)}$$

for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$, resulting in the Lipschitz property

$$\begin{aligned}
&d_\Pi(\tilde{\Phi}_\eta(\mu), \tilde{\Phi}_\eta(\mu')) \\
&\quad = \max_{x \in \mathcal{X}} \max_{t \in \mathcal{T}} \sum_{u \in \mathcal{U}} \left|\tilde{\pi}_t^{\mu,\eta}(u \mid x) - \tilde{\pi}_t^{\mu',\eta}(u \mid x))\right| \\
&\quad \leq \sum_{u \in \mathcal{U}} \frac{(|\mathcal{U}| - 1) K_{\tilde{Q}_\eta} q_{\max}^2}{2\eta q_{\min}^2} \cdot d_{\mathcal{M}}(\mu, \mu') = \frac{|\mathcal{U}|\,(|\mathcal{U}| - 1) K_{\tilde{Q}_\eta} q_{\max}^2}{2\eta q_{\min}^2} \cdot d_{\mathcal{M}}(\mu, \mu'),
\end{aligned}$$

where we let $q_{\max} = \max_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}} q_t(u \mid x)$ and $q_{\min} = \min_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}} q_t(u \mid x)$.

By Lemma A.7.4, $\Psi(\pi)$ is Lipschitz with some Lipschitz constant K_Ψ. Therefore, the resulting Lipschitz constant of the composition $\tilde{\Gamma}_\eta = \Psi \circ \tilde{\Phi}_\eta$ is $\frac{|\mathcal{U}|(|\mathcal{U}|-1) K_{\tilde{Q}_\eta} K_\Psi q_{\max}^2}{2\eta q_{\min}^2}$ and leads to a contraction for any

$$\eta > \max\left(\eta_{\min}, \frac{|\mathcal{U}|\,(|\mathcal{U}| - 1) K_{\tilde{Q}_\eta} K_\Psi q_{\max}^2}{2 q_{\min}^2}\right).$$

Analogously for Boltzmann MFE, by Lemma A.7.3 the mapping $\mu \to Q^*(\mu, t, x, u)$ is Lipschitz with some Lipschitz constant K_{Q^*} for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$. For $\pi^{\mu,\eta} \equiv \Phi_\eta(\mu)$, we have

$$\begin{aligned}
&\left|\pi_t^{\mu,\eta}(u \mid x) - \pi_t^{\mu',\eta}(u \mid x))\right| \\
&= \left| \frac{q_t(u \mid x) \exp\left(\frac{Q^*(\mu,t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{Q^*(\mu,t,x,u')}{\eta}\right)} - \frac{q_t(u \mid x) \exp\left(\frac{Q^*(\mu',t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{Q^*(\mu',t,x,u')}{\eta}\right)} \right|.
\end{aligned}$$

We obtain the Lipschitz constant of Φ_η by applying Lemma A.7.5 to each of the maps given by

$$\mu \mapsto \frac{q_t(u \mid x) \exp\left(\frac{Q^*(\mu,t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{Q^*(\mu,t,x,u')}{\eta}\right)}$$

for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$, resulting in the Lipschitz property

$$\begin{aligned} &d_\Pi(\Phi_\eta(\mu), \Phi_\eta(\mu')) \\ &\quad= \max_{x\in\mathcal{X}} \max_{t\in\mathcal{T}} \sum_{u\in\mathcal{U}} \left|\pi_t^{\mu,\eta}(u \mid x) - \pi_t^{\mu',\eta}(u \mid x))\right| \\ &\quad\leq \sum_{u\in\mathcal{U}} \frac{(|\mathcal{U}|-1)K_{Q^*}q_{\max}^2}{2\eta q_{\min}^2} \cdot d_{\mathcal{M}}(\mu,\mu') = \frac{|\mathcal{U}|\,(|\mathcal{U}|-1)K_{Q^*}q_{\max}^2}{2\eta q_{\min}^2} \cdot d_{\mathcal{M}}(\mu,\mu')\,. \end{aligned}$$

By Lemma A.7.4, $\Psi(\pi)$ is Lipschitz with some Lipschitz constant K_Ψ. The resulting Lipschitz constant of the composition $\Gamma_\eta = \Psi \circ \Phi_\eta$ is $\frac{|\mathcal{U}|(|\mathcal{U}|-1)K_{Q^*}K_\Psi q_{\max}^2}{2\eta q_{\min}^2}$ and leads to a contraction for any

$$\eta > \frac{|\mathcal{U}|\,(|\mathcal{U}|-1)K_{Q^*}K_\Psi q_{\max}^2}{2q_{\min}^2}$$

where for the uniform prior policy, $q_{\max} = q_{\min}$. If required, the Lipschitz constants can be computed recursively according to Lemma A.2.1. □

A.8 PROOF OF THEOREM 3.1.4

Proof. Consider any sequence $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ of η_n-Boltzmann or η_n-RelEnt MFE with $\eta_n \to 0^+$ as $n \to \infty$. Note that a pair (π_n^*, μ_n^*) is completely specified by μ_n^*, since $\pi_n^* = \Phi_{\eta_n}(\mu_n^*)$ or $\pi_n^* = \tilde{\Phi}_{\eta_n}(\mu_n^*)$ uniquely. Therefore, it suffices to show that the associated functions $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu,t,x,u))_{n\in\mathbb{N}}$ and $(\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu,t,x,u))_{n\in\mathbb{N}}$ converge uniformly to $\mu \mapsto Q^*(\mu,t,x,u)$, from which the desired result will follow. For definitions of the different action-value functions, see Appendix A.7.

Note that pointwise convergence is insufficient, since there is no guarantee that μ_n^* itself will converge as $n \to \infty$. However, we can obtain uniform convergence by pointwise convergence and equicontinuity. For RelEnt MFE, we will additionally require uniform convergence of the sequence $(\mu \mapsto \tilde{Q}_{\eta_n}(\mu,t,x,u))_{n\in\mathbb{N}}$ with $\eta_n \to 0^+$. We begin with pointwise convergence of $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu,t,x,u))_{n\in\mathbb{N}}$ to the optimal action-value function $\mu \mapsto Q^*(\mu,t,x,u)$.

Lemma A.8.1. *Any sequence of functions $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu,t,x,u))_{n\in\mathbb{N}}$ with $\eta_n \to 0^+$ converges pointwise to $\mu \mapsto Q^*(\mu,t,x,u)$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.*

Proof. Fix $\mu \in \mathcal{M}$. We make the induction hypothesis for arbitrary $t \in \mathcal{T}$ that for all $x \in \mathcal{X}, u \in \mathcal{U}, \varepsilon > 0$, there exists $n' \in \mathbb{N}$ such that for any $n > n'$ we have

$$\left|Q^{\Phi_{\eta_n}(\mu)}(\mu,t,x,u) - Q^*(\mu,t,x,u)\right| < \varepsilon\,.$$

The induction hypothesis is fulfilled for $t = T-1$, as by definition

$$\left|Q^{\Phi_{\eta_n}(\mu)}(\mu,t,x,u) - Q^*(\mu,t,x,u)\right| = |r(x,u,\mu_t) - r(x,u,\mu_t)| = 0\,.$$

Assume that the induction hypothesis is fulfilled for $t+1$, then at time t let $x \in \mathcal{X}, u \in \mathcal{U}, \varepsilon > 0$ arbitrary. Furthermore, let $x' \in \mathcal{X}$ arbitrary. Collect all optimal actions into a set $\mathcal{U}_{\text{opt}}^{x'} \subseteq \mathcal{U}$, i.e. for $u' \in \mathcal{U}_{\text{opt}}^{x'}$ we have

$$Q^*(\mu,t,x',u_{\text{opt}}) = \max_{u\in\mathcal{U}} Q^*(\mu,t,x',u)\,.$$

We define the minimal action gap

$$\Delta Q_{\min}^{x',\mu} \equiv \min_{u_{\text{opt}} \in \mathcal{U}_{\text{opt}}^{x'}, u_{\text{sub}} \in \mathcal{U} \setminus \mathcal{U}_{\text{opt}}^{x'}} \left(Q^*(\mu, t, x', u_{\text{opt}}) - Q^*(\mu, t, x', u_{\text{sub}}) \right) > 0$$

such that for arbitrary suboptimal actions $u_{\text{sub}} \in \mathcal{U} \setminus \mathcal{U}_{\text{opt}}^{x'}$ and optimal actions $u_{\text{opt}} \in \mathcal{U}_{\text{opt}}^{x'}$,

$$Q^*(\mu, t, x', u_{\text{opt}}) - Q^*(\mu, t, x', u_{\text{sub}}) \geq \Delta Q_{\min}^{x',\mu}.$$

This is well defined if there are suboptimal actions, since there is always at least one optimal action. If all actions are optimal, we can skip bounding the probability of taking suboptimal actions and the result will hold trivially. Thus, we assume henceforth that there exists a suboptimal action.

It follows that the probability of taking suboptimal actions $u_{\text{sub}} \in \mathcal{U} \setminus \mathcal{U}_{\text{opt}}^{x'}$ disappears, since

$$\begin{aligned}
(\Phi_{\eta_n}(\mu))_t(u_{\text{sub}} \mid x') &= \frac{q_t(u_{\text{sub}} \mid x)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{Q^*(\mu,t,x,u') - Q^*(\mu,t,x,u_{\text{sub}})}{\eta}\right)} \\
&\leq \frac{1}{1 + \sum_{u' \in \mathcal{U}} \frac{q_t(u'|x)}{q_t(u_{\text{sub}}|x)} \exp\left(\frac{Q^*(\mu,t,x,u') - Q^*(\mu,t,x,u_{\text{sub}})}{\eta}\right)} \\
&\leq \frac{1 \mid x)}{1 + \frac{q_t(u_{\text{opt}}|x)}{q_t(u_{\text{sub}}|x)} \exp\left(\frac{Q^*(\mu,t,x,u_{\text{opt}}) - Q^*(\mu,t,x,u_{\text{sub}})}{\eta}\right)} \\
&\leq \frac{1 \mid x)}{1 + \frac{q_t(u_{\text{opt}}|x)}{q_t(u_{\text{sub}}|x)} \exp\left(\frac{\Delta Q_{\min}^{x',\mu}}{\eta}\right)} \to 0
\end{aligned}$$

as $\eta \to 0^+$ for some arbitrary optimal action $u_{\text{opt}} \in \mathcal{U}_{\text{opt}}^{x'}$. Since $x' \in \mathcal{X}$ was arbitrary, this holds for all $x' \in \mathcal{X}$. Therefore, by finiteness of $\mathcal{X}$ and $\mathcal{U}$ we can choose $n_1 \in \mathbb{N}$ such that for all $n > n_1$ and for all $u_{\text{sub}} \in \mathcal{U} \setminus \mathcal{U}_{\text{opt}}^{x'}$ we have η_n sufficiently small such that

$$(\Phi_{\eta_n}(\mu))_t(u_{\text{sub}} \mid x') < \frac{\varepsilon}{2|\mathcal{U}|M_Q}$$

where M_Q is the uniform bound of $Q^{\Phi_{\eta_n}(\mu)}$.

Further, by induction assumption, we can choose $n_{x',u'}$ for any $x' \in \mathcal{X}, u' \in \mathcal{U}$ such that for all $n > n_{x',u'}$ we have

$$\left| Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, x', u') - Q^*(\mu, t+1, x', u') \right| < \frac{\varepsilon}{3}$$

Therefore, as long as $n > n' \equiv \max(n_1, \max_{x' \in \mathcal{X}, u' \in \mathcal{U}} n_{x',u'})$, we have

$$\begin{aligned}
&\left| Q^{\Phi_{\eta_n}(\mu)}(\mu, t, x, u) - Q^*(\mu, t, x, u) \right| \\
&= \left| \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \left(\sum_{u' \in \mathcal{U}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, x', u') - \max_{u'' \in \mathcal{U}} Q^*(\mu, t+1, x', u'') \right) \right| \\
&\leq \max_{x' \in \mathcal{X}} \left| \sum_{u' \in \mathcal{U}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, x', u') - \max_{u'' \in \mathcal{U}} Q^*(\mu, t+1, x', u'') \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{x'\in\mathcal{X}} \left| \sum_{u'\in\mathcal{U}^{x'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, x', u') - \max_{u''\in\mathcal{U}} Q^*(\mu, t+1, x', u'') \right| \\
&\quad + \max_{x'\in\mathcal{X}} \left| \sum_{u'\in\mathcal{U}\setminus\mathcal{U}^{x'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, x', u') \right| \\
&\leq \max_{x'\in\mathcal{X}} \left| \sum_{u'\in\mathcal{U}^{x'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, x', u') \right. \\
&\qquad\qquad \left. - \sum_{u'\in\mathcal{U}^{x'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') \max_{u''\in\mathcal{U}} Q^*(\mu, t+1, x', u'') \right| \\
&\quad + \max_{x'\in\mathcal{X}} \left| \sum_{u'\in\mathcal{U}^{x'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') \max_{u''\in\mathcal{U}} Q^*(\mu, t+1, x', u'') - \max_{u''\in\mathcal{U}} Q^*(\mu, t+1, x', u'') \right| \\
&\quad + \max_{x'\in\mathcal{X}} \left| \sum_{u'\in\mathcal{U}\setminus\mathcal{U}^{x'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, x', u') \right| \\
&\leq \max_{x'\in\mathcal{X}} \max_{u'\in\mathcal{U}^{x'}_{\mathrm{opt}}} \left| Q^{\Phi_{\eta_n}(\mu)}(\mu, t+1, x', u') - \max_{u''\in\mathcal{U}} Q^*(\mu, t+1, x', u'') \right| \\
&\quad + \max_{x'\in\mathcal{X}} M_Q \left| - \sum_{u'\in\mathcal{U}\setminus\mathcal{U}^{x'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') \right| + \max_{x'\in\mathcal{X}} M_Q \left| \sum_{u'\in\mathcal{U}\setminus\mathcal{U}^{x'}_{\mathrm{opt}}} (\Phi_{\eta_n}(\mu))_t(u' \mid x') \right| \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3|\mathcal{U}|M_Q} \cdot |\mathcal{U}| M_Q + \frac{\varepsilon}{3|\mathcal{U}|M_Q} \cdot |\mathcal{U}| M_Q = \varepsilon \,.
\end{aligned}
$$

Since $x \in \mathcal{X}, u \in \mathcal{U}, \varepsilon > 0$ were arbitrary, the desired result follows immediately by induction. ■

As we have no control over μ_n^* and the sequence $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ may not even converge, pointwise convergence is insufficient. To obtain uniform convergence, we shall use compactness of $\mathcal{M}$ and equicontinuity.

Lemma A.8.2. *The family of functions $\mathcal{F} \equiv \{\mu \mapsto Q^{\Phi_\eta(\mu)}(\mu, t, x, u)\}_{\eta>0, t\in\mathcal{T}, x\in\mathcal{X}, u\in\mathcal{U}}$ is equicontinuous, i.e. for any $\varepsilon > 0$ and any $\mu \in \mathcal{M}$, we can choose a $\delta > 0$ such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta$ and any $f \in \mathcal{F}$ we have*

$$\big|f(\mu) - f(\mu')\big| < \varepsilon \,.$$

Proof. Fix an arbitrary $\mu \in \mathcal{M}$. We make the (backwards in time) induction hypothesis for all $t \in \mathcal{T}$ that for any $x \in \mathcal{X}, u \in \mathcal{U}, \varepsilon_{t,x,u} > 0$, there exists $\delta_{t,x,u} > 0$ such that for any $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta_{t,x,u}$ and any $f \in \mathcal{F}$ we have

$$\left|Q^{\Phi_\eta(\mu)}(\mu, t, x, u) - Q^{\Phi_\eta(\mu')}(\mu', t, x, u)\right| < \varepsilon_{t,x,u} \,.$$

The induction hypothesis is fulfilled for $t = T-1$, as by assumption, $\nu \to r(x,u,\nu_t)$ is Lipschitz with constant $C_r > 0$. Therefore, for all $x \in \mathcal{X}, u \in \mathcal{U}$ we can choose $\delta_{T-1,x,u} = \frac{\varepsilon_{t,x,u}}{C_r}$ such that for any μ, μ' with $d_{\mathcal{M}}(\mu,\mu') < \delta'$ we have

$$\left|Q^{\Phi_\eta(\mu)}(\mu,t,x,u) - Q^{\Phi_\eta(\mu')}(\mu',t,x,u)\right| = \left|r(x,u,\mu_t) - r(x,u,\mu_t')\right| \leq C_r d_{\mathcal{M}}(\mu,\mu') < \varepsilon_{t,x,u}\,.$$

Assume that the induction hypothesis holds for $t+1$, then at time t let $\varepsilon_{t,x,u} > 0, x \in \mathcal{X}, u \in \mathcal{U}$ arbitrary. By definition, we have

$$\begin{aligned}
&\left|Q^{\Phi_\eta(\mu)}(\mu,t,x,u) - Q^{\Phi_\eta(\mu')}(\mu',t,x,u)\right| \\
&= \Bigg| r(x,u,\mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x,u,\mu_t) \sum_{u' \in \mathcal{U}} (\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u') \\
&\quad - r(x,u,\mu_t') - \sum_{x' \in \mathcal{X}} p(x' \mid x,u,\mu_t') \sum_{u' \in \mathcal{U}} (\Phi_\eta(\mu'))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u') \Bigg| \\
&\leq \left|r(x,u,\mu_t) - r(x,u,\mu_t')\right| \\
&\quad + \sum_{x' \in \mathcal{X}} \left| \left(p(x' \mid x,u,\mu_t) - p(x' \mid x,u,\mu_t')\right) \sum_{u' \in \mathcal{U}} (\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u') \right| \\
&\quad + \sum_{x' \in \mathcal{X}} \Bigg| p(x' \mid x,u,\mu_t') \sum_{u' \in \mathcal{U}} \Big((\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u') \\
&\qquad\qquad - (\Phi_\eta(\mu'))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u') \Big) \Bigg| \\
&\leq \left|r(x,u,\mu_t) - r(x,u,\mu_t')\right| \\
&\quad + \sum_{x' \in \mathcal{X}} \left| \left(p(x' \mid x,u,\mu_t) - p(x' \mid x,u,\mu_t')\right) \sum_{u' \in \mathcal{U}} (\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u') \right| \\
&\quad + \max_{x' \in \mathcal{X}} \Bigg| \sum_{u' \in \mathcal{U}^{x'}_{\mathrm{opt}}} \Big((\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u') \\
&\qquad\qquad - (\Phi_\eta(\mu'))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u') \Big) \Bigg| \\
&\quad + \max_{x' \in \mathcal{X}} \Bigg| \sum_{u' \in \mathcal{U} \setminus \mathcal{U}^{x'}_{\mathrm{opt}}} \Big((\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u') \\
&\qquad\qquad - (\Phi_\eta(\mu'))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u') \Big) \Bigg|
\end{aligned}$$

where we define $\mathcal{U}^{x'}_{\mathrm{opt}} \subseteq \mathcal{U}$ for any $x' \in \mathcal{X}$ to include all optimal actions $u_{\mathrm{opt}} \in \mathcal{U}^{x'}_{\mathrm{opt}}$ such that

$$Q^*(\mu,t,x',u_{\mathrm{opt}}) = \max_{u \in \mathcal{U}} Q^*(\mu,t,x',u)\,.$$

We bound each of the four terms separately.

For the first term, we choose $\delta^1_{t,x,u} = \frac{\varepsilon_{t,x,u}}{4C_r}$ by Lipschitz continuity such that

$$\left|r(x,u,\mu_t) - r(x,u,\mu_t')\right| < \frac{\varepsilon_{t,x,u}}{4}$$

for all μ' with $d_{\mathcal{M}}(\mu, \mu') < \delta^1_{t,x,u}$.

For the second term, we choose $\delta^2_{t,x,u} = \frac{1}{4|\mathcal{X}|M_Q C_p}$ such that for any $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta^2_{t,x,u}$ we have

$$\begin{aligned}
&\sum_{x' \in \mathcal{X}} \left| \big(p(x' \mid x, u, \mu_t) - p(x' \mid x, u, \mu'_t)\big) \sum_{u' \in \mathcal{U}} (\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') \right| \\
&\leq |\mathcal{X}| C_p d_{\mathcal{M}}(\mu, \mu') M_Q < \frac{\varepsilon_{t,x,u}}{4}
\end{aligned}$$

where M_Q denotes the uniform bound of Q and C_p is the Lipschitz constant of $\nu \mapsto p(x' \mid x, u, \nu_t)$.

For the third and fourth term, we first fix $x' \in \mathcal{X}$ and define the minimal action gap as

$$\Delta Q^{x',\mu}_{\min} \equiv \min_{u_{\text{opt}} \in \mathcal{U}^{x'}_{\text{opt}}, u_{\text{sub}} \in \mathcal{U} \setminus \mathcal{U}^{x'}_{\text{opt}}} \big(Q^*(\mu, t, x', u_{\text{opt}}) - Q^*(\mu, t, x', u_{\text{sub}})\big) \,.$$

This is well defined if there are suboptimal actions, since there is always at least one optimal action. If all actions are optimal, we can skip bounding the probability of taking suboptimal actions and the result will still hold. Henceforth, we assume that there exists a suboptimal action.

By Lipschitz continuity of $\mu \mapsto Q^*(\mu, t, x, u)$ from Lemma A.7.3 implying uniform continuity, there exists some $\delta^{3,x'}_{t,x,u} > 0$ such that

$$\big|Q^*(\mu', t, x', u) - Q^*(\mu, t, x', u)\big| < \frac{\Delta Q^{x',\mu}_{\min}}{4}$$

for all $\mu' \in \mathcal{M}, u \in \mathcal{U}$ where $d_{\mathcal{M}}(\mu, \mu') < \delta^{3,x'}_{t,x,u}$, and thus

$$\Delta Q^{x',\mu'}_{\min} = \min_{u_{\text{opt}} \in \mathcal{U}^{x'}_{\text{opt}}, u_{\text{sub}} \in \mathcal{U} \setminus \mathcal{U}^{x'}_{\text{opt}}} \big(Q^*(\mu', t, x', u_{\text{opt}}) - Q^*(\mu', t, x', u_{\text{sub}})\big) > \frac{\Delta Q^{x',\mu}_{\min}}{2} \,.$$

Under this condition, we can now show that the probability of any suboptimal action can be controlled. Define $R^{\min}_q \equiv \min_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}, u' \in \mathcal{U}} \frac{q_t(u'|x)}{q_t(u|x)} > 0$ and $R^{\max}_q \equiv \max_{t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}, u' \in \mathcal{U}} \frac{q_t(u'|x)}{q_t(u|x)} > 0$. Let $u_{\text{sub}} \in \mathcal{U} \setminus \mathcal{U}^{x'}_{\text{opt}}$, then we either have

$$\begin{aligned}
&\big|(\Phi_\eta(\mu))_{t+1}(u_{\text{sub}} \mid x') - (\Phi_\eta(\mu'))_{t+1}(u_{\text{sub}} \mid x')\big| \\
&= \left| \frac{1}{1 + \sum_{u' \neq u_{\text{sub}}} \frac{q_t(u'|x')}{q_t(u_{\text{sub}}|x')} \exp\left(\frac{Q^*(\mu,t,x',u') - Q^*(\mu,t,x',u_{\text{sub}})}{\eta}\right)} \right. \\
&\quad \left. - \frac{1}{1 + \sum_{u' \neq u_{\text{sub}}} \frac{q_t(u'|x')}{q_t(u_{\text{sub}}|x')} \exp\left(\frac{Q^*(\mu',t,x',u') - Q^*(\mu',t,x',u_{\text{sub}})}{\eta}\right)} \right| \\
&\leq \frac{1}{1 + \max_{u' \neq u_{\text{sub}}} R^{\min}_q \exp\left(\frac{Q^*(\mu,t,x',u') - Q^*(\mu,t,x',u_{\text{sub}})}{\eta}\right)} \\
&\quad + \frac{1}{1 + \max_{u' \neq u_{\text{sub}}} R^{\min}_q \exp\left(\frac{Q^*(\mu',t,x',u') - Q^*(\mu',t,x',u_{\text{sub}})}{\eta}\right)} \\
&< \frac{1}{1 + R^{\min}_q \exp\left(\frac{\Delta Q^{x',\mu}_{\min}}{\eta}\right)} + \frac{1}{1 + R^{\min}_q \exp\left(\frac{\Delta Q^{x',\mu}_{\min}}{2\eta}\right)}
\end{aligned}$$

$$\leq \frac{2}{1 + R_q^{\min} \exp\left(\frac{\Delta Q_{\min}^{x',\mu}}{2\eta}\right)} < \frac{\varepsilon_{t,x,u}}{8M_Q|\mathcal{U}|}$$

if $\varepsilon_{t,x,u} > 16M_Q|\mathcal{U}|$ trivially, or otherwise if $\eta < \eta_{\min}^{x'}$ with

$$\eta_{\min}^{x'} \equiv \frac{\Delta Q_{\min}^{x',\mu}}{2\log\left(\frac{16M_Q|\mathcal{U}|}{\varepsilon_{t,x,u}R_q^{\min}} - \frac{1}{R_q^{\min}}\right)},$$

in which case we arbitrarily define $\delta_{t,x,u}^{4,x'} = 1$, or if neither apply, then $\eta \geq \eta_{\min}^{x'}$ and thus

$$\begin{aligned}
&|(\Phi_\eta(\mu))_{t+1}(u_{\text{sub}} \mid x') - (\Phi_\eta(\mu'))_{t+1}(u_{\text{sub}} \mid x')| \\
&= \left| \frac{1}{1 + \sum_{u' \neq u_{\text{sub}}} \frac{q_t(u'|x')}{q_t(u_{\text{sub}}|x')} \exp\left(\frac{Q^*(\mu,t,x',u') - Q^*(\mu,t,x',u_{\text{sub}})}{\eta}\right)} \right. \\
&\quad \left. - \frac{1}{1 + \sum_{u' \neq u_{\text{sub}}} \frac{q_t(u'|x')}{q_t(u_{\text{sub}}|x')} \exp\left(\frac{Q^*(\mu',t,x',u') - Q^*(\mu',t,x',u_{\text{sub}})}{\eta}\right)} \right| \\
&= \left| \frac{\sum_{u' \neq u_{\text{sub}}} \frac{q_t(u'|x)}{q_t(u_{\text{sub}}|x')} \left(\exp\left(\frac{Q^*(\mu',t,x',u') - Q^*(\mu',t,x',u_{\text{sub}})}{\eta}\right) - \exp\left(\frac{Q^*(\mu,t,x',u') - Q^*(\mu,t,x',u_{\text{sub}})}{\eta}\right)\right)}{(1 + \cdots) \cdot (1 + \cdots)} \right| \\
&\leq R_q^{\max} \sum_{u' \neq u_{\text{sub}}} \left| \exp\left(\frac{Q^*(\mu',t,x',u') - Q^*(\mu',t,x',u_{\text{sub}})}{\eta}\right) \right. \\
&\qquad \left. - \exp\left(\frac{Q^*(\mu,t,x',u') - Q^*(\mu,t,x',u_{\text{sub}})}{\eta}\right) \right| \\
&\leq R_q^{\max} \sum_{u' \neq u_{\text{sub}}} \left| \frac{1}{\eta} \exp\left(\frac{\xi_{u'}}{\eta}\right) \right| \\
&\qquad \cdot |(Q^*(\mu',t,x',u') - Q^*(\mu',t,x',u_{\text{sub}})) - (Q^*(\mu,t,x',u') - Q^*(\mu,t,x',u_{\text{sub}}))| \\
&\leq R_q^{\max}|\mathcal{U}| \cdot \frac{1}{\eta_{\min}^{x'}} \exp\left(\frac{2M_Q}{\eta_{\min}^{x'}}\right) \cdot 2K_Q d_{\mathcal{M}}(\mu,\mu') < \frac{\varepsilon_{t,x,u}}{8M_Q|\mathcal{U}|}
\end{aligned}$$

by the mean value theorem with some $\xi_{u'} \in [-2M_Q, 2M_Q]$ for all $u' \in \mathcal{U}$, where we abbreviated the denominator $(1 + \cdots) \cdot (1 + \cdots) \geq 1$, as long as we choose

$$\delta_{t,x,u}^{4,x'} = \frac{\varepsilon_{t,x,u}\eta_{\min}^{x'}}{8M_Q|\mathcal{U}|^2 R_q^{\max} \cdot \exp\left(\frac{2M_Q}{\eta_{\min}^{x'}}\right) \cdot 2K_Q}$$

and $d_{\mathcal{M}}(\mu,\mu') < \delta_{t,x,u}^{4,x'}$, where K_Q is the Lipschitz constant of $\mu \mapsto Q^*(\mu,t,x,u)$ given by Lemma A.7.3.

Since $x' \in \mathcal{X}$ was arbitrary, we now define $\delta_{t,x,u}^3 \equiv \min_{x' \in \mathcal{X}} \delta_{t,x,u}^{3,x'}$, $\delta_{t,x,u}^4 \equiv \min_{x' \in \mathcal{X}} \delta_{t,x,u}^{4,x'}$ and let $d_{\mathcal{M}}(\mu,\mu') < \min(\delta_{t,x,u}^3, \delta_{t,x,u}^4)$. Under these assumptions, for the third term we have approximate optimality for all optimal actions in $\mathcal{U}_{\text{opt}}^{x'}$, since by induction assumption we can choose $\delta_{t+1,x',u'}$ for all $x' \in \mathcal{X}, u' \in \mathcal{U}$ such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu,\mu') < \delta_{t+1,x',u'}$ it holds that

$$\left| Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u') - Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u') \right| < \frac{\varepsilon_{t,x,u}}{16|\mathcal{U}| + 8}.$$

and therefore for all $\mu' \in \mathcal{M}$, as long as $d_{\mathcal{M}}(\mu, \mu') < \min_{x' \in \mathcal{X}, u' \in \mathcal{U}} \delta_{t+1,x',u'}$, we have

$$
\begin{aligned}
&\max_{x' \in \mathcal{X}} \Bigg| \sum_{u' \in \mathcal{U}_{\mathrm{opt}}^{x'}} (\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') \\
&\qquad\qquad - \sum_{u' \in \mathcal{U}_{\mathrm{opt}}^{x'}} (\Phi_\eta(\mu'))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu')}(\mu', t+1, x', u') \Bigg| \\
&\leq \max_{x' \in \mathcal{X}} \Bigg| \sum_{u' \in \mathcal{U}_{\mathrm{opt}}^{x'}} (\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') \\
&\qquad\qquad - \sum_{u' \in \mathcal{U}_{\mathrm{opt}}^{x'}} (\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu')}(\mu', t+1, x', u') \Bigg| \\
&\quad + \max_{x' \in \mathcal{X}} \Bigg| \sum_{u' \in \mathcal{U}_{\mathrm{opt}}^{x'}} (\Phi_\eta(\mu))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu')}(\mu', t+1, x', u') \\
&\qquad\qquad - \sum_{u' \in \mathcal{U}_{\mathrm{opt}}^{x'}} (\Phi_\eta(\mu'))_{t+1}(u' \mid x') Q^{\Phi_\eta(\mu')}(\mu', t+1, x', u') \Bigg| \\
&\leq \max_{x' \in \mathcal{X}} \max_{u' \in \mathcal{U}} \left| Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') - Q^{\Phi_\eta(\mu')}(\mu', t+1, x', u') \right| \\
&\quad + \max_{x' \in \mathcal{X}} \Bigg| \sum_{u' \in \mathcal{U}_{\mathrm{opt}}^{x'}} \left((\Phi_\eta(\mu))_{t+1}(u' \mid x') - (\Phi_\eta(\mu'))_{t+1}(u' \mid x') \right) \\
&\qquad\qquad \cdot \left(Q^{\Phi_\eta(\mu')}(\mu', t+1, x', u') - Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') \right) \Bigg| \\
&\quad + \max_{x' \in \mathcal{X}} \left| \sum_{u' \in \mathcal{U}_{\mathrm{opt}}^{x'}} \left((\Phi_\eta(\mu))_{t+1}(u' \mid x') - (\Phi_\eta(\mu'))_{t+1}(u' \mid x') \right) Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') \right| \\
&\leq \max_{x' \in \mathcal{X}} \max_{u' \in \mathcal{U}} \left| Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') - Q^{\Phi_\eta(\mu')}(\mu', t+1, x', u') \right| \\
&\quad + \max_{x' \in \mathcal{X}} \max_{u' \in \mathcal{U}} 2|\mathcal{U}| \left| Q^{\Phi_\eta(\mu')}(\mu', t+1, x', u') - Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') \right| \\
&\quad + \max_{x' \in \mathcal{X}} \max_{u'' \in \mathcal{U}} \left| Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u'') \right| \\
&\qquad\qquad \cdot \left| \sum_{u' \in \mathcal{U} \setminus \mathcal{U}_{\mathrm{opt}}^{x'}} \left((\Phi_\eta(\mu'))_{t+1}(u' \mid x') - (\Phi_\eta(\mu))_{t+1}(u' \mid x') \right) \right| \\
&< (1 + 2|\mathcal{U}|) \cdot \frac{\varepsilon_{t,x,u}}{16|\mathcal{U}| + 8} + M_Q |\mathcal{U}| \cdot \frac{\varepsilon_{t,x,u}}{8 M_Q |\mathcal{U}|} < \frac{\varepsilon_{t,x,u}}{4}
\end{aligned}
$$

where we use that for any $u' \in \mathcal{U}_{\mathrm{opt}}^{x'}$ we have

$$
Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u') = \max_{u'' \in \mathcal{U}} Q^{\Phi_\eta(\mu)}(\mu, t+1, x', u'') \,.
$$

Analogously, for the fourth term we have

$$\begin{aligned}
&\max_{x'\in\mathcal{X}}\Bigg|\sum_{u'\in\mathcal{U}\setminus\mathcal{U}_{\text{opt}}^{x'}}((\Phi_\eta(\mu))_{t+1}(u'\mid x')Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u')\\
&\qquad\qquad -(\Phi_\eta(\mu'))_{t+1}(u'\mid x')Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u'))\Bigg|\\
&\leq \max_{x'\in\mathcal{X}}\sum_{u'\in\mathcal{U}\setminus\mathcal{U}_{\text{opt}}^{x'}}\Big|(\Phi_\eta(\mu))_{t+1}(u'\mid x')Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u')\\
&\qquad\qquad -(\Phi_\eta(\mu))_{t+1}(u'\mid x')Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u')\Big|\\
&\quad+\max_{x'\in\mathcal{X}}\sum_{u'\in\mathcal{U}\setminus\mathcal{U}_{\text{opt}}^{x'}}\Big|(\Phi_\eta(\mu))_{t+1}(u'\mid x')Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u')\\
&\qquad\qquad -(\Phi_\eta(\mu'))_{t+1}(u'\mid x')Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u')\Big|\\
&\leq \max_{x'\in\mathcal{X}}\max_{u'\in\mathcal{U}}\Big|Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u')-Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u')\Big|\\
&\quad+\max_{x'\in\mathcal{X}}M_Q\sum_{u'\in\mathcal{U}\setminus\mathcal{U}_{\text{opt}}^{x'}}\big|(\Phi_\eta(\mu))_{t+1}(u'\mid x')-(\Phi_\eta(\mu'))_{t+1}(u'\mid x')\big|\\
&<\frac{\varepsilon_{t,x,u}}{8}+M_Q|\mathcal{U}|\cdot\frac{\varepsilon_{t,x,u}}{8M_Q|\mathcal{U}|}=\frac{\varepsilon_{t,x,u}}{4}
\end{aligned}$$

under the previous conditions, since as long as we have $d_{\mathcal{M}}(\mu,\mu')<\delta_{t+1,x',u'}$ for all $x'\in\mathcal{X},u'\in\mathcal{U}$ from before, we have

$$\Big|Q^{\Phi_\eta(\mu)}(\mu,t+1,x',u')-Q^{\Phi_\eta(\mu')}(\mu',t+1,x',u')\Big|<\frac{\varepsilon_{t,x,u}}{16|\mathcal{U}|+8}<\frac{\varepsilon_{t,x,u}}{8}\,.$$

Finally, by choosing $\delta_{t,x,u}$ such that all conditions are fulfilled, i.e.

$$\delta_{t,x,u}\equiv\min\left(\delta^1_{t,x,u},\delta^2_{t,x,u},\delta^3_{t,x,u},\delta^4_{t,x,u},\min_{x'\in\mathcal{X},u'\in\mathcal{U}}\delta_{t+1,x',u'}\right)>0\,,$$

the induction hypothesis is fulfilled, since then for any μ' with $d_{\mathcal{M}}(\mu,\mu')<\delta_{t,x,u}$ we have

$$\Big|Q^{\Phi_\eta(\mu)}(\mu,t,x,u)-Q^{\Phi_\eta(\mu')}(\mu',t,x,u)\Big|<\varepsilon_{t,x,u}\,.$$

Since $\eta>0$ is arbitrary, the desired result follows immediately, as we can set $\varepsilon_{t,x,u}=\varepsilon$ for each $t\in\mathcal{T},x\in\mathcal{X},u\in\mathcal{U}$ and obtain $\delta\equiv\max_{t\in\mathcal{T},x\in\mathcal{X},u\in\mathcal{U}}\delta_{t,x,u}$, fulfilling the required equicontinuity property at μ. ■

From equicontinuity, we get the desired uniform convergence via compactness.

Lemma A.8.3. *If $(f_n)_{n\in\mathbb{N}}$ with $f_n:\mathcal{M}\to\mathbb{R}$ is an equicontinuous sequence of functions and for all $\mu\in\mathcal{M}$ we have $f_n(\mu)\to f(\mu)$ pointwise, then $f_n(\mu)\to f(\mu)$ uniformly.*

Proof. Let $\varepsilon>0$ arbitrary, then there exists by equicontinuity for any point $\mu\in\mathcal{M}$ a $\delta(\mu)$ such that for all $\mu'\in\mathcal{M}$ with $d_{\mathcal{M}}(\mu,\mu')<\delta(\mu)$ we have for all $n\in\mathbb{N}$

$$\big|f_n(\mu)-f_n(\mu')\big|<\frac{\varepsilon}{3}$$

which via pointwise convergence implies

$$\left|f(\mu) - f(\mu')\right| \leq \frac{\varepsilon}{3}\,.$$

Since $\mathcal{M}$ is compact, it is separable, i.e. there exists a countable dense subset $(\mu_j)_{j\in\mathbb{N}}$ of $\mathcal{M}$. Let $\delta(\mu)$ be as defined above and cover $\mathcal{M}$ by the open balls $(B_{\delta(\mu_j)}(\mu_j))_{j\in\mathbb{N}}$. By the compactness of $\mathcal{M}$, finitely many of these balls $B_{\delta(\mu_{n_1})}(\mu_{n_1}), \ldots, B_{\delta(\mu_{n_k})}(\mu_{n_k})$ cover $\mathcal{M}$. By pointwise convergence, for any $i = 1, \ldots, k$ we can find an integer n_i such that for all $n > n_i$ we have

$$|f_n(\mu_{n_i}) - f(\mu_{n_i})| < \frac{\varepsilon}{3}\,.$$

Taken together, we find that for $n > \max_{i=1,\ldots,k} n_i$ and arbitrary $\mu \in \mathcal{M}$, we have

$$\begin{aligned}|f_n(\mu) - f(\mu)| &\leq |f_n(\mu) - f_n(\mu_{n_i})| + |f_n(\mu_{n_i}) - f(\mu_{n_i})| + |f(\mu_{n_i}) - f(\mu)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon\end{aligned}$$

for some center point μ_{n_i} of a ball containing μ from the finite cover. ■

Therefore, a sequence of Boltzmann MFE with vanishing η is approximately optimal in the MFG.

Lemma A.8.4. *For any sequence $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ of η_n-Boltzmann MFE with $\eta_n \to 0^+$ and for any $\varepsilon > 0$ there exists integer $N \in \mathbb{N}$ such that for all integers $n > N$ we have*

$$J^{\mu_n^*}(\pi_n^*) \geq \max_{\pi} J^{\mu_n^*}(\pi) - \varepsilon\,.$$

Proof. By Lemma A.8.2, $\mathcal{F} \equiv (\mu \mapsto Q^{\Phi_\eta(\mu)}(\mu, t, x, u))_{\eta>0, t\in\mathcal{T}, x\in\mathcal{X}, u\in\mathcal{U}}$ is equicontinuous. Therefore, any sequence $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu, t, x, u))_{n\in\mathbb{N}}$ with $\eta_n \to 0^+$ is also equicontinuous for any $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.

Furthermore, by Lemma A.8.1, the sequence $(\mu \mapsto Q^{\Phi_{\eta_n}(\mu)}(\mu, t, x, u))_{n\in\mathbb{N}}$ converges pointwise to $\mu \to Q^*(\mu, t, x, u)$ for any $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.

By Lemma A.8.3, we thus have $\left|Q^{\Phi_{\eta_n}(\mu)}(\mu, t, x, u) - Q^*(\mu, t, x, u)\right| \to 0$ uniformly. Therefore, for any $\varepsilon > 0$, there exists an integer N by uniform convergence such that for all integers $n > N$ we have

$$Q^{\pi_n^*}(\mu_n^*, t, x, u) \geq Q^*(\mu_n^*, t, x, u) - \varepsilon\,,$$

and by the same argument as in Lemma A.8.1, using an action gap and separating out the suboptimal actions to obtain vanishing mass on suboptimal actions via uniform convergence of $Q^{\pi_n^*} \to Q^*$,

$$\left|\sum_{u\in\mathcal{U}} \pi_{n,t}^*(u \mid x) Q^{\pi_n^*}(\mu_n^*, t, x, u) - \max_{u\in\mathcal{U}} Q^*(\mu_n^*, t, x, u)\right| \to 0$$

such that the desired result follows immediately by $J^{\mu_n^*}(\pi_n^*) = \sum_{x\in\mathcal{X}} \mu_0(x) \sum_{u\in\mathcal{U}} \pi_{n,t}^*(u \mid x) Q^{\pi_n^*}(\mu_n^*, t, x, u)$ and $\max_\pi J^{\mu_n^*}(\pi) = \sum_{x\in\mathcal{X}} \mu_0(x) \max_{u\in\mathcal{U}} Q^*(\mu_n^*, t, x, u)$. ■

Finally, we show approximate optimality in the actual N-agent game as long as a pair $(\pi^*, \mu^*) \in \Pi \times \mathcal{M}$ with $\mu^* = \Psi(\pi^*)$ has vanishing exploitability in the MFG. By Lemma A.8.4, for any sequence $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ of η_n-Boltzmann MFE with $\eta_n \to 0^+$ and for any $\varepsilon > 0$ there exists an integer $n' \in \mathbb{N}$ such that for all integers $n > n'$ we have

$$J^{\mu_n^*}(\pi_n^*) \geq \max_{\pi} J^{\mu_n^*}(\pi) - \varepsilon \,.$$

Let $\varepsilon' > 0$ be arbitrary and choose a sequence of optimal policies $\{\pi^N\}_{N\in\mathbb{N}}$ such that for all $N \in \mathbb{N}$ we have

$$\pi^N \in \arg\max_{\pi\in\Pi} J_1^N(\pi, \pi_n^*, \ldots, \pi_n^*) \,.$$

By Lemma A.5.6 there exists $N' \in \mathbb{N}$ such that for all $N > N'$ and all $n > n'$, we have

$$\begin{aligned}
\max_{\pi\in\Pi} J_1^N(\pi, \pi_n^*, \ldots, \pi_n^*) - \varepsilon - \varepsilon' &\leq \max_{\pi\in\Pi} J^{\mu_n^*}(\pi) - \varepsilon - \frac{\varepsilon'}{2} \\
&\leq J^{\mu_n^*}(\pi_n^*) - \frac{\varepsilon'}{2} \\
&\leq J_1^N(\pi_n^*, \pi_n^*, \ldots, \pi_n^*)
\end{aligned}$$

which is the desired approximate Nash equilibrium property since $\varepsilon, \varepsilon'$ are arbitrary. This applies by symmetry to all agents.

For RelEnt MFE, the same can be done by first showing the uniform convergence of the soft action-value function to the usual action-value function. For this, note that the smooth maximum Bellman recursion converges to the hard maximum Bellman recursion for any fixed μ.

Lemma A.8.5. *For any $f : \mathcal{U} \to \mathbb{R}$ and any $g : \mathcal{U} \to \mathbb{R}$ with $g(u) > 0$ for all $u \in \mathcal{U}$, we have*

$$\lim_{\eta\to 0^+} \eta \log \sum_{u\in\mathcal{U}} g(u) \exp \frac{f(u)}{\eta} = \max_{u\in\mathcal{U}} f(u) \,.$$

Proof. Let $\delta = \frac{1}{\eta} \to +\infty$. Then, by L'Hospital's rule we have

$$\begin{aligned}
\lim_{\delta\to+\infty} \frac{\log \sum_{u\in\mathcal{U}} g(u) \exp\left(\delta f(u)\right)}{\delta} &= \lim_{\delta\to+\infty} \frac{\sum_{u\in\mathcal{U}} g(u) \exp\left(\delta f(u)\right) f(u)}{\sum_{u\in\mathcal{U}} g(u) \exp\left(\delta f(u)\right)} \\
&= \lim_{\delta\to+\infty} \frac{\sum_{u\in\mathcal{U}} g(u) \exp\left(\delta (f(u) - \max_{u\in\mathcal{U}} f(u))\right) f(u)}{\sum_{u\in\mathcal{U}} g(u) \exp\left(\delta (f(u) - \max_{u\in\mathcal{U}} f(u))\right)} \\
&= \frac{|\mathcal{U}_{\max}| \max_{u\in\mathcal{U}} f(u)}{|\mathcal{U}_{\max}|} = \max_{u\in\mathcal{U}} f(u)
\end{aligned}$$

where $|\mathcal{U}_{\max}|$ is the number of elements in $\mathcal{U}$ that maximize f. ∎

Using this result, we can show pointwise convergence of the soft action-value function to the action-value function.

Lemma A.8.6. *Any sequence of functions $(\mu \mapsto \tilde{Q}_{\eta_n}(\mu, t, x, u))_{n\in\mathbb{N}}$ with $\eta_n \to 0^+$ converges pointwise to $\mu \mapsto Q^*(\mu, t, x, u)$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.*

Proof. Fix $\mu \in \mathcal{M}$. We show by induction that for any $\varepsilon > 0$, there exists $\eta_t > 0$ such that for all $\eta < \eta_t$ we have $\left|\tilde{Q}_\eta(\mu, t, x, u) - Q^*(\mu, t, x, u)\right| < \varepsilon$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$. This holds for $t = T - 1$ and arbitrary $x \in \mathcal{X}, u \in \mathcal{U}$ by Lemma A.8.5, since $r(x, u, \mu_{T-1})$ is independent of η. Assume this holds for $t + 1$ and consider t. Then, by the induction assumption we can choose $\eta_{t+1} > 0$ such that for $\eta < \eta_{t+1}$, as $\eta \to 0^+$ we have

$$\begin{aligned}
&\tilde{Q}_\eta(\mu, t, x, u) \\
&\quad = r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \eta \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right) \\
&\quad \leq r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \eta \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{Q^*(\mu, t+1, x', u') + \frac{\varepsilon}{2}}{\eta}\right) \\
&\quad \to r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \max_{u' \in \mathcal{U}} Q^*(\mu, t+1, x', u') + \frac{\varepsilon}{2}
\end{aligned}$$

by Lemma A.8.5 and monotonicity of $\log$ and $\exp$. Analogously,

$$\begin{aligned}
&\tilde{Q}_\eta(\mu, t, x, u) \\
&\quad \geq r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \eta \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{Q^*(\mu, t+1, x', u') - \frac{\varepsilon}{2}}{\eta}\right) \\
&\quad \to r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \max_{u' \in \mathcal{U}} Q^*(\mu, t+1, x', u') - \frac{\varepsilon}{2}\,.
\end{aligned}$$

Therefore, we can choose $\eta_t < \eta_{t+1}$ such that for all $\eta < \eta_t$ we have

$$\begin{aligned}
&\left|\tilde{Q}_\eta(\mu, t, x, u) - Q^*(\mu, t, x, u)\right| \\
&\quad = \left|\tilde{Q}_\eta(\mu, t, x, u) - \left(r(x, u, \mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \max_{u' \in \mathcal{U}} Q^*(\mu, t+1, x', u')\right)\right| < \varepsilon
\end{aligned}$$

which is the desired result. ∎

We can now show that the soft action-value function converges uniformly to the action-value function as $\eta \to 0^+$.

Lemma A.8.7. *Any sequence of functions $(\mu \mapsto \tilde{Q}_{\eta_n}(\mu, t, x, u))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges uniformly to $\mu \mapsto Q^*(\mu, t, x, u)$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.*

Proof. First, we show that $\tilde{Q}_\eta(\mu, t, x, u)$ is monotonically decreasing in η for $\eta > 0$, i.e. $\frac{\partial}{\partial \eta} \tilde{Q}_\eta(\mu, t, x, u) \leq 0$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$. This is the case for $t = T - 1$ and arbitrary $x \in \mathcal{X}, u \in \mathcal{U}$, since $\tilde{Q}_\eta(\mu, T-1, x, u)$ is constant. Assume this holds for $t + 1$, then for t and arbitrary $x \in \mathcal{X}, u \in \mathcal{U}$ we have

$$\begin{aligned}
\frac{\partial}{\partial \eta} \tilde{Q}_\eta(\mu, t, x, u) &= \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right) \\
&\quad + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mu_t) \eta \frac{\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}\right)}
\end{aligned}$$

$$\cdot \left(-\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta^2} + \frac{1}{\eta}\frac{\partial}{\partial \eta}\tilde{Q}_\eta(\mu, t+1, x', u') \right)$$

$$\leq \max_{x' \in \mathcal{X}} \left(\log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta} \right) - \frac{\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta} \right) \frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta}}{\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta} \right)} \right)$$

by induction hypothesis. Let $\xi_{u'} \equiv \frac{\tilde{Q}_\eta(\mu, t+1, x', u')}{\eta} \in \mathbb{R}$ and $x' \in \mathcal{X}$ arbitrary, then by Jensen's inequality applied to the convex function $\phi(x) = x \log x$ we have

$$\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \phi(\exp \xi_{u'}) \geq \phi\left(\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp \xi_{u'} \right)$$

$$\iff \log\left(\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp \xi_{u'} \right) - \frac{\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \xi_{u'} \exp \xi_{u'}}{\left(\sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp \xi_{u'} \right)} \leq 0\,,$$

such that $\tilde{Q}_\eta(\mu, t, x, u)$ is monotonically decreasing for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ by induction.

Furthermore, $\mathcal{M}$ is compact and both $\tilde{Q}_\eta$ and Q are compositions, sums, products and finite maxima of continuous functions in μ and therefore continuous in μ by the standing assumptions. Since $(\mu \mapsto \tilde{Q}_{\eta_n}(\mu, t, x, u))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges pointwise to $\mu \mapsto Q^*(\mu, t, x, u)$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ by Lemma A.8.6, by Dini's theorem the convergence is uniform. ■

Now that $\tilde{Q}_\eta$ converges uniformly against Q, we can show that RelEnt MFE have vanishing exploitability by replicating the proof for Boltzmann MFE.

Lemma A.8.8. *Any sequence of functions $(\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, x, u))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges pointwise to $\mu \mapsto Q^*(\mu, t, x, u)$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.*

Proof. The proof is the same as in Lemma A.8.1. The only difference is that we additionally choose $n_2 \in \mathbb{N}$ in each induction step such that for all $n > n_2$ we have

$$\left| \tilde{Q}_\eta(\mu, t, x, u) - Q^*(\mu, t, x, u) \right| \leq \frac{\Delta Q_{\min}^{x', \mu}}{4}$$

for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$, which is possible, since by Lemma A.8.7, $\tilde{Q}_\eta$ converges uniformly against Q. As long as we choose $n' \equiv \max(n_1, n_2, \max_{x' \in \mathcal{X}, u' \in \mathcal{U}} n_{x', u'})$, the rest of the proof will apply. ■

Lemma A.8.9. *Any sequence of functions $(\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, x, u))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ fulfills equicontinuity for large enough n: For any $\varepsilon > 0$ and any $\mu \in \mathcal{M}$, we can choose a $\delta > 0$ and an integer $n' \in \mathbb{N}$ such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta$ and for all $n > n'$ we have*

$$\left| Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, x, u) - Q^{\tilde{\Phi}_{\eta_n}(\mu')}(\mu', t, x, u) \right| < \varepsilon\,.$$

Proof. To obtain the desired property, we replicate the proof of Lemma A.8.2 by setting $\mathcal{F} = (\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, x, u))_{n\in\mathbb{N}}$. Any bounds for $\tilde{Q}_\eta$ can be instantiated by the corresponding bound for Q and then bounding the distance between both by uniform convergence. The only differences lie in bounding the terms

$$\left|(\tilde{\Phi}_{\eta_n}(\mu)(u_{\text{sub}} \mid x') - (\tilde{\Phi}_{\eta_n}(\mu')(u_{\text{sub}} \mid x')\right|$$

where the action-value function has been replaced with the soft action-value function. Since $\tilde{Q}_{\eta_n}$ uniformly converges to Q, we instantiate additional requirements $N^{x'}_{t,x,u}, \tilde{N}^{x'}_{t,x,u}$ to let $n > N^{x'}_{t,x,u}, n > \tilde{N}^{x'}_{t,x,u}$ large enough such that η is sufficiently small enough.

The first difference is to obtain

$$\left|\tilde{Q}_{\eta_n}(\mu', t, x, u) - \tilde{Q}_{\eta_n}(\mu, t, x, u)\right| < \frac{\Delta Q^{x',\mu}_{\min}}{4}$$

for all $\mu' \in \mathcal{M}, t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ with $d_{\mathcal{M}}(\mu, \mu')$ sufficiently small. We choose $\hat{\delta}^3_{t,x,u}$ slightly stronger than in the original proof, such that if $d_{\mathcal{M}}(\mu, \mu') < \hat{\delta}^3_{t,x,u}$, we have

$$\left|Q^*(\mu', t, x, u) - Q^*(\mu, t, x, u)\right| < \frac{\Delta Q^{x',\mu}_{\min}}{12}.$$

We must then additionally choose $N^{x'}_{t,x,u} \in \mathbb{N}$ for each induction step via uniform convergence from Lemma A.8.7 such that as long as $n > N^{x'}_{t,x,u}$, we have

$$\left|\tilde{Q}_{\eta_n}(\mu, t, x, u) - Q^*(\mu, t, x, u)\right| < \frac{\Delta Q^{x',\mu}_{\min}}{12}.$$

This implies the required inequality

$$\begin{aligned}
&\left|\tilde{Q}_{\eta_n}(\mu', t, x, u) - \tilde{Q}_{\eta_n}(\mu, t, x, u)\right| \\
&\leq \left|\tilde{Q}_{\eta_n}(\mu', t, x, u) - Q^*(\mu', t, x, u)\right| + \left|Q^*(\mu', t, x, u) - Q^*(\mu, t, x, u)\right| \\
&\quad + \left|Q^*(\mu, t, x, u) - \tilde{Q}_{\eta_n}(\mu, t, x, u)\right| < \frac{\Delta Q^{x',\mu}_{\min}}{4}
\end{aligned}$$

and we can proceed as in the original proof.

The second difference lies in choosing $\delta^{4,x'}_{t,x,u}$. Note that $\tilde{Q}_{\eta_n}$ is still bounded by M_Q, see Lemma A.7.1. However, since $\tilde{Q}_{\eta_n}$ might no longer be Lipschitz with the same constant as Q^*, we choose an additional integer $\tilde{N}^{x'}_{t,x,u} \in \mathbb{N}$ for each induction step by Lemma A.8.7, such that as long as $n > \tilde{N}^{x'}_{t,x,u}$, we have

$$\left|\tilde{Q}_{\eta_n}(\mu, t, x, u) - Q^*(\mu, t, x, u)\right| \leq \Delta^{x'}_Q \equiv \frac{\frac{\varepsilon_{t,x,u}}{16M_Q|\mathcal{U}|}}{4R^{\max}_q|\mathcal{U}| \cdot \frac{1}{\eta^{x'}_{\min}} \exp\left(\frac{2M_Q}{\eta^{x'}_{\min}}\right)}$$

for any $\mu' \in \mathcal{M}, t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$. The required bound then follows immediately from

$$\left|(\Phi_{\eta_n}(\mu)(u_{\text{sub}} \mid x') - (\Phi_{\eta_n}(\mu')(u_{\text{sub}} \mid x')\right|$$

$$
\begin{aligned}
&\leq R_q^{\max} \sum_{u' \neq u_{\text{sub}}} \left| \exp\left(\frac{\tilde{Q}_{\eta_n}(\mu', t, x', u') - \tilde{Q}_{\eta_n}(\mu', t, x', u_{\text{sub}})}{\eta} \right) \right. \\
&\qquad \left. - \exp\left(\frac{\tilde{Q}_{\eta_n}(\mu, t, x', u') - \tilde{Q}_{\eta_n}(\mu, t, x', u_{\text{sub}})}{\eta} \right) \right| \\
&\leq R_q^{\max} \sum_{u' \neq u_{\text{sub}}} \left| \frac{1}{\eta} \exp\left(\frac{\xi_{u'}}{\eta} \right) \right| \left| (\tilde{Q}_{\eta_n}(\mu', t, x', u') - \tilde{Q}_{\eta_n}(\mu', t, x', u_{\text{sub}})) \right. \\
&\qquad \left. - (\tilde{Q}_{\eta_n}(\mu, t, x', u') - \tilde{Q}_{\eta_n}(\mu, t, x', u_{\text{sub}})) \right| \\
&\leq R_q^{\max} |\mathcal{U}| \cdot \frac{1}{\eta_{\min}^{x'}} \exp\left(\frac{2M_Q}{\eta_{\min}^{x'}} \right) \left(\left| \tilde{Q}_{\eta_n}(\mu', t, x', u') - \tilde{Q}_{\eta_n}(\mu, t, x', u') \right| \right. \\
&\qquad \left. + \left| \tilde{Q}_{\eta_n}(\mu, t, x', u_{\text{sub}}) - \tilde{Q}_{\eta_n}(\mu', t, x', u_{\text{sub}}) \right| \right) \\
&\leq R_q^{\max} |\mathcal{U}| \cdot \frac{1}{\eta_{\min}^{x'}} \exp\left(\frac{2M_Q}{\eta_{\min}^{x'}} \right) \cdot \left(2K_Q d_{\mathcal{M}}(\mu, \mu') + 4\Delta_Q^{x'} \right) \\
&\leq R_q^{\max} |\mathcal{U}| \cdot \frac{1}{\eta_{\min}^{x'}} \exp\left(\frac{2M_Q}{\eta_{\min}^{x'}} \right) \cdot \left(2K_Q d_{\mathcal{M}}(\mu, \mu') \right) + \frac{\varepsilon_{t,x,u}}{16 M_Q |\mathcal{U}|} < \frac{\varepsilon_{t,x,u}}{8 M_Q |\mathcal{U}|}
\end{aligned}
$$

as in the original proof by letting $d_{\mathcal{M}}(\mu, \mu') < \delta_{t,x,u}^{4,x'}$ and choosing

$$
\delta_{t,x,u}^{4,x'} = \frac{\varepsilon_{t,x,u} \eta_{\min}^{x'}}{16 M_Q |\mathcal{U}|^2 R_q^{\max} \cdot \exp\left(\frac{2M_Q}{\eta_{\min}^{x'}} \right) \cdot 2K_Q} .
$$

The rest of the proof is analogous. We obtain the additional requirement $n > N_{t,x,u}^{x'}$, $n > \tilde{N}_{t,x,u}^{x'}$ for some integers $N_{t,x,u}^{x'}, \tilde{N}_{t,x,u}^{x'}$ and each $t \in \mathcal{T}, x \in \mathcal{X}, x' \in \mathcal{X}, u \in \mathcal{U}$. By choosing $n' \equiv \max_{t \in \mathcal{T}, x \in \mathcal{X}, x' \in \mathcal{X}, u \in \mathcal{U}} \max(N_{t,x,u}^{x'}, \tilde{N}_{t,x,u}^{x'})$, the desired result holds as long as $n > n'$. ■

From this property, we again obtain the desired uniform convergence via compactness of $\mathcal{M}$.

Lemma A.8.10. *Any sequence of functions $(\mu \mapsto Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, x, u))_{n \in \mathbb{N}}$ with $\eta_n \to 0^+$ converges uniformly to $\mu \mapsto Q^*(\mu, t, x, u)$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.*

Proof. Fix $\varepsilon > 0, t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$. Then, there exists by Lemma A.8.9 for any point $\mu \in \mathcal{M}$ both $\delta(\mu)$ and n' such that for all $\mu' \in \mathcal{M}$ with $d_{\mathcal{M}}(\mu, \mu') < \delta(\mu)$ for all $n > n'$ we have

$$
\left| Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu, t, x, u) - Q^{\tilde{\Phi}_{\eta_n}(\mu')}(\mu', t, x, u) \right| < \frac{\varepsilon}{3}
$$

which via pointwise convergence from Lemma A.8.8 implies

$$
\left| Q^*(\mu, t, x, u) - Q^*(\mu', t, x, u) \right| \leq \frac{\varepsilon}{3} .
$$

Since $\mathcal{M}$ is compact, it is separable, i.e. there exists a countable dense subset $(\mu_j)_{j \in \mathbb{N}}$ of $\mathcal{M}$. Let $\delta(\mu)$ be as defined above and cover $\mathcal{M}$ by the open balls $(B_{\delta(\mu_j)}(\mu_j))_{j \in \mathbb{N}}$. By the compactness of $\mathcal{M}$, finitely many of these balls $B_{\delta(\mu_{n_1})}(\mu_{n_1}), \ldots, B_{\delta(\mu_{n_k})}(\mu_{n_k})$ cover $\mathcal{M}$. By pointwise convergence from Lemma A.8.8, for any $i = 1, \ldots, k$ we can find integers m_i such that for all $n > m_i$ we have

$$
\left| Q^{\tilde{\Phi}_{\eta_n}(\mu_{n_i})}(\mu_{n_i}, t, x, u) - Q^*(\mu_{n_i}, t, x, u) \right| < \frac{\varepsilon}{3} .
$$

Taken together, we find that for $n > \max(n', \max_{i=1,\dots,k} m_i)$ and arbitrary $\mu \in \mathcal{M}$, we have

$$\begin{aligned}\left|Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu,t,x,u) - Q^*(\mu,t,x,u)\right| &< \left|Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu,t,x,u) - Q^{\tilde{\Phi}_{\eta_n}(\mu_{n_i})}(\mu_{n_i},t,x,u)\right| \\ &\quad + \left|Q^{\tilde{\Phi}_{\eta_n}(\mu_{n_i})}(\mu_{n_i},t,x,u) - Q^*(\mu_{n_i},t,x,u)\right| \\ &\quad + |Q^*(\mu_{n_i},t,x,u) - Q^*(\mu,t,x,u)| \\ &< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon\end{aligned}$$

for some center point μ_{n_i} of a ball containing μ from the finite cover. ■

As a result, a sequence of RelEnt MFE with $\eta \to 0^+$ is approximately optimal in the MFG.

Lemma A.8.11. *For any sequence $(\pi_n^*, \mu_n^*)_{n\in\mathbb{N}}$ of η_n-RelEnt MFE with $\eta_n \to 0^+$ and for any $\varepsilon > 0$ there exists integer $n' \in \mathbb{N}$ such that for all integers $n > n'$ we have*

$$J^{\mu_n^*}(\pi_n^*) \geq \max_{\pi} J^{\mu_n^*}(\pi) - \varepsilon\,.$$

Proof. By Lemma A.8.10, we have $\left|Q^{\tilde{\Phi}_{\eta_n}(\mu)}(\mu,t,x,u) - Q^*(\mu,t,x,u)\right| \to 0$ uniformly. Therefore, for any $\varepsilon > 0$, there exists by uniform convergence an integer n' such that for all integers $n > n'$ we have

$$Q^{\pi_n^*}(\mu_n^*,t,x,u) \geq Q^*(\mu_n^*,t,x,u) - \varepsilon = \max_{\pi\in\Pi} Q^{\pi}(\mu_n^*,t,x,u) - \varepsilon\,,$$

and since by Lemma A.3.1, we have

$$\begin{aligned}J^{\mu_n^*}(\pi_n^*) &= \sum_{x\in\mathcal{X}} \mu_0(x) \cdot \sum_{u\in\mathcal{U}} Q^{\pi_n^*}(\mu_n^*,t,x,u) \\ &\geq \sum_{x\in\mathcal{X}} \mu_0(x) \cdot \max_{\pi\in\Pi} \sum_{u\in\mathcal{U}} Q^{\pi}(\mu_n^*,t,x,u) - \varepsilon \\ &= \max_{\pi\in\Pi} J^{\mu_n^*}(\pi) - \varepsilon\,,\end{aligned}$$

the desired result follows immediately. ■

By repeating the previous argumentation for Boltzmann MFE with Lemma A.5.6 and replacing Lemma A.8.4 with Lemma A.8.11, we obtain the desired result for RelEnt MFE. □

A.9 RELATIVE ENTROPY MEAN FIELD GAMES

We show that the necessary conditions for optimality hold for the candidate solution. (For further insight, see also [376], [377] and references therein.) Fix a MF $\mu \in \mathcal{M}$ and formulate the induced problem as an optimization problem, with $\rho_t(x)$ as the probability of our representative agent visiting state $x \in \mathcal{X}$ at time $t \in \mathcal{T}$, to obtain

$$\max_{\rho,\pi} \quad \sum_{t=0}^{T-1} \sum_{x\in\mathcal{X}} \rho_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) r(x,u,\mu_t)$$

$$
\begin{aligned}
\text{s.t.} \quad & \rho_{t+1}(x') = \sum_{x\in\mathcal{X}} \rho_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) p(x' \mid x, u, \mu_t) \qquad \forall x' \in \mathcal{X}, t \in \{0, \dots, T-2\}, \\
& 1 = \sum_{x\in\mathcal{X}} \rho_t(x) \qquad \forall t \in \{0, \dots, T-1\}, \\
& 1 = \sum_{u\in\mathcal{U}} \pi_t(u \mid x) \qquad \forall x \in \mathcal{X}, t \in \{0, \dots, T-1\}, \\
& 0 \le \rho_t(x), 0 \le \pi_t(u \mid x) \qquad \forall x \in \mathcal{X}, u \in \mathcal{U}, t \in \{0, \dots, T-1\}, \\
& \mu_0(x) = \rho_0(x) \qquad \forall x \in \mathcal{X}.
\end{aligned}
$$

Note that if the agent follows the MF policy of the other agents, we have $\rho_t = \mu_t$. The optimized objective is just the expectation $\mathbb{E}\left[\sum_{t=0}^{T-1} r(x_t, u_t)\right]$. As in [378], we change this objective to include a KL-divergence penalty weighted by the state-visitation distribution ρ_t by introducing the temperature $\eta > 0$ and prior policy $q \in \Pi$ to obtain

$$
\begin{aligned}
\max_{\rho_t, \pi_t} \quad & \sum_{t=0}^{T-1} \sum_{x\in\mathcal{X}} \rho_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) r(x, u, \mu_t) - \eta \sum_{t=0}^{T-1} \sum_{x\in\mathcal{X}} \rho_t(x) \mathrm{KL}(\pi_t(\cdot \mid x) \parallel q_t(\cdot \mid x)) \\
\text{s.t.} \quad & \rho_{t+1}(x') = \sum_{x\in\mathcal{X}} \rho_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) p(x' \mid x, u, \mu_t) \qquad \forall x' \in \mathcal{X}, t \in \{0, \dots, T-2\}, \\
& 1 = \sum_{x\in\mathcal{X}} \rho_t(x) \qquad \forall t \in \{0, \dots, T-1\}, \\
& 1 = \sum_{u\in\mathcal{U}} \pi_t(u \mid x) \qquad \forall x \in \mathcal{X}, t \in \{0, \dots, T-1\}, \\
& 0 \le \rho_t(x), 0 \le \pi_t(u \mid x) \qquad \forall x \in \mathcal{X}, u \in \mathcal{U}, t \in \{0, \dots, T-1\}, \\
& \mu_0(x) = \rho_0(x) \qquad \forall x \in \mathcal{X}.
\end{aligned}
$$

We ignore the constraints $0 \le \pi_t(u \mid x)$ and $0 \le \rho_t(x)$ and see later that they will hold automatically. This results in the simplified optimization problem

$$
\begin{aligned}
\max_{\rho_t, \pi_t} \quad & \sum_{t=0}^{T-1} \sum_{x\in\mathcal{X}} \rho_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) r(x, u, \mu_t) - \eta \sum_{t=0}^{T-1} \sum_{x\in\mathcal{X}} \rho_t(x) \mathrm{KL}(\pi_t(\cdot \mid x) \parallel q_t(\cdot \mid x)) \\
\text{s.t.} \quad & \rho_{t+1}(x') = \sum_{x\in\mathcal{X}} \rho_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) p(x' \mid x, u, \mu_t) \qquad \forall x' \in \mathcal{X}, t \in \{0, \dots, T-2\}, \\
& 1 = \sum_{x\in\mathcal{X}} \rho_t(x) \qquad \forall t \in \{0, \dots, T-1\}, \\
& 1 = \sum_{u\in\mathcal{U}} \pi_t(u \mid x) \qquad \forall x \in \mathcal{X}, t \in \{0, \dots, T-1\}, \\
& \mu_0(x) = \rho_0(x) \qquad \forall x \in \mathcal{X},
\end{aligned}
$$

for which we introduce Lagrange multipliers $\lambda_1(t, s)$, $\lambda_2(t)$, $\lambda_3(t, s)$, $\lambda_4(x)$ and the Lagrangian

$$
\begin{aligned}
& L(\rho, \pi, \lambda_1, \lambda_2, \lambda_3, \lambda_4) \\
& \quad = \sum_{t=0}^{T-1} \sum_{x\in\mathcal{X}} \rho_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) \left(r(x, u, \mu_t) - \eta \log \frac{\pi_t(u \mid x)}{q_t(u \mid x)} \right) \\
& \quad - \sum_{t=0}^{T-1} \sum_{x'\in\mathcal{X}} \lambda_1(t, x') \left(\rho_{t+1}(x') - \sum_{x\in\mathcal{X}} \rho_t(x) \sum_{u\in\mathcal{U}} \pi_t(u \mid x) p(x' \mid x, u, \mu_t) \right)
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{t=0}^{T-1} \lambda_2(t) \left(1 - \sum_{x \in \mathcal{X}} \rho_t(x)\right) \\
&- \sum_{t=0}^{T-1} \sum_{x \in \mathcal{X}} \lambda_3(t,s) \left(\sum_{u \in \mathcal{U}} \pi_t(u \mid x) - 1\right) \\
&- \sum_{x \in \mathcal{X}} \lambda_4(x) \left(\mu_0(x) - \rho_0(x)\right)
\end{aligned}
$$

with the artificial constraint $\lambda_1(T-1,s) \equiv 0$, which allows us to formulate the following necessary conditions for optimality. For $\nabla_{\pi_t(u|x)} L \stackrel{!}{=} 0$ and all $x \in \mathcal{X}, u \in \mathcal{U}, t \in \{0, \dots, T-1\}$, we obtain

$$
\begin{aligned}
\nabla_{\pi_t} L &= \rho_t(x) \left(r(x,u,\mu_t) - \eta \log \frac{\pi_t(u \mid x)}{q_t(u \mid x)} - \eta + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u,\mu_t)\right) - \lambda_3(t,s) \\
\implies \pi_t^*(u \mid x) &= q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) - \eta + \sum_{x'} \lambda_1(t,x') p(x' \mid x,u,\mu_t) - \frac{\lambda_3(t,s)}{\rho_t(x)}}{\eta}\right).
\end{aligned}
$$

For $\nabla_{\lambda_3} L \stackrel{!}{=} 0$ and all $x \in \mathcal{X}, t \in \{0, \dots, T-1\}$, by inserting π_t^* we obtain

$$
\begin{aligned}
\nabla_{\lambda_3(t,s)} L &= 1 - \sum_{u \in \mathcal{U}} \pi_t(u \mid x) \\
\implies 1 &= \sum_{u \in \mathcal{U}} q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) - \eta + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u,\mu_t) - \frac{\lambda_3(t,s)}{\rho_t(x)}}{\eta}\right)
\end{aligned}
$$

which is fulfilled by choosing

$$
\lambda_3^*(t,s) = \eta \rho_t(x) \log \sum_{u \in \mathcal{U}} q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) - \eta + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u,\mu_t)}{\eta}\right)
$$

since it fulfills the required equation

$$
\begin{aligned}
&\sum_{u \in \mathcal{U}} q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) - \eta + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u,\mu_t) - \frac{\lambda_3^*(t,s)}{\rho_t(x)}}{\eta}\right) \\
&= \sum_{u \in \mathcal{U}} q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) - \eta + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u,\mu_t)}{\eta}\right) \\
&\quad \cdot \left(\sum_{u \in \mathcal{U}} q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) - \eta + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u,\mu_t)}{\eta}\right)\right)^{-1} = 1 .
\end{aligned}
$$

Finally, inserting λ_3^* and π^*, for $\nabla_{\rho_t(x)} L$ we obtain

$$
\begin{aligned}
&\nabla_{\rho_t(x)} L \\
&\quad = \sum_{u \in \mathcal{U}} \pi_t(u \mid x) \left(\eta + \lambda_2(t) + \frac{\lambda_3(t,s)}{\rho_t(x)}\right) - \lambda_1(t-1,s) \stackrel{!}{=} 0
\end{aligned}
$$

which implies

$\lambda_1^*(t-1,s)$

$$= \eta + \lambda_2(t) + \eta \log \sum_{u \in \mathcal{U}} q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) - \eta + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u,\mu_t)}{\eta}\right).$$

We can subtract $\lambda_2(t)$ and shift the time index to obtain the soft value function $\tilde{V}_\eta(\mu,t,s)$ defined via terminal condition $\tilde{V}_\eta(\mu,T,s) \equiv 0$ and the recursion

$$\tilde{V}_\eta(\mu,t,s) = \eta \log \sum_{u \in \mathcal{U}} q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) + \sum_{x' \in \mathcal{X}} \tilde{V}_\eta(\mu,t+1,x') p(x' \mid x,u,\mu_t)}{\eta}\right)$$

since then, by normalization the optimal policy for all $x \in \mathcal{X}, u \in \mathcal{U}, t \in \{0,\dots,T-1\}$ is equivalent to

$$\begin{aligned}\pi_t^*(u \mid x) &= \frac{q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u,\mu_t)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{r(x,u',\mu_t) + \sum_{x' \in \mathcal{X}} \lambda_1(t,x') p(x' \mid x,u',\mu_t)}{\eta}\right)} \\ &= \frac{q_t(u \mid x) \exp\left(\frac{r(x,u,\mu_t) + \sum_{x' \in \mathcal{X}} \tilde{V}_\eta(\mu,t+1,x') p(x' \mid x,u,\mu_t)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{r(x,u',\mu_t) + \sum_{x' \in \mathcal{X}} \tilde{V}_\eta(\mu,t+1,x') p(x' \mid x,u',\mu_t)}{\eta}\right)}.\end{aligned}$$

To obtain a recursion in $\tilde{Q}_\eta$, define

$$\begin{aligned}&\tilde{Q}_\eta(\mu,t,x,u) \\ &\equiv r(x,u,\mu_t) + \sum_{x' \in \mathcal{X}} p(x' \mid x,u,\mu_t) \eta \log \sum_{u' \in \mathcal{U}} q_{t+1}(u' \mid x') \exp\left(\frac{\tilde{Q}_\eta(\mu,t+1,x',u')}{\eta}\right)\end{aligned}$$

with terminal condition $\tilde{Q}_\eta(\mu,T,x,u) \equiv 0$ to obtain

$$\pi_t^*(u \mid x) = \frac{q_t(u \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,x,u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u' \mid x) \exp\left(\frac{\tilde{Q}_\eta(\mu,t,x,u')}{\eta}\right)}$$

which is the desired result as π^* fulfills all constraints and determines ρ uniquely. For the uniform prior $q_t(u \mid x) = 1/|\mathcal{U}|$, we obtain the maximum entropy solution.

A.10 IMPLEMENTATION DETAILS

For all the DQN experiments, we use the configurations given in Table A.1 and hyperparameters given in Table A.2. Note that we add epsilon scheduling and a discount factor to DQN for stability reasons, i.e. the loss term has an additional factor smaller than one before the maximum operation, cf. [33]. For the action-value network, we use a fully connected dueling architecture ([379]) with one shared hidden layer of 256 neurons, and one separate hidden layer of 256 neurons for value and advantage stream each. As the activation function, we use ReLU. Further, we use gradient norm clipping and the ADAM optimizer. To allow for time-dependent policies, we append the current time to the observations. The precise algorithms are given in Algorithms 4 to 9.

We transform all discrete-valued observations except time to corresponding one-hot vectors, except in the intractably large Taxi environment where we simply observe one value in $\{0,1\}$ for each tile's

Algorithm 4 Exact FPI

Initialize $\mu^0 = \Psi(q)$ as the MF induced by the uniformly random policy q.
for $k = 0, 1, \cdots$ **do**
  Compute the Q-function $Q^*(\mu^k, t, x, u)$ for fixed μ^k.
  Choose $\pi^k \in \Pi$ such that $\pi_t^k(u \mid x) \implies u \in \arg\max_{u \in \mathcal{U}} Q^k(\mu^k, t, x, u)$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$ by putting all probability mass on the first optimal action, or evenly on all optimal actions.
  Optionally: Overwrite $\pi^k \leftarrow \frac{1}{k+1}\pi^k + \frac{k}{k+1}\pi^{k-1}$. (FP averaged policy)
  Compute the MF $\mu^{k+1} = \Psi(\pi^k)$ induced by π^k.
  Optionally: Overwrite $\mu^{k+1} \leftarrow \frac{1}{k+1}\mu^{k+1} + \frac{k}{k+1}\mu^k$. (FP averaged MFs)

Algorithm 5 Boltzmann / RelEnt iteration

Input: Temperature $\eta > 0$, prior policy $q \in \Pi$.
Initialize $\mu^0 = \Psi(q)$ as the MF induced by q.
for $k = 0, 1, \cdots$ **do**
  Compute the Q-function (Boltzmann) or soft Q-function (RelEnt) $Q(\mu^k, t, x, u)$ for fixed μ^k.
  Define π^k by $\pi_t^k(u \mid x) = \frac{q_t(u|x) \exp\left(\frac{Q(\mu^k, t, x, u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u'|x) \exp\left(\frac{Q(\mu^k, t, x, u')}{\eta}\right)}$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.
  Optionally: Overwrite $\pi^k \leftarrow \frac{1}{k+1}\pi^k + \frac{k}{k+1}\pi^{k-1}$. (FP averaged policy)
  Compute the MF $\mu^{k+1} = \Psi(\pi^k)$ induced by π^k.
  Optionally: Overwrite $\mu^{k+1} \leftarrow \frac{1}{k+1}\mu^{k+1} + \frac{k}{k+1}\mu^k$. (FP averaged MFs)

passenger status. For evaluation of exploitability, we compare the values of the optimal policy and the evaluated policy in the MDP induced by the MF generated by the evaluated policy. In intractable cases, we use DQN to approximately obtain the optimal policy. In this case, we obtain the values by averaging over many episodes in the MDP induced by the MF generated by the evaluated policy via Algorithm 8.

A.11 PROBLEMS

Summarizing properties of the considered problems are given in Table A.3.

Algorithm 6 Boltzmann DQN iteration

Input: Temperature $\eta > 0$, prior policy $q \in \Pi$.
Input: Simulation parameters, DQN hyperparameters.
Initialize $\mu^0 \approx \Psi(q)$ as the MF induced by q using Algorithm 8.
for $k = 0, 1, \cdots$ **do**
  Approximate the Q-function $Q^*(\mu^k, t, x, u)$ using Algorithm 7 on the MDP induced by μ^k.
  Define π^k by $\pi_t^k(u \mid x) = \frac{q_t(u|x) \exp\left(\frac{Q^*(\mu^k, t, x, u)}{\eta}\right)}{\sum_{u' \in \mathcal{U}} q_t(u'|x) \exp\left(\frac{Q^*(\mu^k, t, x, u')}{\eta}\right)}$ for all $t \in \mathcal{T}, x \in \mathcal{X}, u \in \mathcal{U}$.
  Approximately simulate MFs $\mu^{k+1} \approx \Psi(\pi^k)$ induced by π^k using Algorithm 8.

Algorithm 7 DQN

Input: Number of epochs L, mini-batch size N, target update frequency M, replay buffer size D.
Input: Probability of random action ϵ, Discount factor γ, ADAM and gradient clipping parameters.
Initialize network Q_θ, target network $Q_{\theta'} \leftarrow Q_\theta$ and replay buffer $\mathcal{D}$ of size D.
for L epochs **do**
  for $t = 1, \dots, \mathcal{T}$ **do**
    One environment step
      Let new action $a_t \leftarrow \arg\max_{u \in \mathcal{U}} Q_\theta(t, x, u)$, or with probability ϵ sample uniformly random instead.
      Sample new state $x_{t+1} \sim p(x_{t+1} \mid x_t, u_t)$.
      Add transition tuple $(x_t, u_t, r(x_t, u_t), x_{t+1})$ to replay buffer $\mathcal{D}$.
    One mini-batch descent step
      Sample from the replay buffer: $\{(x_t^i, u_t^i, r_t^i, x_{t+1}^i)\}_{i=1,\dots,N} \sim \mathcal{D}$.
      Compute loss $J_Q = \sum_{i=1}^N \left(r_t^i + \gamma \max_{u' \in \mathcal{U}} Q(t+1, x_{t+1}^i, u') - Q(t, x_t^i, u_t^i)\right)^2$.
      Update θ according to $\nabla_\theta J_Q$ using ADAM with gradient norm clipping.
    if number of steps $\bmod M = 0$ **then**
      Update target network $\theta' \leftarrow \theta$.

Algorithm 8 Stochastic MF simulation

Input: Number of MFs K, number of particles M, policy π.
for $k = 1, \dots, K$ **do**
  Initialize particles $x_m^0 \sim \mu_0$ for all $m = 1, \dots, M$.
  for $t \in \mathcal{T}$ **do**
    Define empirical measure $\mu_t^k \leftarrow \frac{1}{M} \sum_{m=1}^M \delta_{x_m^t}$.
    for $m = 1, \dots, M$ **do**
      Sample action $u \sim \pi_t(u \mid x_m^t)$.
      Sample new particle state $x_m^{t+1} \sim p(x_m^{t+1} \mid x_m^t, u, \mu_t^k)$.
return average empirical MF $(\frac{1}{K} \sum_{k=1}^K \mu_t^k)_{t \in \mathcal{T}}$

LR. Similar to the example mentioned in the main text, we let a large number of agents choose simultaneously between going left (L) or right (R). Afterwards, each agent shall be punished proportional to the number of agents that chose the same action, but more-so for choosing right than left.

More formally, let $\mathcal{X} = \{C, L, R\}, \mathcal{U} = \mathcal{X} \setminus \{C\}$, $\mu_0(C) = 1$, $r(x, u, \mu_t) = -\mathbf{1}_{\{L\}}(x) \cdot \mu_t(L) - 2 \cdot \mathbf{1}_{\{R\}}(x) \cdot \mu_t(R)$ and $\mathcal{T} = \{0, 1\}$. Note the difference to the toy example in the main text: right is punished more than left. The transition function allows picking the next state directly, i.e. for all $s, x' \in \mathcal{X}, u \in \mathcal{U}$,

$$\mathbb{P}(x_{t+1} = x' \mid x_t = x, u_t = u) = \mathbf{1}_{\{x'\}}(u)\,.$$

For this example, we have $K_Q = 1$ since the return Q of the initial state changes linearly with μ_1 and lies between 0 and -2, while the distance between two MFs is also bounded by 2. Analogously, $K_\Psi = 1$ since $(\Psi(\pi))_1$ similarly changes linearly with π_0, and both can change at most by 2. Thus, we obtain guaranteed convergence via Boltzmann iteration if $\eta > 1$. In numerical evaluations, we see convergence already for $\eta \geq 0.7$.

Algorithm 9 Prior descent

Input: Number of outer iterations I.
Input: Initial prior policy $q \in \Pi$.
for outer iteration $i = 1, \ldots, I$ **do**
  Find η heuristically or minimally such that Algorithm 5 with temperature η and prior q converges.
  if no such η exists **then**
    return q
  $q \leftarrow$ solution of Algorithm 5 with temperature η and prior q.

TABLE A.1: Hyperparameter configurations for Boltzmann DQN Iteration.

Parameter	RPS	SIS	Taxi
FPI count	1000	50	15
Number of particles for MF	1000	1000	200
Number of MFs	5	5	5
Number of episodes for evaluation	2000	2000	500

RPS. This game is inspired by [380] and their generalized non-zero-sum version of Rock-Paper-Scissors, for which classical FP would not converge. Each of the agents can choose between rock, paper and scissors, and obtains a reward proportional to double the number of beaten agents minus the number of agents beating the agent. We modify the proportionality factors such that a uniformly random prior policy does not constitute a MFE.

Let $\mathcal{X} = \{0, R, P, S\}, \mathcal{U} = \mathcal{X} \setminus \{0\}, \mu_0(0) = 1, \mathcal{T} = \{0, 1\}$, and for any $u \in \mathcal{U}, \mu_t \in \mathcal{P}(\mathcal{X})$,

$$
\begin{aligned}
r(R, u, \mu_t) &= 2 \cdot \mu_t(S) - 1 \cdot \mu_t(P), \\
r(P, u, \mu_t) &= 4 \cdot \mu_t(R) - 2 \cdot \mu_t(S), \\
r(S, u, \mu_t) &= 6 \cdot \mu_t(P) - 3 \cdot \mu_t(R)\,.
\end{aligned}
$$

The transition function allows picking the next state directly, i.e. for all $x, x' \in \mathcal{X}, u \in \mathcal{U}$,

$$
\mathbb{P}(x_{t+1} = x' \mid x_t = x, u_t = u) = \mathbf{1}_{\{x'\}}(u)\,.
$$

TABLE A.2: Hyperparameter configurations for DQN.

Hyperparameter	Value
Replay buffer size	10000
ADAM Learning rate	0.0005
Discount factor	0.99
Target update frequency	500
Gradient clipping norm	40
Mini-batch size	128
Epsilon schedule	1 linearly down to 0.02 at 0.8 times maximum steps
Total epochs	1000

TABLE A.3: Overview of problem properties.

Problem	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{U}\rvert$
LR	2	3	2
RPS	2	4	3
SIS	50	2	2
Taxi	100	$\sim 2^{27}$	5

SIS. In this problem, a large number of agents can choose between social distancing (D) or going out (U). If a susceptible (S) agent chooses social distancing, they may not become infected (I). Otherwise, an agent may become infected with a probability proportional to the number of agents being infected. If infected, an agent will recover with a fixed chance every time step. Both social distancing and being infected have an associated cost.

Let $\mathcal{X} = \{S, I\}$, $\mathcal{U} = \{U, D\}$, $\mu_0(I) = 0.6$, $r(x, u, \mu_t) = -\mathbf{1}_{\{I\}}(x) - 0.5 \cdot \mathbf{1}_{\{D\}}(u)$ and $\mathcal{T} = \{0, \dots, 50\}$. We find that similar parameters produce similar results, and set the transition probability mass functions as

$$\begin{aligned}\mathbb{P}(x_{t+1} = S \mid x_t = I) &= 0.3\\ \mathbb{P}(x_{t+1} = I \mid x_t = S, u_t = U) &= 0.9^2 \cdot \mu_t(I)\\ \mathbb{P}(x_{t+1} = I \mid x_t = S, u_t = D) &= 0\,.\end{aligned}$$

TAXI. In this problem, we consider a $K \times L$ grid. The state is described by a tuple (x, y, x', y', p, B) where (x, y) is the agent's position, (x', y') indicates the current desired destination of the passenger or is $(0, 0)$ otherwise, and $p \in \{0, 1\}$ indicates whether a passenger is in the taxi or not. Finally, B is a $K \times L$ matrix indicating whether a new passenger is available for the taxi on the corresponding tile. All taxis start on the same tile and have no passengers in the queue or on the map at the beginning. The problem runs for 100 time steps.

The taxi can choose between five actions W, U, D, L, R, where W (Wait) allows the taxi to pick up / deliver passengers, and U, D, L, R (Up, Down, Left, Right) allows it to move in all four directions. As there are many taxis, there is a chance of a jam on tile s given by $\min(0.7, 10 \cdot \mu_t(x))$, i.e. the taxi will not move with this probability. The taxi also cannot move into walls or back into the starting tile, in which case it will stay on its current tile. With a probability of 0.8, a new passenger spawns on one randomly chosen free tile of each region. On picking up a passenger, the destination is generated by randomly picking any free tile of the same region. Delivering passengers to a destination and picking them up gives a reward of 1 in region 1 and 1.2 in region 2.

For our experiments, we use the following small map, where S denotes the starting tile, 1 denotes a free tile from region 1, 2 denotes a free tile from region 2 and H denotes an impassable wall:

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ H & S & H \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$$

This produces a similar situation as in LR, where a fraction of taxis should choose each region so the values balance out, while also requiring solution of a problem that is intractable to solve exactly via dynamic programming.

A.12 ADDITIONAL EXPERIMENTS

In Figure A.1, we observe that prior descent for both Boltzmann and RelEnt MFE with the same uniform prior policy performs qualitatively similarly, and coincide in LR and SIS except for numerical inaccuracies. It can be seen that using a temperature sufficiently low to converge in LR and RPS allows prior descent to descend to the exact MFE iteratively. In SIS on the other hand, picking a fixed temperature that converges for the initial uniform prior policy does not guarantee monotonic improvement of exploitability afterwards. Instead, by applying the heuristic

$$\eta_{i+1} = \eta_i \cdot c$$

for each outer iteration i, where $c \geq 1$ adjusts the temperature after each outer iteration, we avoid scanning over all temperatures in each step and reach convergence to a good approximate MFE for both Boltzmann and MaxEnt iteration.

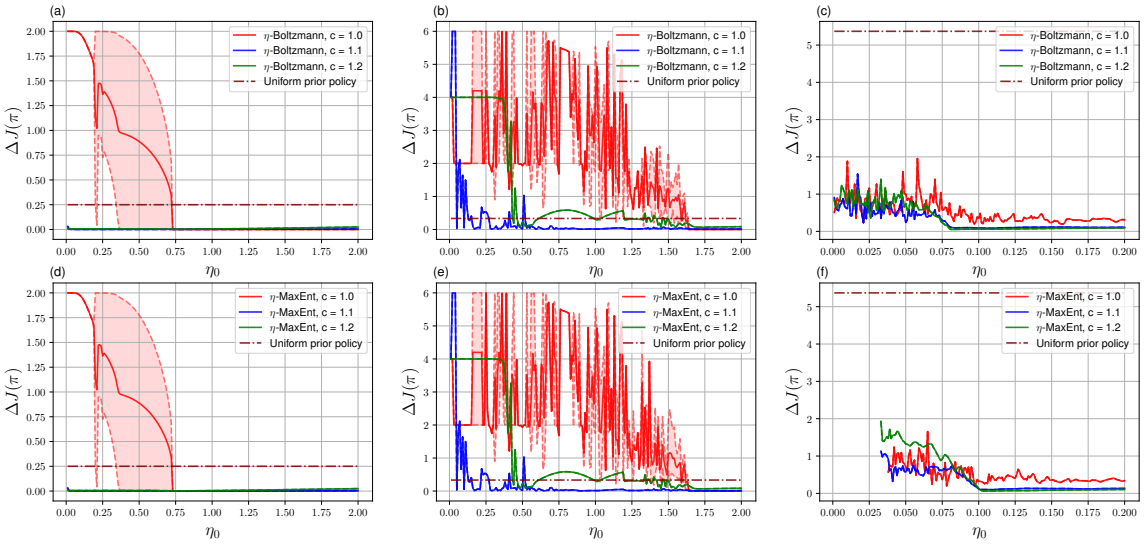


FIGURE A.1: Convergence in exploitability of prior iteration algorithm. Mean exploitability (straight lines), maximum and minimum (dashed lines) over the final 10 iterations of the last outer iteration. 50 outer iterations and 100 inner iterations each; (a, d) LR; (b, e) RPS; (c, f) SIS. Maximum entropy (MaxEnt) results begin at higher temperatures due to limited floating point accuracy. The exploitability of the initial uniform prior policy is indicated by the dashed horizontal line.

In Figure A.2 empirical results are shown for FP variants averaging only policy or MF. In the simple one-step toy problems LR and RPS, averaging the policies appears to converge to the exact solution without regularization and to the regularized solution with regularization. Averaging the MF on the other hand fails, since this method can only produce deterministic policies. By applying any amount of regularization, averaging the MF is led to success in LR and SIS. Nonetheless, both methods fail to converge to the MFE in SIS and produce worse results than obtained by prior descent in Figure A.1.

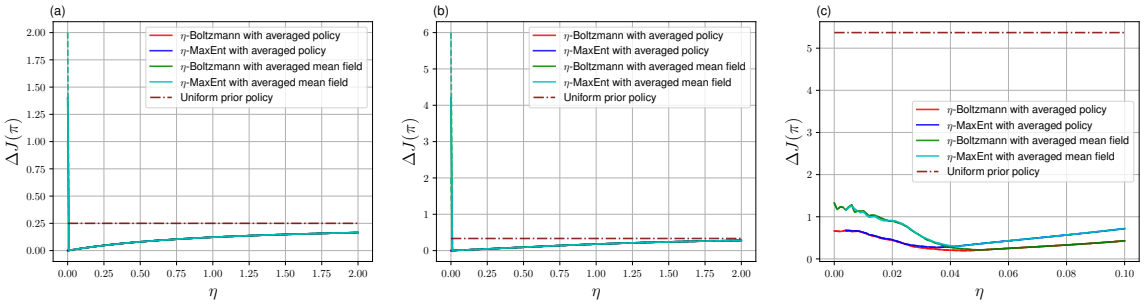


FIGURE A.2: Convergence in exploitability of FP algorithm. Mean exploitability over the final 10 iterations. Dashed lines represent maximum and minimum over the final 10 iterations. (a) LR, 10000 iterations; (b) RPS, 10000 iterations; (c) SIS, 1000 iterations. The exploitability of the uniform prior policy is indicated by the dashed horizontal line.

In Figure A.3 we depict the convergence of exploitability and MF of MaxEnt iteration in SIS. The results are qualitatively similar with Boltzmann iteration and, as in the main text, show the convergence behaviour near the critical temperature leading to convergence.

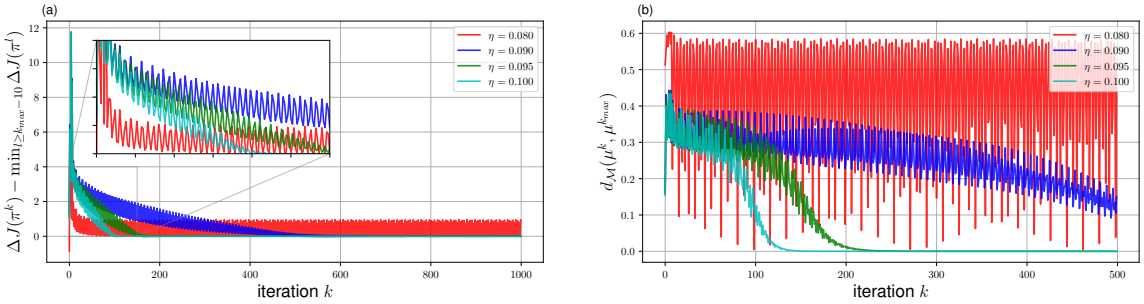


FIGURE A.3: Change in exploitability and MF over iterations. (a) Difference between current and final minimum exploitability over the last 10 iterations; (b) Distance between current and final MF, cut off at 500 iterations for readability. Plotted for the η-RelEnt iterations in SIS for the indicated temperature settings and uniform prior policy.

In Figure A.4 we depict the convergence of exploitability for Boltzmann DQN iteration in SIS and Taxi during one of the runs. All 4 other runs show similar qualitative behaviour. As can be seen, the highest temperature of 0.2 shows less oscillatory behaviour, stabilizing Boltzmann DQN iteration. In Taxi, it can be seen that the used temperatures are insufficient to allow Boltzmann DQN iteration to converge. We believe that using prior descent could allow for better results. We could not verify this due to the high computational cost, as this includes repeatedly and sequentially solving an expensive RL problem.

Finally, in Figure A.5 we depict the resulting behavior in the SIS case. In the Boltzmann iteration result, at the beginning the number of infected is high enough to make social distancing the optimal action to take. As the number of infected falls, it reaches an equilibrium point where both social distancing or potentially getting infected are of equal value. Finally, as the game ends at time $t = T = 50$, there is no point in social distancing any more. Our approach yields intuitive results here, while exact FPI and FP fail to converge.

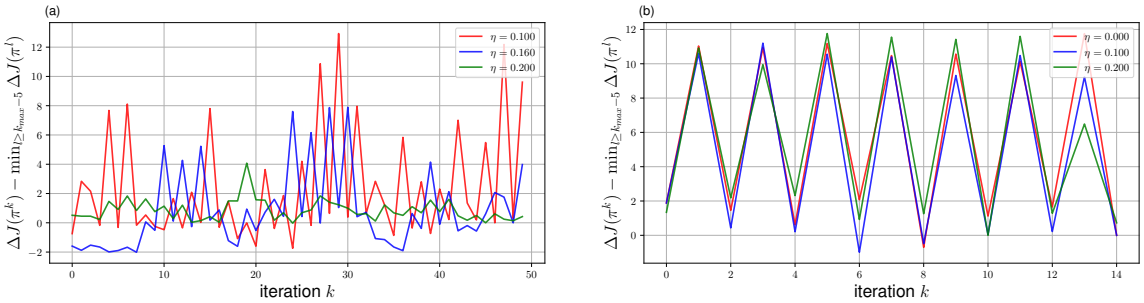


FIGURE A.4: Change in exploitability over Boltzmann DQN iterations. Difference between current and final estimated minimum exploitability over the last 5 iterations. (a) SIS, 50 iterations; (b) Taxi, 15 iterations. Plotted for the η-Boltzmann DQN iteration for the indicated temperature settings and uniform prior policy.

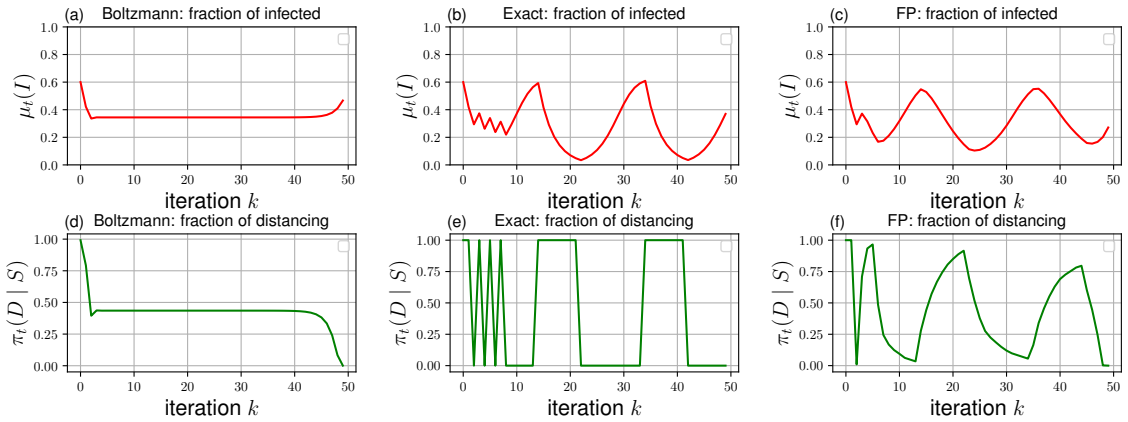


FIGURE A.5: Qualitative behavior in SIS. Fraction of infected agents and fraction of susceptible agents picking social distancing over time. (a, d): Boltzmann iteration ($\eta = 0.07$); (b, e): exact FPI; (c, f): FP (averaging both policy and MF) results in SIS after 500 iterations. More iterations and averaging only policy or MF show same qualitative results.

B

APPENDIX B: SUPPLEMENTARY DETAILS ON SECTION 3.2

B.1 THEORETICAL DETAILS

In this section, we will give all intermediate results required to prove the results in the main text, as well as additional results, e.g. for the uncontrolled case. For convenience, we first state all obtained theoretical result. Proofs for each of the theorems and corollaries can be found in their own sections further below.

Note that except for Theorem 3.2.1, as mentioned in the main text we can also slightly weaken Assumption 3.2.2 to block-wise Lipschitz continuous W, i.e. there exist $L_W > 0$ and disjoint intervals $\{\mathcal{I}_1, \dots, \mathcal{I}_Q\}, \cup_i \mathcal{I}_i = \mathcal{I}$ s.t. $\forall i, j \in \{1, \dots, Q\}$,

$$|W(x,y) - W(\tilde{x},\tilde{y})| \leq L_W(|x-\tilde{x}| + |y-\tilde{y}|), \quad \forall (x,y),(\tilde{x},\tilde{y}) \in \mathcal{I}_i \times \mathcal{I}_j \tag{B.1.1}$$

which is fulfilled e.g. for block-wise Lipschitz-continuous or block-wise constant graphons.

For $\alpha \in \mathcal{I}$, define the α-neighborhood maps $\mathbb{G}^\alpha \colon \mathcal{M}_t \to \mathcal{P}(\mathcal{X})$ and empirical α-neighborhood maps $\mathbb{G}^\alpha_N \colon \mathcal{M}_t \to \mathcal{P}(\mathcal{X})$ as

$$\mathbb{G}^\alpha(\boldsymbol{\mu}_t) \coloneqq \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^\beta \,\mathrm{d}\beta, \quad \mathbb{G}^\alpha_N(\boldsymbol{\mu}_t) \coloneqq \int_{\mathcal{I}} W_N(\alpha,\beta)\mu_t^\beta \,\mathrm{d}\beta \tag{B.1.2}$$

and note how we naturally have $\mathbb{G}^\alpha_t = \mathbb{G}^\alpha(\boldsymbol{\mu}_t)$ in the MF system and $\mathbb{G}^i_t = \mathbb{G}^{\frac{i}{N}}_N(\boldsymbol{\mu}^N_t)$ in the finite system. Finally, for $\boldsymbol{\nu}, \boldsymbol{\nu}' \in \mathcal{M}_t$, $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ and graphon W, define the ensemble transition kernel operator $P^{\boldsymbol{\pi}}_{t,\boldsymbol{\nu},W} \colon \mathcal{M}_t \to \mathcal{M}_t$ via

$$\left(\boldsymbol{\nu} P^{\boldsymbol{\pi}}_{t,\boldsymbol{\nu}',W}\right)^\alpha \equiv \sum_{x\in\mathcal{X}} \nu^\alpha(x) \sum_{u\in\mathcal{U}} \pi^\alpha_t(u \mid x) p\left(\cdot \mid x, u, \int_{\mathcal{I}} W(\alpha,\beta)\nu'^\beta \,\mathrm{d}\beta\right) \tag{B.1.3}$$

and note how we have $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t,W}$ in the MF system.

After showing Theorem 3.2.2, we continue by showing convergence of the law of deviating agent state x^i_t to the law of the corresponding auxiliary MF systems given by

$$\hat{x}^{\frac{i}{N}}_0 \sim \mu_0, \quad \hat{u}^{\frac{i}{N}}_t \sim \hat{\pi}_t(\hat{u}^{\frac{i}{N}}_t \mid \hat{x}^{\frac{i}{N}}_t), \quad \hat{x}^{\frac{i}{N}}_{t+1} \sim p(\hat{x}^{\frac{i}{N}}_{t+1} \mid \hat{x}^{\frac{i}{N}}_t, \hat{u}^{\frac{i}{N}}_t, \mathbb{G}^{\frac{i}{N}}_t), \quad \forall t \in \mathcal{T} \tag{B.1.4}$$

for almost all agents i as $N \to \infty$.

Lemma B.1.1. *Consider Lipschitz continuous $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ up to a finite number of discontinuities D_π, with associated MF ensemble $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$. Under Assumptions 3.2.1 and 3.2.2 and the N-agent policy $(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) \in \Pi^N$ where $(\pi^1, \pi^2, \dots, \pi^N) = \Gamma_N(\boldsymbol{\pi}) \in \Pi^N$, $\hat{\pi} \in \Pi$ arbitrary, for any uniformly bounded family of functions $\mathcal{G}$ from $\mathcal{X}$ to $\mathbb{R}$ and any $\varepsilon, \delta > 0$, $t \in \mathcal{T}$, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have*

$$\sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[g(x^i_t)\right] - \mathbb{E}\left[g(\hat{x}^{\frac{i}{N}}_t)\right] \right| < \varepsilon \tag{B.1.5}$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$.

Similarly, for any uniformly Lipschitz, uniformly bounded family of measurable functions $\mathcal{H}$ from $\mathcal{X} \times \mathcal{B}_1(\mathcal{X})$ to $\mathbb{R}$ and any $\varepsilon, \delta > 0$, $t \in \mathcal{T}$, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\sup_{h\in\mathcal{H}} \left| \mathbb{E}\left[h(x^i_t, \mathbb{G}^{\frac{i}{N}}_N(\boldsymbol{\mu}^N_t))\right] - \mathbb{E}\left[h(\hat{x}^{\frac{i}{N}}_t, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] \right| < \varepsilon \tag{B.1.6}$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$.

As a direct implication of the above results, the objective functions of almost all agents converge uniformly to the MF objectives.

Corollary B.1.1. *Consider Lipschitz continuous $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ up to a finite number of discontinuities D_π, with associated MF ensemble $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$. Under Assumptions 3.2.1 and 3.2.2 and the N-agent policy $(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) \in \Pi^N$ where $(\pi^1, \pi^2, \dots, \pi^N) = \Gamma_N(\boldsymbol{\pi}) \in \Pi^N$, $\hat{\pi} \in \Pi$ arbitrary, for any $\varepsilon, \delta > 0$, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have*

$$\left| J^N_i(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \pi^N) - J^{\boldsymbol{\mu}}_{\frac{i}{N}}(\hat{\pi}) \right| < \varepsilon \tag{B.1.7}$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$.

The approximate Nash property (Theorem 3.2.3) of a GMFE $(\boldsymbol{\pi}, \boldsymbol{\mu})$ then follows immediately from the definition of a GMFE, since $\boldsymbol{\pi}$ is by definition optimal under $\boldsymbol{\mu}$.

As a corollary, we also obtain results for the uncontrolled case without actions, which is equivalent to the case where $|\mathcal{U}| = 1$, i.e. there being only one trivial policy that is always optimal.

Corollary B.1.2. *Under Assumption 3.2.1 and $|\mathcal{U}| = 1$, we have for all measurable functions $f\colon \mathcal{X} \times \mathcal{I} \to \mathbb{R}$ uniformly bounded by $|f| \leq M_f$ and all $t \in \mathcal{T}$ that*

$$\mathbb{E}\left[\left|\boldsymbol{\mu}_t^N(f) - \boldsymbol{\mu}_t(f)\right|\right] \to 0\,. \tag{B.1.8}$$

Furthermore, if the convergence in Assumption 3.2.1 is at rate $\mathcal{O}(1/\sqrt{N})$, the rate of convergence is also at $\mathcal{O}(1/\sqrt{N})$.

If further Assumption 3.2.2 holds, then for any uniformly bounded family of functions $\mathcal{G}$ from $\mathcal{X}$ to $\mathbb{R}$ and any $\varepsilon, \delta > 0$, $t \in \mathcal{T}$, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}\left[g(x_t^i)\right] - \mathbb{E}\left[g(\hat{x}_t^{\frac{i}{N}})\right] \right| < \varepsilon \tag{B.1.9}$$

uniformly over $i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$, and similarly for any uniformly Lipschitz, uniformly bounded family of measurable functions $\mathcal{H}$ from $\mathcal{X} \times \mathcal{B}_1(\mathcal{X})$ to $\mathbb{R}$ and any $\varepsilon, \delta > 0$, $t \in \mathcal{T}$, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\sup_{h \in \mathcal{H}} \left| \mathbb{E}\left[h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N))\right] - \mathbb{E}\left[h(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] \right| < \varepsilon \tag{B.1.10}$$

uniformly over $i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$.

B.2 PROOF OF THEOREM 3.2.1

Proof. First, we will verify [24], Assumption 1 for the MFG with dynamics given by Eq. (3.2.14). For this purpose, as in [24] let us metrize the product space with the sup-metric, and equip the space $\mathcal{P}(\mathcal{X} \times \mathcal{I})$ with the weak topology. Note that the results hold for both the finite and infinite horizon setting, see [24], Remark 6.

(a) The reward function $\tilde{r}((x,\alpha), u, \tilde{\mu}) \coloneqq r(x, u, \int_{\mathcal{I}} W(\alpha_t, \beta)\tilde{\mu}_t(\cdot, \beta)\,\mathrm{d}\beta)$ is continuous, since for $((x_n, \alpha_n), u_n, \tilde{\mu}_n) \to ((x,\alpha), u, \tilde{\mu})$ we have

$$\int_{\mathcal{I}} W(\alpha_n, \beta)\tilde{\mu}_n(\cdot, \beta)\,\mathrm{d}\beta \to \int_{\mathcal{I}} W(\alpha, \beta)\tilde{\mu}(\cdot, \beta)\,\mathrm{d}\beta$$

by Lipschitz continuity of W and weak convergence of $\tilde{\mu}_n$, and therefore

$$r(x_n, u_n, \int_{\mathcal{I}} W(\alpha_n, \beta)\tilde{\mu}_n(\cdot, \beta)\,\mathrm{d}\beta) \to r(x, u, \int_{\mathcal{I}} W(\alpha, \beta)\tilde{\mu}(\cdot, \beta)\,\mathrm{d}\beta)$$

by Assumption 3.2.2.

(b) The action space is compact and the state space is locally compact.

(c) Consider the moment function $w(x, \alpha) \equiv 2$. In this case, we can choose $\zeta = 1$ (we use ζ instead of α in [24]).

(d) The stochastic kernel $\tilde{p}$ that fulfills Eq. (3.2.14) such that $(\tilde{x}_{t+1}, \alpha_{t+1}) \sim \tilde{p}(\tilde{x}_{t+1}, \alpha_{t+1} \mid (\tilde{x}_t, \alpha_t), \tilde{u}_t, \tilde{\mu}_t)$ is weakly continuous, since for $((x_n, \alpha_n), u_n, \tilde{\mu}_n) \to ((x, \alpha), u, \tilde{\mu})$ we again have

$$\int_{\mathcal{I}} W(\alpha_n, \beta)\tilde{\mu}_n(\cdot, \beta)\,\mathrm{d}\beta \to \int_{\mathcal{I}} W(\alpha, \beta)\tilde{\mu}(\cdot, \beta)\,\mathrm{d}\beta$$

and therefore for any Lipschitz and bounded $f\colon \mathcal{X} \times \mathcal{I} \to \mathbb{R}$,

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathcal{I}} f(y, \alpha)\,\tilde{p}(\mathrm{d}(y, \alpha) \mid (x_n, \alpha_n), u_n, \tilde{\mu}_n) \\
&\quad= \int_{\mathcal{I}}\int_{\mathcal{X}} f(y, \alpha)\,p(\mathrm{d}y \mid (x_n, \alpha), u_n, \tilde{\mu}_n)\,\delta_{\alpha_n}(\mathrm{d}\alpha) \\
&\quad= \int_{\mathcal{X}} f(y, \alpha_n)\,p(\mathrm{d}y \mid x_n, u_n, \int_{\mathcal{I}} W(\alpha_n, \beta)\tilde{\mu}_n(\cdot, \beta)\,\mathrm{d}\beta) \\
&\quad\to \int_{\mathcal{X}} f(y, \alpha)\,p(\mathrm{d}y \mid x, u, \int_{\mathcal{I}} W(\alpha, \beta)\tilde{\mu}(\cdot, \beta)\,\mathrm{d}\beta) \\
&\quad= \int_{\mathcal{X}\times\mathcal{I}} f(y, \alpha)\,\tilde{p}(\mathrm{d}(y, \alpha) \mid (x, \alpha), u, \tilde{\mu})
\end{aligned}$$

by disintegration of $\tilde{p}$ and Eq. (3.2.14).

(e) By boundedness of r, we trivially have $v(x) \equiv 1 \leq \infty$.

(f) By boundedness of r, we can trivially choose $\beta = 1$ (we flip the usage of β and γ).

(g) As a result of the above choices, $\zeta\gamma\beta = \gamma < 1$ trivially.

By [24], Theorem 3.3 we have the existence of a MFE $(\tilde{\pi}, \tilde{\mu})$ with some Markovian feedback policy $\tilde{\pi}$ acting on the state $(\tilde{x}_t, \alpha_t)$. By defining the MF and policy ensembles $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ via $\pi_t^\alpha(u \mid x) = \tilde{\pi}_t(u \mid x, \alpha)$, $\mu_t^\alpha(x) = \tilde{\mu}_t(x, \alpha)$, we obtain existence of the α-a.e. optimal policy ensemble $\boldsymbol{\pi}$, since at any time $t \in \mathcal{T}$, the joint state-action distribution $\tilde{\mu}_t \otimes \tilde{\pi}_t$ puts mass 1 on optimal state-action pairs (see [24], Theorem 3.6), implying that for a.e. α the policy must be optimal, as otherwise there exists a non-null set $\tilde{\mathcal{I}}_0 \subseteq \mathcal{I}$ such that for all $\alpha \in \tilde{\mathcal{I}}_0$, there is some suboptimality $\varepsilon > 0$, which directly contradicts the prequel.

For the remaining suboptimal $\alpha \in \mathcal{I}_0$ in the null set $\mathcal{I}_0 \subseteq \mathcal{I}$, we redefine $\boldsymbol{\pi}$ optimally for those α (always possible in our case, see e.g. [68]). This policy ensemble generates $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$ α-a.e. uniquely, and we need only consider its α-a.e. unique equivalence class for optimality, implying $\boldsymbol{\pi} \in \Phi(\boldsymbol{\mu})$. Furthermore, $\boldsymbol{\mu}$ is always measurable by definition, whereas $\boldsymbol{\pi}$ is measurable because $\tilde{\pi}_t$ is by definition a Markov kernel, and thus $\tilde{\pi}_t(u \mid \cdot, \cdot) = \tilde{\pi}_t(\{u\} \mid \cdot, \cdot)$ for Borel set $\{u\}$ is a measurable function, which implies measurability of $\tilde{\pi}_t(u \mid x, \cdot)$ (see e.g. [381], Appendix E). Therefore, we have proven existence of the GMFE $(\boldsymbol{\pi}, \boldsymbol{\mu})$. □

B.3 PROOF OF THEOREM 3.2.2

Proof. The proof is by induction as follows.

INITIAL CASE. For $t = 0$, we trivially have for all measurable functions $f\colon \mathcal{X} \times \mathcal{I} \to \mathbb{R}$ uniformly bounded by $|f| \leq M_f$ a LLN result

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_0^N(f) - \boldsymbol{\mu}_0(f)\right|\right] \\
&\quad= \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_0^{N,\alpha}(x)\, f(x,\alpha) - \sum_{x\in\mathcal{X}} \mu_0^{\alpha}(x)\, f(x,\alpha)\,\mathrm{d}\alpha\right|\right] \\
&\quad= \mathbb{E}\left[\left|\frac{1}{N}\sum_{i\in\mathcal{V}_N}\left(\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_0^i,\alpha)\,\mathrm{d}\alpha - \mathbb{E}\left[\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_0^i,\alpha)\,\mathrm{d}\alpha\right]\right)\right|\right] \\
&\quad\leq \left(\mathbb{E}\left[\left(\frac{1}{N}\sum_{i\in\mathcal{V}_N}\left(\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_0^i,\alpha)\,\mathrm{d}\alpha - \mathbb{E}\left[\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_0^i,\alpha)\,\mathrm{d}\alpha\right]\right)\right)^2\right]\right)^{\frac{1}{2}} \\
&\quad= \left(\frac{1}{N^2}\sum_{i\in\mathcal{V}_N}\mathbb{E}\left[\left(\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_0^i,\alpha)\,\mathrm{d}\alpha - \mathbb{E}\left[\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_0^i,\alpha)\,\mathrm{d}\alpha\right]\right)^2\right]\right)^{\frac{1}{2}} \leq \frac{2M_f}{\sqrt{N}}
\end{aligned}$$

by definition of $\boldsymbol{\mu}_t^N$, independence of $\{x_0^i\}_{i\in\mathcal{V}_N}$ and $x_0^i \sim \mu_0 = \mu_0^\alpha$ for all $i \in \mathcal{V}_N, \alpha \in \mathcal{I}$, where the second equality follows from Fubini's theorem.

INDUCTION STEP. Assume that the induction assumption holds at t. Then by definition of $\boldsymbol{\mu}_t^N$, for all bounded function $f : \mathcal{X} \times \mathcal{I} \to \mathbb{R}$ with $|f| \leq M_f$,

$$\begin{aligned}
\mathbb{E}\left[\left|\boldsymbol{\mu}_{t+1}^N(f) - \boldsymbol{\mu}_{t+1}(f)\right|\right] \leq\ & \mathbb{E}\left[\left|\boldsymbol{\mu}_{t+1}^N(f) - \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}^N}_{t,\boldsymbol{\mu}_t^N,W_N}(f)\right|\right] \\
&+ \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}^N}_{t,\boldsymbol{\mu}_t^N,W_N}(f) - \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}^N}_{t,\boldsymbol{\mu}_t^N,W}(f)\right|\right] \\
&+ \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}^N}_{t,\boldsymbol{\mu}_t^N,W}(f) - \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t^N,W}(f)\right|\right] \\
&+ \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t^N,W}(f) - \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t,W}(f)\right|\right] \\
&+ \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t,W}(f) - \boldsymbol{\mu}_{t+1}(f)\right|\right] .
\end{aligned}$$

FIRST TERM. We have by definition of $\boldsymbol{\mu}_t^N$

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_{t+1}^N(f) - \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}^N}_{t,\boldsymbol{\mu}_t^N,W_N}(f)\right|\right] \\
&\quad= \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_{t+1}^{N,\alpha}(x)\, f(x,\alpha)\,\mathrm{d}\alpha\right.\right. \\
&\quad\quad\left.\left. - \int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u\in\mathcal{U}} \pi_t^{N,\alpha}(u \mid x) \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W_N(\alpha,\beta)\mu_t^{N,\beta}\,\mathrm{d}\beta\right)\, f(x',\alpha)\,\mathrm{d}\alpha\right|\right] \\
&\quad= \mathbb{E}\left[\left|\frac{1}{N}\sum_{i\in\mathcal{V}_N}\left(\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_{t+1}^i,\alpha)\,\mathrm{d}\alpha - \mathbb{E}\left[\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_{t+1}^i,\alpha)\,\mathrm{d}\alpha \,\middle|\, \mathbf{x}_t\right]\right)\right|\right] \\
&\quad\leq \left(\mathbb{E}\left[\left(\frac{1}{N}\sum_{i\in\mathcal{V}_N}\left(\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_{t+1}^i,\alpha)\,\mathrm{d}\alpha - \mathbb{E}\left[\int_{(\frac{i-1}{N},\frac{i}{N}]} f(x_{t+1}^i,\alpha)\,\mathrm{d}\alpha \,\middle|\, \mathbf{x}_t\right]\right)\right)^2\right]\right)^{\frac{1}{2}}
\end{aligned}$$

$$= \left(\frac{1}{N^2} \sum_{i \in \mathcal{V}_N} \mathbb{E} \left[\left(\int_{(\frac{i-1}{N}, \frac{i}{N}]} f(x_{t+1}^i, \alpha) \, \mathrm{d}\alpha - \mathbb{E} \left[\int_{(\frac{i-1}{N}, \frac{i}{N}]} f(x_{t+1}^i, \alpha) \, \mathrm{d}\alpha \,\middle|\, \mathbf{x}_t \right] \right)^2 \right] \right)^{\frac{1}{2}}$$
$$\leq \frac{2M_f}{\sqrt{N}}$$

where the last equality follows from conditional independence of $\{x_{t+1}^i\}_{i \in \mathcal{V}_N}$ given $\mathbf{x}_t \equiv \{x_t^i\}_{i \in \mathcal{V}_N}$ and the law of total expectation.

SECOND TERM. We have

$$\mathbb{E}\left[\left| \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}^N}_{t, \boldsymbol{\mu}_t^N, W_N}(f) - \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}^N}_{t, \boldsymbol{\mu}_t^N, W}(f) \right| \right]$$
$$= \mathbb{E}\left[\left| \int_{\mathcal{I}} \sum_{x \in \mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u \in \mathcal{U}} \pi_t^{N,\alpha}(u \mid x) \sum_{x' \in \mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W_N(\alpha, \beta) \mu_t^{N,\beta} \, \mathrm{d}\beta \right) f(x', \alpha) \, \mathrm{d}\alpha \right.\right.$$
$$\left.\left. - \int_{\mathcal{I}} \sum_{x \in \mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u \in \mathcal{U}} \pi_t^{N,\alpha}(u \mid x) \sum_{x' \in \mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha, \beta) \mu_t^{N,\beta} \, \mathrm{d}\beta \right) f(x', \alpha) \, \mathrm{d}\alpha \right| \right]$$
$$\leq |\mathcal{X}| M_f L_p \mathbb{E}\left[\int_{\mathcal{I}} \left\| \int_{\mathcal{I}} W_N(\alpha, \beta) \mu_t^{N,\beta} \, \mathrm{d}\beta - \int_{\mathcal{I}} W(\alpha, \beta) \mu_t^{N,\beta} \, \mathrm{d}\beta \right\| \, \mathrm{d}\alpha \right]$$
$$\leq |\mathcal{X}|^2 M_f L_p \sup_{x \in \mathcal{X}} \mathbb{E}\left[\int_{\mathcal{I}} \left| \int_{\mathcal{I}} W_N(\alpha, \beta) \mu_t^{N,\beta}(x) - W(\alpha, \beta) \mu_t^{N,\beta}(x) \, \mathrm{d}\beta \right| \, \mathrm{d}\alpha \right] \to 0$$

by Assumption 3.2.1 and $\mu_t^{N,\beta}(x)$ trivially being bounded by 1. If the convergence in Assumption 3.2.1 is at rate $\mathcal{O}(1/\sqrt{N})$, then this convergence is also at rate $\mathcal{O}(1/\sqrt{N})$.

THIRD TERM. We have

$$\mathbb{E}\left[\left| \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}^N}_{t, \boldsymbol{\mu}_t^N, W}(f) - \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}}_{t, \boldsymbol{\mu}_t^N, W}(f) \right| \right]$$
$$= \mathbb{E}\left[\left| \int_{\mathcal{I}} \sum_{x \in \mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u \in \mathcal{U}} \pi_t^{N,\alpha}(u \mid x) \sum_{x' \in \mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha, \beta) \mu_t^{N,\beta} \, \mathrm{d}\beta \right) f(x', \alpha) \, \mathrm{d}\alpha \right.\right.$$
$$\left.\left. - \int_{\mathcal{I}} \sum_{x \in \mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u \in \mathcal{U}} \pi_t^{\alpha}(u \mid x) \sum_{x' \in \mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha, \beta) \mu_t^{N,\beta} \, \mathrm{d}\beta \right) f(x', \alpha) \, \mathrm{d}\alpha \right| \right]$$
$$\leq |\mathcal{X}||\mathcal{U}| M_f \mathbb{E}\left[\int_{\mathcal{I}} \left| \pi_t^{N,\alpha}(u \mid x) - \pi_t^{\alpha}(u \mid x) \right| \, \mathrm{d}\alpha \right]$$
$$= |\mathcal{X}||\mathcal{U}| M_f \mathbb{E}\left[\sum_{j \in \mathcal{V}_N \setminus \{i\}} \int_{(\frac{j-1}{N}, \frac{j}{N}]} \left| \pi_t^{\frac{\lceil N\alpha \rceil}{N}}(u \mid x) - \pi_t^{\alpha}(u \mid x) \right| \, \mathrm{d}\alpha \right]$$
$$+ |\mathcal{X}||\mathcal{U}| M_f \mathbb{E}\left[\int_{(\frac{i-1}{N}, \frac{i}{N}]} |\hat{\pi}_t(u \mid x) - \pi_t^{\alpha}(u \mid x)| \, \mathrm{d}\alpha \right]$$
$$\leq |\mathcal{X}||\mathcal{U}| M_f \cdot \frac{L_\pi}{N} + |\mathcal{X}||\mathcal{U}| M_f \cdot \frac{2|D_\pi|}{N} + |\mathcal{X}||\mathcal{U}| M_f \cdot \frac{2}{N}$$

by assumption of Lipschitz continuous π up to a finite number of discontinuities D_π as well as the deviating agent i's error term, for which the integrands are bounded by 2.

FOURTH TERM. We have

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t^N,W}(f) - \boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t,W}(f)\right|\right] \\
&\quad= \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{N,\beta}\,\mathrm{d}\beta\right) f(x',\alpha)\,\mathrm{d}\alpha\right.\right. \\
&\qquad\left.\left. - \int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}\,\mathrm{d}\beta\right) f(x',\alpha)\,\mathrm{d}\alpha\right|\right] \\
&\quad\le M_f|\mathcal{X}|\mathbb{E}\left[\sup_{x,u}\int_{\mathcal{I}}\left|p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{N,\beta}\,\mathrm{d}\beta\right)\right.\right. \\
&\qquad\qquad\left.\left. - p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}\,\mathrm{d}\beta\right)\right|\,\mathrm{d}\alpha\right] \\
&\quad\le M_f|\mathcal{X}|L_p \sum_{x'\in\mathcal{X}} \mathbb{E}\left[\int_{\mathcal{I}}\left|\int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{N,\beta}(x')\,\mathrm{d}\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}(x')\,\mathrm{d}\beta\right|\,\mathrm{d}\alpha\right] \\
&\quad\le M_f|\mathcal{X}|^2 L_p \cdot \frac{C'(1)}{\sqrt{N}}
\end{aligned}$$

in the case of rate $\mathcal{O}(1/\sqrt{N})$, or uniformly to zero otherwise, from Lipschitz P by defining the functions $f'_{x',\alpha}(x,\beta) = W(\alpha,\beta)\cdot\mathbf{1}_{x=x'}$ for any $(x',\alpha)\in\mathcal{X}\times\mathcal{I}$ and using the induction assumption on $f'_{x',\alpha}$ to obtain

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathcal{I}}\left|\int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{N,\beta}(x')\,\mathrm{d}\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}(x')\,\mathrm{d}\beta\right|\,\mathrm{d}\alpha\right] \\
&\quad= \int_{\mathcal{I}} \mathbb{E}\left[\left|\int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{N,\beta}(x')\,\mathrm{d}\beta - \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}(x')\,\mathrm{d}\beta\right|\right]\,\mathrm{d}\alpha \\
&\quad= \int_{\mathcal{I}} \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N(f'_{x',\alpha}) - \boldsymbol{\mu}_t(f'_{x',\alpha})\right|\right]\,\mathrm{d}\alpha \le \frac{C'(1)}{\sqrt{N}}
\end{aligned}$$

for some $C'(1) > 0$ uniformly over all f' bounded by 1 if the convergence in Assumption 3.2.1 is at rate $\mathcal{O}(1/\sqrt{N})$, or uniformly to zero otherwise.

FIFTH TERM. We have

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t,W}(f) - \boldsymbol{\mu}_t P^{\boldsymbol{\pi}}_{t,\boldsymbol{\mu}_t,W}(f)\right|\right] \\
&\quad= \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}\,\mathrm{d}\beta\right) f(x',\alpha)\,\mathrm{d}\alpha\right.\right. \\
&\qquad\left.\left. - \int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{\alpha}(x) \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}\,\mathrm{d}\beta\right) f(x',\alpha)\,\mathrm{d}\alpha\right|\right] \\
&\quad= \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{N,\alpha}(x) f'(x,\alpha)\,\mathrm{d}\alpha - \int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{\alpha}(x) f'(x,\alpha)\,\mathrm{d}\alpha\right|\right] \\
&\quad= \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N(f') - \boldsymbol{\mu}_t(f')\right|\right] \le \frac{C'(M_f)}{\sqrt{N}}\,.
\end{aligned}$$

in the case of rate $\mathcal{O}(1/\sqrt{N})$, or uniformly to zero otherwise, again by induction assumption applied to the function

$$f'(x,\alpha) = \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \int_{\mathcal{I}} W(\alpha,\beta)\mu_t^{\beta}\,\mathrm{d}\beta\right) f(x',\alpha)$$

bounded by M_f. This completes the proof by induction. □

B.4 PROOF OF LEMMA B.1.1

Proof. First, we will show that Eq. (B.1.5) implies Eq. (B.1.6).

PROOF OF (B.1.5) $\implies$ (B.1.6). We consider a uniformly Lipschitz, uniformly bounded family of measurable functions $\mathcal{H}$ from $\mathcal{X} \times \mathcal{B}_1(\mathcal{X})$ to $\mathbb{R}$. Let M_h be the uniform bound of functions in $\mathcal{H}$ and L_h be the uniform Lipschitz constant. Then, for arbitrary $h \in \mathcal{H}$ we have

$$\begin{aligned}
&\left|\mathbb{E}\left[h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N))\right] - \mathbb{E}\left[h(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| \\
&\quad = \left|\mathbb{E}\left[h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N))\right] - \mathbb{E}\left[h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| \\
&\qquad + \left|\mathbb{E}\left[h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] - \mathbb{E}\left[h(x_t^i, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| \\
&\qquad + \left|\mathbb{E}\left[h(x_t^i, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] - \mathbb{E}\left[h(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right|
\end{aligned}$$

which we will analyze in the following.

FIRST TERM. We have

$$\begin{aligned}
&\left|\mathbb{E}\left[h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N))\right] - \mathbb{E}\left[h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| \\
&\quad \le \mathbb{E}\left[\mathbb{E}\left[\left|h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N)) - h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right| \,\middle|\, x_t^i\right]\right] \\
&\quad \le L_h \mathbb{E}\left[\left\|\mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N) - \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t)\right\|\right] \\
&\quad = L_h \sum_{x \in \mathcal{X}} \mathbb{E}\left[\left|\int_{\mathcal{I}} W_N(\frac{i}{N}, \beta)\mu_t^{N,\beta}(x)\,\mathrm{d}\beta - \int_{\mathcal{I}} W_N(\frac{i}{N}, \beta)\mu_t^{\beta}(x)\,\mathrm{d}\beta\right|\right] \le \frac{C(1)}{\sqrt{N}}
\end{aligned}$$

by Theorem 3.2.2 applied to the functions $f'_{N,i,x}(x', \beta) = W_N(\frac{i}{N}, \beta) \cdot \mathbf{1}_{x=x'}$ uniformly bounded by 1.

SECOND TERM. Similarly, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[h(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] - \mathbb{E}\left[h(x_t^i, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| \\
&\quad \le L_h \|\mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t) - \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)\|_1 \\
&\quad \le L_h \sum_{x \in \mathcal{X}} \left|\int_{\mathcal{I}} \left(W_N(\frac{i}{N}, \beta) - W(\frac{i}{N}, \beta)\right) \mu_t^{\beta}(x)\,\mathrm{d}\beta\right| \\
&\quad \le L_h \sum_{x \in \mathcal{X}} \left|\int_{\mathcal{I}} \left(W_N(\frac{i}{N}, \beta) - N\int_{(\frac{i-1}{N}, \frac{i}{N}]} W(\alpha, \beta)\,\mathrm{d}\alpha\right) \mu_t^{\beta}(x)\,\mathrm{d}\beta\right|
\end{aligned}$$

$$+ L_h \sum_{x\in\mathcal{X}} \left| \int_{\mathcal{I}} \left(N \int_{(\frac{i-1}{N}, \frac{i}{N}]} W(\alpha,\beta)\,\mathrm{d}\alpha - W(\frac{i}{N},\beta) \right) \mu_t^\beta(x)\,\mathrm{d}\beta \right|$$

where the latter term can be bounded as

$$\begin{aligned}
&L_h \sum_{x\in\mathcal{X}} \left| \int_{\mathcal{I}} \left(N \int_{(\frac{i-1}{N}, \frac{i}{N}]} W(\alpha,\beta)\,\mathrm{d}\alpha - W(\frac{i}{N},\beta) \right) \mu_t^\beta(x)\,\mathrm{d}\beta \right| \\
&\leq L_h \sum_{x\in\mathcal{X}} \left| \int_{\mathcal{I}} N \int_{(\frac{i-1}{N}, \frac{i}{N}]} \left(W(\alpha,\beta) - W(\frac{\lceil N\alpha\rceil}{N},\beta) \right) \mu_t^\beta(x)\,\mathrm{d}\alpha\,\mathrm{d}\beta \right| \\
&\leq L_h |\mathcal{X}| N \cdot \frac{1}{N} \cdot \frac{L_W}{N} = \frac{L_W L_h |\mathcal{X}|}{N}
\end{aligned}$$

by Assumption 3.2.2.

Alternatively, if we assumed the weaker block-wise Lipschitz condition on W in Eq. (B.1.1), we can obtain the same result for almost all $i \in \mathcal{V}_N$, i.e. for any $\delta_0 > 0$ there exists $N' \in \mathbb{N}$ such that for any $N > N'$, there exists a set $\mathcal{J}_N^0$, $|\mathcal{J}_N^0| \geq \lfloor (1-\delta_0)N \rfloor$ such that for all $i \in \mathcal{J}_N^0$ the above is true: Since by Eq. (B.1.1) there exist only a finite number Q of intervals and therefore jumps, there can be only Q many i for which the above fails, while for all other i we again have

$$\begin{aligned}
&\left| \int_{\mathcal{I}} N \int_{(\frac{i-1}{N}, \frac{i}{N}]} \left(W(\alpha,\beta) - W(\frac{\lceil N\alpha\rceil}{N},\beta) \right) \mu_t^\beta(x)\,\mathrm{d}\alpha\,\mathrm{d}\beta \right| \\
&\leq \sum_{j\in\{1,\dots,Q\}} \left| \int_{\mathcal{I}_j} N \int_{(\frac{i-1}{N}, \frac{i}{N}]} \left(W(\alpha,\beta) - W(\frac{\lceil N\alpha\rceil}{N},\beta) \right) \mu_t^\beta(x)\,\mathrm{d}\alpha\,\mathrm{d}\beta \right| \\
&\leq N \cdot \frac{1}{N} \cdot \frac{L_W}{N} = \frac{L_W L_h |\mathcal{X}|}{N}
\end{aligned}$$

by Eq. (B.1.1), as $(\frac{i-1}{N}, \frac{i}{N}] \times \mathcal{I}_j \subseteq \mathcal{I}_k \times \mathcal{I}_j$ for some $k \in \{1,\dots,Q\}$.

For the former term we observe that

$$\begin{aligned}
&L_h \sum_{x\in\mathcal{X}} \left| \int_{\mathcal{I}} \left(W_N(\frac{i}{N},\beta) - N \int_{(\frac{i-1}{N}, \frac{i}{N}]} W(\alpha,\beta)\,\mathrm{d}\alpha \right) \mu_t^\beta(x)\,\mathrm{d}\beta \right| \\
&\leq L_h \sum_{x\in\mathcal{X}} N \int_{(\frac{i-1}{N}, \frac{i}{N}]} \left| \int_{\mathcal{I}} \left(W_N(\alpha,\beta) - W(\alpha,\beta) \right) \mu_t^\beta(x)\,\mathrm{d}\beta \right| \mathrm{d}\alpha
\end{aligned}$$

and by defining for any $x \in \mathcal{X}$ the terms $I_i^N(x)$ via

$$I_i^N(x) \coloneqq N \int_{(\frac{i-1}{N}, \frac{i}{N}]} \left| \int_{\mathcal{I}} \left(W_N(\alpha,\beta) - W(\alpha,\beta) \right) \mu_t^\beta(x)\,\mathrm{d}\beta \right| \mathrm{d}\alpha$$

and noticing that we have

$$\frac{1}{N} \sum_{i=1}^{N} I_i^N(x) = \int_{\mathcal{I}} \left| \int_{\mathcal{I}} \left(W_N(\alpha,\beta) - W(\alpha,\beta) \right) \mu_t^\beta(x)\,\mathrm{d}\beta \right| \mathrm{d}\alpha \to 0$$

by Assumption 3.2.1, we can conclude that for any $\varepsilon_1, \delta_1 > 0$ there exists $N' \in \mathbb{N}$ such that for any $N > N'$, there exists a set $\mathcal{J}_N^1$, $|\mathcal{J}_N^1| \geq \lfloor (1-\delta_1)N \rfloor$ such that for all $i \in \mathcal{J}_N^1$ we have

$$I_i^N(x) < \varepsilon_1,$$

since by the above we can choose $N' \in \mathbb{N}$ such that for any $N > N'$ we have $\frac{1}{N}\sum_{i=1}^{N} I_i^N(x) < \varepsilon_1\delta_1$, and from $I_i^N(x) \geq 0$ it would otherwise follow that $\frac{1}{N}\sum_{i=1}^{N} I_i^N(x) \geq \frac{1}{N} \cdot \lceil \delta_1 N \rceil \, \varepsilon_1 \geq \varepsilon_1\delta_1$ which would be a direct contradiction. Therefore, for all $i \in \mathcal{J}_N^1$, we have uniformly

$$L_h \sum_{x\in\mathcal{X}} N \int_{(\frac{i-1}{N}, \frac{i}{N}]} \left| \int_{\mathcal{I}} (W_N(\alpha,\beta) - W(\alpha,\beta))\, \mu_t^\beta(x)\, \mathrm{d}\beta \right| \mathrm{d}\alpha = L_h \sum_{x\in\mathcal{X}} I_i^N(x) \to 0\,.$$

THIRD TERM. By Eq. (B.1.5), for any $\varepsilon_2, \delta_2 > 0$ there exists a set $\mathcal{J}_N^2$, $|\mathcal{J}_N^2| \geq \lfloor (1-\delta_2)N \rfloor$ such that for all $i \in \mathcal{J}_N^2$ we have

$$\left| \mathbb{E}\left[h(x_t^i, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] - \mathbb{E}\left[h(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| < \varepsilon_2$$

independent of $\hat{\pi} \in \Pi$.

The intersection of $\mathcal{J}_N^0$, $\mathcal{J}_N^1$, $\mathcal{J}_N^2$ has at least $N - \lceil \delta_0 N \rceil - \lceil \delta_1 N \rceil - \lceil \delta_2 N \rceil$ agents fulfilling Eq. (B.1.6), which completes the proof of (B.1.5) $\implies$ (B.1.6) for almost all agents by choosing $\varepsilon_1, \varepsilon_2$ sufficiently small and $\delta_0, \delta_1, \delta_2 < \frac{\delta}{3}$ such that $N - \lceil \delta_0 N \rceil - \lceil \delta_1 N \rceil - \lceil \delta_2 N \rceil \geq \lfloor (1-\delta)N \rfloor$, which is equivalent to $1 - \frac{\lceil \delta_0 N \rceil}{N} - \frac{\lceil \delta_1 N \rceil}{N} - \frac{\lceil \delta_2 N \rceil}{N} \geq \frac{\lfloor (1-\delta)N \rfloor}{N}$ and is true for sufficiently large N, since in the limit, $1 - \frac{\lceil \delta_0 N \rceil}{N} - \frac{\lceil \delta_1 N \rceil}{N} - \frac{\lceil \delta_2 N \rceil}{N} \to 1 - \delta_0 - \delta_1 - \delta_2$ and $\frac{\lfloor (1-\delta)N \rfloor}{N} \to 1 - \delta$ as $N \to \infty$.

PROOF OF (B.1.5). All that remains is to show Eq. (B.1.5), which will automatically imply Eq. (B.1.6) at all times $t \in \mathcal{T}$ by the prequel. We will show Eq. (B.1.5) by induction.

INITIAL CASE. At $t = 0$, $\mathcal{L}(x_t^i) = \mu_0 = \mathcal{L}(\hat{x}_t^{\frac{i}{N}})$ by definition. Thus, trivially

$$\left| \mathbb{E}\left[g(x_0^i) \right] - \mathbb{E}\left[g(\hat{x}_0^{\frac{i}{N}}) \right] \right| = 0 < \varepsilon\,.$$

INDUCTION STEP. For any uniformly bounded family of functions $\mathcal{G}$ from $\mathcal{X}$ to $\mathbb{R}$ with bound M_g, we will show that for any $\varepsilon, \delta > 0$, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\left| \mathbb{E}\left[g(x_{t+1}^i) \right] - \mathbb{E}\left[g(\hat{x}_{t+1}^{\frac{i}{N}}) \right] \right| < \varepsilon$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$. Observe that

$$\left| \mathbb{E}\left[g(x_{t+1}^i) \right] - \mathbb{E}\left[g(\hat{x}_{t+1}^{\frac{i}{N}}) \right] \right| = \left| \mathbb{E}\left[l_{N,t}(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N)) \right] - \mathbb{E}\left[l_{N,t}(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right|$$

where we defined the uniformly bounded, uniformly Lipschitz functions

$$l_{N,t}(x,\nu) \equiv \sum_{u\in\mathcal{U}} \hat{\pi}_t(u \mid x) \sum_{x'\in\mathcal{X}} p(x' \mid x, u, \nu) g(x')$$

with Lipschitz constant $|\mathcal{X}| M_g L_p$ and uniform bound M_g. By the induction assumption and (B.1.5) $\implies$ (B.1.6) from the prequel, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\left| \mathbb{E}\left[l_{N,t}(x_t^i, \mathbb{G}_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N)) \right] - \mathbb{E}\left[l_{N,t}(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| < \varepsilon$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$, which completes the proof by induction. □

B.5 PROOF OF COROLLARY B.1.1

Proof. Define the uniformly bounded, uniformly Lipschitz functions

$$r_{\hat{\pi}}(x, \nu) \equiv \sum_{u \in \mathcal{U}} r(x, u, \nu) \hat{\pi}_t(u \mid x)$$

with Lipschitz constant $|U| L_r$ and uniform bound M_r such that by Lemma B.1.1 and Fubini's theorem, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\begin{aligned}
&\left| J_i^N(\pi^1, \dots, \pi^{i-1}, \hat{\pi}, \pi^{i+1}, \dots, \hat{\pi}) - J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\hat{\pi}) \right| \\
&\quad \leq \sum_{t=0}^{T-1} \left| \mathbb{E}\left[r_{\hat{\pi}_t}(x_t^i, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] - \mathbb{E}\left[r_{\hat{\pi}_t}(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| < \varepsilon \,.
\end{aligned}$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$ by choosing the maximum over all N' at each finite time step from Lemma B.1.1.

In case of the infinite horizon discounted objective, we instead first cut off at a time $T > \frac{\log \frac{\varepsilon(1-\gamma)}{4M_r}}{\log \gamma}$ such that trivially

$$\begin{aligned}
&\sum_{t=0}^{T-1} \gamma^t \left| \mathbb{E}\left[r_{\hat{\pi}_t}(x_t^i, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] - \mathbb{E}\left[r_{\hat{\pi}_t}(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| \\
&\qquad + \gamma^T \sum_{t=T}^{\infty} \gamma^{t-T} \left| \mathbb{E}\left[r_{\hat{\pi}_t}(x_t^i, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] - \mathbb{E}\left[r_{\hat{\pi}_t}(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| \\
&\quad < \sum_{t=0}^{T-1} \gamma^t \left| \mathbb{E}\left[r_{\hat{\pi}_t}(x_t^i, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] - \mathbb{E}\left[r_{\hat{\pi}_t}(\hat{x}_t^{\frac{i}{N}}, \mathbb{G}^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| + \frac{\varepsilon}{2}
\end{aligned}$$

and then handle the remaining term analogously to the finite horizon case. □

B.6 PROOF OF THEOREM 3.2.3

Proof. By Corollary B.1.1, for any $\varepsilon, \delta > 0$ there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\begin{aligned}
&\max_{\pi \in \Pi} \left(J_i^N(\pi^1, \dots, \pi^{i-1}, \pi, \pi^{i+1}, \dots, \pi^N) - J_i^N(\pi^1, \dots, \pi^N) \right) \\
&\quad \leq \max_{\pi \in \Pi} \left(J_i^N(\pi^1, \dots, \pi^{i-1}, \pi, \pi^{i+1}, \dots, \pi^N) - J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi) \right) \\
&\qquad + \max_{\pi \in \Pi} \left(J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi) - J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi^{\frac{i}{N}}) \right) \\
&\qquad + \left(J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi^{\frac{i}{N}}) - J_i^N(\pi^1, \dots, \pi^N) \right) \\
&\quad < \frac{\varepsilon}{2} + 0 + \frac{\varepsilon}{2} = \varepsilon
\end{aligned}$$

uniformly over $i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$, since $\pi^{\frac{i}{N}} \in \arg\max_\pi J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi)$ by definition of a GMFE. Reordering completes the proof. □

B.7 PROOF OF COROLLARY B.1.2

Proof. The proof follows immediately from Theorem 3.2.2 and Lemma B.1.1 by considering the trivial policy $\boldsymbol{\pi}$ that always chooses the only action available together with its generated MF $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$. □

B.8 PROOF OF PROPOSITION 3.2.2

Proof. The set $\boldsymbol{\Pi}$ is a complete metric space, since existence of limits follows from completeness of $\mathbb{R}$, pointwise limits of measurable functions are measurable, and policies will remain normalized. Banach's fixed point theorem applied to $\hat{\Phi} \circ \hat{\Psi}$ gives us the desired result. □

B.9 PROOF OF THEOREM 3.2.4

Proof. Formally, we approximate the MFs by $\hat{\Psi}(\boldsymbol{\pi}) = \sum_{i=1}^{M} \mathbf{1}_{\alpha \in \tilde{\mathcal{I}}_i} \hat{\mu}^{\alpha_i}$ for any fixed policy ensemble $\boldsymbol{\pi}$, and similarly policies $\hat{\Phi}(\boldsymbol{\mu}) = \sum_{i=1}^{M} \mathbf{1}_{\alpha \in \tilde{\mathcal{I}}_i} \pi^{\alpha_i}$ where π^{α_i} is the softmax policy of α_i for fixed $\boldsymbol{\mu}$, i.e.

$$\pi_t^{\alpha}(u \mid x) = \frac{\exp\left(\frac{Q_{\alpha}^{\boldsymbol{\mu}}(t,x,u)}{\eta}\right)}{\sum_{u \in \mathcal{U}} \exp\left(\frac{Q_{\alpha}^{\boldsymbol{\mu}}(t,x,u)}{\eta}\right)}. \tag{B.9.11}$$

By the Bellman equation (3.2.13), $Q_{\alpha}^{\boldsymbol{\mu}}(t,x,u)$ is Lipschitz in $\boldsymbol{\mu}$ for all $(t,x,u) \in \mathcal{T} \times \mathcal{X} \times \mathcal{U}$ under Assumption 3.2.2. Since the Lipschitz constants are shared over all α, by [9], Lemma B.7.5, Eq. (B.9.11) is therefore Lipschitz with Lipschitz constant proportional to $1/\eta$, which immediately implies that $\hat{\Phi}$ is also Lipschitz with Lipschitz constant c_1/η. By its recursive definition as compositions of Lipschitz functions, $\hat{\Psi}$ is Lipschitz as well with some constant c_2. Therefore, the composition of both functions $\hat{\Psi} \circ \hat{\Phi}$ is Lipschitz with constants $c_1 c_2/\eta$, which will be less than 1 for sufficiently large η. By Proposition 3.2.2, the equivalence classes algorithm $\hat{\Psi} \circ \hat{\Phi}$ converges to a fixed point. □

B.10 PROOF OF THEOREM 3.2.5

Proof. First, note that under the equivalence classes method, the distance between any α and its representant α_i uniformly shrinks to zero as $M \to \infty$, i.e. $\max_{i=1,\dots,M} \sup_{\alpha \in \tilde{\mathcal{I}}_i} |\alpha - \alpha_i| \to 0$.

We begin by showing that a solution of the M equivalence classes method $(\boldsymbol{\pi}, \boldsymbol{\mu}) \in \boldsymbol{\Pi} \times \boldsymbol{\mathcal{M}}$, $\boldsymbol{\pi} \in \hat{\Phi}(\boldsymbol{\mu})$, $\boldsymbol{\mu} = \hat{\Psi}(\boldsymbol{\pi})$ following Eq. (B.11.12), Eq. (B.11.13) fulfills approximate optimality, i.e. for any $\varepsilon > 0$ there exists M' s.t. for all $M > M'$

$$\sup_{\alpha \in \mathcal{I}} \max_{\pi \in \Pi} \left(J_{\alpha}^{\bar{\boldsymbol{\mu}}}(\pi) - J_{\alpha}^{\bar{\boldsymbol{\mu}}}(\pi^{\alpha}) \right) < \varepsilon,$$

where we introduced the true, exact MF ensemble $\bar{\boldsymbol{\mu}} = \Psi(\boldsymbol{\pi})$ following Eq. (3.2.12) generated by the block-wise solution policy $\sum_{i=1}^{M} \mathbf{1}_{\alpha \in \tilde{\mathcal{I}}_i} \pi^{\alpha_i}$ of the M equivalence classes method, as well as the true MF system under $\bar{\boldsymbol{\mu}}$ and any policy $\pi \in \Pi$

$$\bar{x}_0^{\alpha} \sim \mu_0(\bar{x}_0^{\alpha}), \quad \bar{u}_t^{\alpha} \sim \pi_t(\bar{u}_t^{\alpha} \mid \bar{x}_t^{\alpha}), \quad \bar{x}_{t+1}^{\alpha} \sim p(\bar{x}_{t+1}^{\alpha} \mid \bar{x}_t^{\alpha}, \bar{u}_t^{\alpha}, \bar{\mathbb{G}}_t^{\alpha}), \quad \forall (\alpha, t) \in \mathcal{I} \times \mathcal{T}$$

with $\mathcal{B}_1(\mathcal{X})$-valued $\bar{\mathbb{G}}_t^\alpha := \int_{\mathcal{I}} W(\alpha,\beta)\bar{\mu}_t^\beta \,\mathrm{d}\beta$ and $J_\alpha^{\bar{\boldsymbol{\mu}}}(\pi) \equiv \mathbb{E}\left[\sum_{t=0}^{T-1} r(\bar{x}_t^\alpha, \bar{u}_t^\alpha, \bar{\mathbb{G}}_t^\alpha)\right]$, while system (3.2.9) is to be understood as the system under the approximate MF ensemble $\boldsymbol{\mu}$.

To see this, we will analyze

$$\begin{aligned}
\sup_{\alpha\in\mathcal{I}} \max_{\pi\in\Pi} \left(J_\alpha^{\bar{\boldsymbol{\mu}}}(\pi) - J_\alpha^{\bar{\boldsymbol{\mu}}}(\pi^\alpha)\right) \leq{}& \max_{i=1,\dots,M} \sup_{\alpha\in\tilde{\mathcal{I}}_i} \max_{\pi\in\Pi} \left(J_\alpha^{\bar{\boldsymbol{\mu}}}(\pi) - J_\alpha^{\boldsymbol{\mu}}(\pi)\right) \\
&+ \max_{i=1,\dots,M} \sup_{\alpha\in\tilde{\mathcal{I}}_i} \max_{\pi\in\Pi} \left(J_\alpha^{\boldsymbol{\mu}}(\pi) - J_{\alpha_i}^{\boldsymbol{\mu}}(\pi)\right) \\
&+ \max_{i=1,\dots,M} \sup_{\alpha\in\tilde{\mathcal{I}}_i} \max_{\pi\in\Pi} \left(J_{\alpha_i}^{\boldsymbol{\mu}}(\pi) - J_{\alpha_i}^{\boldsymbol{\mu}}(\pi^{\alpha_i})\right) \\
&+ \max_{i=1,\dots,M} \sup_{\alpha\in\tilde{\mathcal{I}}_i} \left(J_{\alpha_i}^{\boldsymbol{\mu}}(\pi^{\alpha_i}) - J_\alpha^{\boldsymbol{\mu}}(\pi^{\alpha_i})\right) \\
&+ \max_{i=1,\dots,M} \sup_{\alpha\in\tilde{\mathcal{I}}_i} \left(J_\alpha^{\boldsymbol{\mu}}(\pi^\alpha) - J_\alpha^{\bar{\boldsymbol{\mu}}}(\pi^\alpha)\right) .
\end{aligned}$$

FIRST TERM. For any $\pi \in \Pi$, define the uniformly bounded, uniformly Lipschitz functions

$$r_\pi(x,\nu) \equiv \sum_{u\in\mathcal{U}} r(x,u,\nu)\pi_t(u \mid x)$$

with Lipschitz constant $|U|L_r$ and uniform bound M_r such that for the first term, we have

$$\begin{aligned}
&\left(J_\alpha^{\bar{\boldsymbol{\mu}}}(\pi) - J_\alpha^{\boldsymbol{\mu}}(\pi)\right) \leq \left|J_\alpha^{\bar{\boldsymbol{\mu}}}(\pi) - J_\alpha^{\boldsymbol{\mu}}(\pi)\right| \\
&\leq \sum_{t=0}^{T-1} \left|\mathbb{E}\left[r_\pi(\bar{x}_t^\alpha, \bar{\mathbb{G}}_t^\alpha)\right] - \mathbb{E}\left[r_\pi(x_t^\alpha, \mathbb{G}_t^\alpha\right]\right| \\
&\leq \sum_{t=0}^{T-1} \left|\mathbb{E}\left[r_\pi(\bar{x}_t^\alpha, \bar{\mathbb{G}}_t^\alpha) - r_\pi(\bar{x}_t^\alpha, \mathbb{G}_t^\alpha\right]\right| + \sum_{t=0}^{T-1} \left|\mathbb{E}\left[r_\pi(\bar{x}_t^\alpha, \mathbb{G}_t^\alpha\right] - \mathbb{E}\left[r_\pi(x_t^\alpha, \mathbb{G}_t^\alpha\right]\right| \\
&\leq \sum_{t=0}^{T-1} |U|L_r \left\|\int_{\mathcal{I}} W(\alpha,\beta)(\bar{\mu}_t^\beta - \mu_t^\beta)\,\mathrm{d}\beta)\right\| + \sum_{t=0}^{T-1} \left|\mathbb{E}\left[r_\pi(\bar{x}_t^\alpha, \mathbb{G}_t^\alpha\right] - \mathbb{E}\left[r_\pi(x_t^\alpha, \mathbb{G}_t^\alpha\right]\right| .
\end{aligned}$$

For the former term, note that

$$\begin{aligned}
\left\|\int_{\mathcal{I}} W(\alpha,\beta)(\bar{\mu}_t^\beta - \mu_t^\beta)\,\mathrm{d}\beta)\right\| &= \left\|\sum_i \int_{\tilde{\mathcal{I}}_i} W(\alpha,\beta)(\bar{\mu}_t^\beta - \mu_t^{\alpha_i})\,\mathrm{d}\beta)\right\| \\
&\leq \max_{i=1,\dots,M} \sup_{\alpha\in\tilde{\mathcal{I}}_i} |\mathcal{X}|\,\|\bar{\mu}_t^\alpha - \mu_t^{\alpha_i}\|
\end{aligned}$$

and we will show by induction over $t = 0, 1, \dots, T$ that $\sup_{\alpha\in\tilde{\mathcal{I}}_i} \|\bar{\mu}_t^\alpha - \mu_t^{\alpha_i}\| \to 0$ over all α uniformly over all equivalence classes $\tilde{\mathcal{I}}_i$. At $t = 0$, we have trivially $\bar{\mu}_0 = \mu_0$. Assume that $\sup_{\alpha\in\tilde{\mathcal{I}}_i} \|\bar{\mu}_t^\alpha - \mu_t^{\alpha_i}\| \to 0$. Then for $t+1$, we have

$$\begin{aligned}
&\sup_{\alpha\in\tilde{\mathcal{I}}_i} \left\|\bar{\mu}_{t+1}^\alpha - \mu_{t+1}^{\alpha_i}\right\| \\
&= \sup_{\alpha\in\tilde{\mathcal{I}}_i} \left\|\sum_{x\in\mathcal{X}} \bar{\mu}_t^\alpha(x) \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) p(\cdot \mid x, u, \bar{\mathbb{G}}_t^\alpha) - \sum_{x\in\mathcal{X}} \mu_t^{\alpha_i}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha_i}(u \mid x) p(\cdot \mid x, u, \mathbb{G}_t^{\alpha_i})\right\| \\
&\leq \sup_{\alpha\in\tilde{\mathcal{I}}_i} \left\|\sum_{x\in\mathcal{X}} \bar{\mu}_t^\alpha(x) \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) p(\cdot \mid x, u, \bar{\mathbb{G}}_t^\alpha) - \sum_{x\in\mathcal{X}} \bar{\mu}_t^{\alpha_i}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha_i}(u \mid x) p(\cdot \mid x, u, \bar{\mathbb{G}}_t^{\alpha_i})\right\|
\end{aligned}$$

$$
\begin{aligned}
&+ \left\| \sum_{x\in\mathcal{X}} \bar{\mu}_t^{\alpha_i}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha_i}(u \mid x) p(\cdot \mid x, u, \bar{\mathbb{G}}_t^{\alpha_i}) - \sum_{x\in\mathcal{X}} \mu_t^{\alpha_i}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha_i}(u \mid x) p(\cdot \mid x, u, \bar{\mathbb{G}}_t^{\alpha_i}) \right\| \\
&+ \left\| \sum_{x\in\mathcal{X}} \bar{\mu}_t^{\alpha_i}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha_i}(u \mid x) p(\cdot \mid x, u, \bar{\mathbb{G}}_t^{\alpha_i}) - \sum_{x\in\mathcal{X}} \mu_t^{\alpha_i}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha_i}(u \mid x) p(\cdot \mid x, u, \mathbb{G}_t^{\alpha_i}) \right\| \\
&\leq \sup_{\alpha\in\tilde{\mathcal{I}}_i} \left\| \sum_{x\in\mathcal{X}} \bar{\mu}_t^{\alpha}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha}(u \mid x) p(\cdot \mid x, u, \bar{\mathbb{G}}_t^{\alpha}) - \sum_{x\in\mathcal{X}} \bar{\mu}_t^{\alpha_i}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha_i}(u \mid x) p(\cdot \mid x, u, \bar{\mathbb{G}}_t^{\alpha_i}) \right\| \\
&\quad + |\mathcal{X}|^2 \left\| \bar{\mu}_t^{\alpha_i} - \mu_t^{\alpha_i} \right\| + |\mathcal{X}|^2 |\mathcal{U}| L_p \left\| \bar{\mu}_t^{\alpha_i} - \mu_t^{\alpha_i} \right\| \to 0
\end{aligned}
$$

as $M \to \infty$, since the first term is uniformly Lipschitz in α by Eq. (3.2.12) as a recursive composition, finite multiplication and addition of Lipschitz functions, whereas the other terms tend to zero by induction hypothesis. Since the Lipschitz constants do not depend on $\tilde{\mathcal{I}}_i$, the convergence is uniform.

To bound the latter term, we first note that $r_\pi(x, \mathbb{G}_t^\alpha)$ is always bounded by M_r regardless of t, x, α, π, i.e. it again suffices to show that for any family of functions $\mathcal{G}$ from $\mathcal{X}$ to $\mathbb{R}$ uniformly bounded by M_r, we have

$$
\sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[g(\bar{x}_t^\alpha) \right] - \mathbb{E}\left[g(x_t^\alpha) \right] \right| \to 0 \,.
$$

The proof is by induction. At $t = 0$, we trivially have $\mathcal{L}(\bar{x}_t^\alpha) = \mu_0 = \mathcal{L}(x_t^\alpha)$. Assuming that the induction hypothesis holds at t, then at $t + 1$ we have

$$
\sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[g(\bar{x}_{t+1}^\alpha) \right] - \mathbb{E}\left[g(x_{t+1}^\alpha) \right] \right| = \sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[l_t(\bar{x}_t^\alpha, \mathbb{G}_t^\alpha)) \right] - \mathbb{E}\left[l_t(x_t^\alpha, \mathbb{G}_t^\alpha) \right] \right| \to 0
$$

by the induction hypothesis, where we defined the uniformly bounded functions

$$
l_t(x, \nu) \equiv \sum_{u\in\mathcal{U}} \pi_t(u \mid x) \sum_{x'\in\mathcal{X}} p(x' \mid x, u, \nu) g(x')
$$

with uniform bound M_r. Therefore, $\left| J_\alpha^{\bar{\boldsymbol{\mu}}}(\pi) - J_\alpha^{\boldsymbol{\mu}}(\pi) \right| \to 0$ uniformly over all α, π.

SECOND TERM. For the second term, we analogously have

$$
\begin{aligned}
&\left(J_\alpha^{\boldsymbol{\mu}}(\pi) - J_{\alpha_i}^{\boldsymbol{\mu}}(\pi) \right) \leq \left| J_\alpha^{\boldsymbol{\mu}}(\pi) - J_{\alpha_i}^{\boldsymbol{\mu}}(\pi) \right| \\
&\leq \sum_{t=0}^{T-1} |U| L_r \left\| \int_{\mathcal{I}} (W(\alpha, \beta) - W(\alpha_i, \beta)) \mu_t^\beta \, \mathrm{d}\beta) \right\| + \sum_{t=0}^{T-1} \left| \mathbb{E}\left[r_\pi(x_t^\alpha, \mathbb{G}_t^{\alpha_i} \right] - \mathbb{E}\left[r_\pi(x_t^{\alpha_i}, \mathbb{G}_t^{\alpha_i} \right] \right|
\end{aligned}
$$

where the former term uniformly tends to zero as $M \to \infty$ over all α by Lipschitz W from Assumption 3.2.2 and increasingly fine partition intervals $\tilde{\mathcal{I}}_i$, while for the latter term we again show that for any family of functions $\mathcal{G}$ from $\mathcal{X}$ to $\mathbb{R}$ uniformly bounded by M_r, we have

$$
\sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[g(x_t^\alpha) \right] - \mathbb{E}\left[g(x_t^{\alpha_i}) \right] \right| \to 0 \,.
$$

The proof is by induction. At $t = 0$, we trivially have $\mathcal{L}(x_t^\alpha) = \mu_0 = \mathcal{L}(x_t^{\alpha_i})$. Assuming that the induction hypothesis holds at t, then at $t + 1$ we have

$$
\sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[g(x_{t+1}^\alpha) \right] - \mathbb{E}\left[g(x_{t+1}^{\alpha_i}) \right] \right| = \sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[l_t(x_t^\alpha, \mathbb{G}_t^{\alpha_i})) \right] - \mathbb{E}\left[l_t(x_t^{\alpha_i}, \mathbb{G}_t^{\alpha_i}) \right] \right| \to 0
$$

by the induction hypothesis, where we defined the uniformly bounded functions

$$
l_t(x, \nu) \equiv \sum_{u\in\mathcal{U}} \pi_t(u \mid x) \sum_{x'\in\mathcal{X}} p(x' \mid x, u, \nu) g(x')
$$

with uniform bound M_r. Therefore, $\left| J_\alpha^{\boldsymbol{\mu}}(\pi) - J_{\alpha_i}^{\boldsymbol{\mu}}(\pi) \right| \to 0$ uniformly over all α, π.

THIRD TERM. By definition, we have optimality of $\boldsymbol{\pi} \in \hat{\Phi}(\boldsymbol{\mu})$ under the approximate MF $\boldsymbol{\mu}$ at each representative α_i. Therefore, the term $\max_{\pi \in \Pi} \left(J^{\boldsymbol{\mu}}_{\alpha_i}(\pi) - J^{\boldsymbol{\mu}}_{\alpha_i}(\pi^{\alpha_i}) \right)$ is upper bounded by 0, as there is no policy π that improves over π^{α_i}.

FOURTH AND FIFTH TERM. The results follow from the first and second term by inserting π^{α} for π.

VARIATIONS ON THE SETTING. The infinite horizon discounted case is handled as in the proof of Corollary B.1.1, i.e. repeating the above up to some chosen time horizon T and trivially bounding all terms with $t \geq T$. The block-wise Lipschitz graphon case in Eq. (B.1.1) is handled by choosing the equivalence classes $\bar{\mathcal{I}}_i \subseteq \mathcal{I}_j$ such that they are part of at most one block $\mathcal{I}_j$ of the graphon.

PROOF OF THEOREM 3.2.5. Now fix any $\varepsilon, \delta > 0$. As a result of the prequel, we have that there exists M' s.t. for all $M > M'$

$$\sup_{\alpha \in \mathcal{I}} \max_{\pi \in \Pi} |J^{\boldsymbol{\mu}}_{\alpha}(\pi^{\alpha}) - J^{\boldsymbol{\mu}}_{\alpha}(\pi)| < \frac{\varepsilon}{3}.$$

Pick any such $M > M'$. By Corollary B.1.1 (for the first and third term, since $\boldsymbol{\pi}$ is constant with at most M discontinuities) and the prequel (for the second term), there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\begin{aligned}
&\max_{\pi \in \Pi} \left(J^N_i(\pi^1, \dots, \pi^{i-1}, \pi, \pi^{i+1}, \dots, \pi^N) - J^N_i(\pi^1, \dots, \pi^N) \right) \\
&\quad \leq \max_{\pi \in \Pi} \left(J^N_i(\pi^1, \dots, \pi^{i-1}, \pi, \pi^{i+1}, \dots, \pi^N) - J^{\bar{\boldsymbol{\mu}}}_{\frac{i}{N}}(\pi) \right) \\
&\qquad + \max_{\pi \in \Pi} \left(J^{\bar{\boldsymbol{\mu}}}_{\frac{i}{N}}(\pi) - J^{\bar{\boldsymbol{\mu}}}_{\frac{i}{N}}(\pi^{\frac{i}{N}}) \right) \\
&\qquad + \left(J^{\bar{\boldsymbol{\mu}}}_{\frac{i}{N}}(\pi^{\frac{i}{N}}) - J^N_i(\pi^1, \dots, \pi^N) \right) \\
&\quad < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon
\end{aligned}$$

which holds uniformly over $i \in \mathcal{J}_N$ for some $\mathcal{J}_N \subseteq \mathcal{V}_N$ with $|\mathcal{J}_N| \geq \lfloor (1-\delta)N \rfloor$, since $\pi^{\frac{i}{N}} \in \arg\max_{\pi} J^{\boldsymbol{\mu}}_{\frac{i}{N}}(\pi)$ by definition of a GMFE. Reordering completes the proof. □

B.11 EXPERIMENTAL DETAILS

In this section, we will give a full description of all the algorithms and hyperparameters we used during our experiments. For RL, we use PPO [73].

For the approximate equivalence classes, we shall consider grids $(\alpha_m \in [0,1])_{m=1,\dots,M}$ with associated policies $(\pi^{\alpha_m} \in \Pi)_{m=1,\dots,M}$ and MFs $(\mu^{\alpha_m} \in \mathcal{P}(\mathcal{X})^{\mathcal{T}})_{m=1,\dots,M}$. For the grid, we choose the points $\alpha_m = \frac{m}{100}$ with $m = 0, \dots, 100$. Here, an agent α shall use the policy π^{α_m} with the closest α_m.

To be precise, for the approximate MF $\boldsymbol{\mu} = \hat{\Psi}(\boldsymbol{\pi})$ we define $\mu^\alpha \equiv \hat{\mu}^{\alpha_m}$ for the α_m closest to α, i.e. formally, we thus have

$$\hat{\Psi}(\boldsymbol{\pi}) = \sum_{m=1}^{M} \mathbf{1}_{\alpha \in \tilde{\mathcal{I}}_m} \hat{\mu}^{\alpha_m} \tag{B.11.12}$$

for any fixed policy ensemble $\boldsymbol{\pi}$, with $\hat{\mu}$ defined through the recursive equation

$$\hat{\mu}_0^{\alpha_m} \equiv \mu_0, \quad \hat{\mu}_{t+1}^{\alpha_m}(x') \equiv \sum_{x \in \mathcal{X}} \hat{\mu}_t^{\alpha_m}(x) \sum_{u \in \mathcal{U}} \pi_t^{\alpha_m}(u \mid x) p(x' \mid x, u, \hat{\mathbb{G}}_t^{\alpha_m}), \quad m = 1, \dots, M \tag{B.11.13}$$

where under the assumption of equivalence classes $\tilde{\mathcal{I}}_m \equiv [a_m, b_m]$ of size $(b_m - a_m)$, we obtain neighborhood MF via

$$\hat{\mathbb{G}}_t^\alpha = \sum_{m=1}^{M} (b_m - a_m) W(\alpha, \alpha_m) \hat{\mu}_t^{\alpha_m} . \tag{B.11.14}$$

Note that in our algorithms, we shall assume equisized partitions and use $(b_m - a_m) = \frac{1}{M}$. Similarly, the policy ensemble is approximated by

$$\hat{\Phi}(\boldsymbol{\mu}) = \sum_{i=1}^{M} \mathbf{1}_{\alpha \in \tilde{\mathcal{I}}_i} \pi^{\alpha_i} \tag{B.11.15}$$

where π^{α_i} is the optimal policy of α_i for any fixed $\boldsymbol{\mu}$, i.e. the optimal policy and MF of each α is approximated by the optimal solution and MF of the closest α_i, which is an increasingly good approximation for sufficiently fine grids under the standing Lipschitz assumptions. In the case of block-wise Lipschitz continuous graphons via Eq. (B.1.1), a similar justification holds as long as each equivalence class remains constrained to one of the blocks of the graphon, see Theorem 3.2.5.

In Algorithm 10, the learning scheme is described on a high level. In our experiments, we either use approximate equivalence classes via Algorithms 11 and 12, or RL in the form of PPO together with sequential Monte Carlo in Algorithm 13, though in principle one can mix arbitrary methods.

Algorithm 10 FPI

1: Initialize $\boldsymbol{\mu}^0$ as the MF induced by the uniformly random policy $\mathbf{q}$.
2: **for** $k = 0, 1, \dots$ **do**
3: Compute $\boldsymbol{\pi}^k \in \boldsymbol{\Pi}$ either directly by PPO on Eq. (3.2.14), or by computing $\mathbf{Q}^{\boldsymbol{\mu}}$ via Algorithm 11 and using Eq. (B.13.19) to obtain a softmax policy.
4: Compute $\boldsymbol{\mu}^{k+1}$ induced by π^k using Algorithm 12, or for RL the neighborhood MFs $\mathbb{G}_t^\alpha$ directly using Algorithm 13.

We ran each trial of our experiments on a single conventional CPU core, with typical wall-clock times reaching up to at most a few days. We estimate the required compute to approximately 6500 core hours. We did not use any GPUs or TPUs. More specifically, the training of our approximate equivalence class approach took on average approximately 24 hours for 250 iterations in SIS and 50 iterations in Investment. As a result, Figure B.1 for the selection of appropriate temperatures took around 2500 core hours. The PPO experiments took approximately 3 days for each configuration, resulting in approximately 200 core hours for Figure B.3. Finally, for the N-agent evaluations in

Algorithm 11 Backwards induction

Input: Grid $(\alpha_m \in [0,1])_{m=1,\dots,M}$, MF $\boldsymbol{\mu} \in \boldsymbol{\mathcal{M}}$.
for $m = 1, \dots, M$ **do**
  Initialize terminal condition $Q^{\boldsymbol{\mu}}_{\alpha}(T, x, u) \equiv 0$ for all $(x, u) \in \mathcal{X} \times \mathcal{U}$.
  for $t = T-1, \dots, 0$ **do**
    for $(x, u) \in \mathcal{X} \times \mathcal{U}$ **do**
      $Q^{\boldsymbol{\mu}}_{\alpha_m}(t, x, u) \leftarrow r(x, u, \mathbb{G}_t^{\alpha_m}) + \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \mathbb{G}_t^{\alpha_m}) \max_{u'} Q^{\boldsymbol{\mu}}_{\alpha_m}(t+1, x', u')$.
Return $(Q^{\boldsymbol{\mu}}_{\alpha_m})_{m=1,\dots,M}$

Algorithm 12 Forward simulation

Input: Grid $(\alpha_m \in [0,1])_{m=1,\dots,M}$, policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$.
Initialize starting condition $\mu_0^{\alpha_m} \equiv \mu_0$ for all $m = 1, \dots, M$.
for $t = 0, \dots, T-2$ **do**
  for $m = 1, \dots, M$ **do**
    $\mu_{t+1}^{\alpha_m} \leftarrow \sum_{x \in \mathcal{X}} \mu_t^{\alpha_m}(x) \sum_{u \in \mathcal{U}} \pi_t^{\alpha_m}(u \mid x) p(x, u, \frac{1}{M} \sum_{n=1}^{M} W(\alpha_m, \alpha_n) \mu_t^{\alpha_n})$.
Return $(\mu^{\alpha_m})_{m=1,\dots,M}$

Figure 3.8, each run up to 100 agents takes up to 4 core hours. Adding on top of that around 250 core hours for the rest of the experiments results in a total of approximately 4000 core hours.

For PPO, we used the RLlib implementation by [76] (version 1.2.0, Apache-2.0 license). To allow for time-dependent policies, we append the current time to the network inputs. Further, discrete-valued observations are one-hot encoded. Any other parameter configurations are given in Algorithms 10, 11, 12 and 13, as well as in Table B.2.

As for the specific configurations used in the PPO experiments, we give the hyperparameters in Table B.1 and used with a feedforward neural network policy consisting of two hidden layers with 256 nodes and $\tanh$ activations, outputting a softmax policy over all actions.

B.12 PROBLEM DEFINITIONS

For each possible problem setting, we list the applied temperature setting in Table B.2. In the following, let $\mathbb{G} \in \mathcal{P}(\mathcal{X})$.

Algorithm 13 Sequential Monte Carlo

Input: Number of trajectories $K = 5$, number of particles $L = 200$, policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$.
for $k = 1, \dots, K$ **do**
  Initialize particles $\alpha_m \sim \mathrm{Unif}([0,1])$, $x_0^{m,k} \sim \mu_0$ for all $m = 1, \dots, L$.
  for $t = 1, \dots, T-1$ **do**
    for $m = 1, \dots, L$ **do**
      Sample action $u \sim \pi_t^{\alpha_m}(u \mid x_t^{m,k})$.
      Sample new particle state $x_{t+1}^{m,k} \sim p(x_{t+1}^{m,k} \mid x_t^{m,k}, u, \frac{1}{L} \sum_{n=1}^{L} W(\alpha_m, \alpha_n) \delta_{x_t^{n,k}})$.
return neighborhood MF $\mathbb{G}_t^{\alpha} \approx \frac{1}{K} \sum_{k=1}^{K} \frac{1}{L} \sum_{m=1}^{L} W(\alpha, \alpha_m) \delta_{x_t^{m,k}}$.

TABLE B.1: Hyperparameter configurations for PPO.

Symbol	Function	Value
l_r	Learning rate	0.00005
γ	Discount rate	1
λ	GAE lambda	0.99
c_{KL}	KL coefficient	0.2
β	KL target	0.006
c_{ent}	Entropy coefficient	0.01
ϵ	Clip parameter	0.2
B	Training batch size	4000
B_m	Mini batch size	128
I_{SGD}	SGD iterations per training batch	30

SIS-GRAPHON. In the SIS-Graphon game as described in the main text, we have $\mathcal{X} = \{S, I\}$, $\mathcal{U} = \{U, D\}$, $\mu_0(I) = 0.5$, $r(x, u, \mathbb{G}) = -2 \cdot \mathbf{1}_{\{I\}}(x) - 0.5 \cdot \mathbf{1}_{\{D\}}(u)$ and $\mathcal{T} = \{0, \dots, 49\}$. Similar parameters produce similar results, and we set the transition probabilities as

$$\begin{aligned}\mathbb{P}(S \mid I, \cdot, \cdot) &= 0.2,\\ \mathbb{P}(I \mid S, U, \mathbb{G}) &= 0.8 \cdot \mathbb{G}(I),\\ \mathbb{P}(I \mid S, D, \cdot) &= 0\,.\end{aligned}$$

INVESTMENT-GRAPHON. Similarly, in the Investment-Graphon game we have $\mathcal{X} = \{0, 1, \dots 9\}$, $\mathcal{U} = \{I, O\}$, $\mu_0(0) = 1$, $r(x, u, \mathbb{G}) = \frac{0.3x}{1+\sum_{x' \in \mathcal{X}} x' \mathbb{G}(x')} - 2 \cdot \mathbf{1}_{\{I\}}(u)$ and $\mathcal{T} = \{0, \dots, 49\}$. We set the transition probabilities for $x = 0, 1, \dots, 8$ as

$$\begin{aligned}\mathbb{P}(x+1 \mid x, I, \cdot) &= \frac{9-x}{10},\\ \mathbb{P}(x \mid x, I, \cdot) &= \frac{1+x}{10},\\ \mathbb{P}(x \mid x, O, \cdot) &= 1,\end{aligned}$$

while for $x = 9$ the next state is always $x = 9$.

B.13 EXPLOITABILITY AND TEMPERATURE CHOICE

In the following, we will explain our choice of temperatures in Table B.2 by approximately evaluating the average exploitability of GMFE candidates $(\boldsymbol{\pi}, \boldsymbol{\mu})$ – as it is intractable to approximately evaluate the maximum exploitability over all $\alpha \in \mathcal{I}$ – defined by

$$\Delta J(\boldsymbol{\pi}, \boldsymbol{\mu}) = \int_{\mathcal{I}} \sup_{\pi^* \in \Pi} J_\alpha^{\boldsymbol{\mu}}(\pi^*) - J_\alpha^{\boldsymbol{\mu}}(\pi^\alpha)\, \mathrm{d}\alpha\,. \tag{B.13.16}$$

More specifically, when using approximate equivalence classes, we compute the exploitability of some policy $\boldsymbol{\pi}$ by computing the optimal policy $\boldsymbol{\pi}^*$ obtained via Algorithm 11, under the fixed MF $\boldsymbol{\mu}$ generated by $\boldsymbol{\pi}$ via Algorithm 12, inserting $\pi^{*,\alpha}$ into Eq. (B.13.16) and then approximating by

$$\int_{\mathcal{I}} J_\alpha^{\boldsymbol{\mu}}(\pi)\, \mathrm{d}\alpha \approx \frac{1}{M} \sum_{m=1,\dots,M} \sum_{x \in \mathcal{X}} \mu_0(x) \sum_{u \in \mathcal{U}} \pi_0(u \mid x) Q_{\alpha_m}^{\boldsymbol{\mu},\boldsymbol{\pi}}(0, x, u)\,. \tag{B.13.17}$$

Here, we defined for any policy $\pi \in \Pi$ and $\alpha \in \mathcal{I}$ the policy evaluation functions $Q_\alpha^{\boldsymbol{\mu},\pi}$ as usual via

$$Q_\alpha^{\boldsymbol{\mu},\pi}(t,x,u) = r(x,u,\mathbb{G}_t^\alpha) + \sum_{x' \in \mathcal{X}} p(x' \mid x,u,\mathbb{G}_t^\alpha) \sum_{u' \in \mathcal{U}} \pi_0(u' \mid x) Q_\alpha^{\boldsymbol{\mu},\pi}(t+1,x',u') \tag{B.13.18}$$

with terminal condition $Q_\alpha^{\boldsymbol{\mu},\pi}(T,x,u) \equiv 0$, which can be computed as in Algorithm 11, see also [68] for a review.

To achieve convergence of FPI to approximate equilibria, for previous $\boldsymbol{\mu}^n \in \boldsymbol{\mathcal{M}}$ we compute the action value function $Q_\alpha^{\boldsymbol{\mu}^n}$ via Algorithm 11 using approximate equivalence classes and then define the next policy $\boldsymbol{\pi}^{n+1} = \hat{\Phi}(\boldsymbol{\mu}^n)$ for every $\alpha \in \mathcal{I}$ via the softmax function

$$\pi_t^{n+1,\alpha_i}(u \mid x) = \frac{\exp\left(\frac{Q_{\alpha_i}^{\boldsymbol{\mu}^n}(t,x,u)}{\eta}\right)}{\sum_{u \in \mathcal{U}} \exp\left(\frac{Q_{\alpha_i}^{\boldsymbol{\mu}^n}(t,x,u)}{\eta}\right)} \tag{B.13.19}$$

for the closest α_i with some temperature $\eta > 0$ chosen minimally for convergence.

For choosing the temperature, we evaluate the approximate final exploitability at various temperatures. The results can be seen in Figure B.1, where we plot the average, minimum and maximum exploitability over the last 10 iterations of the fixed point learning scheme. The reasoning behind choosing our temperatures as in Table B.2 is that we can see no fluctuations (indicating convergence of our learning scheme) together with a low approximate exploitability at the indicated temperatures.

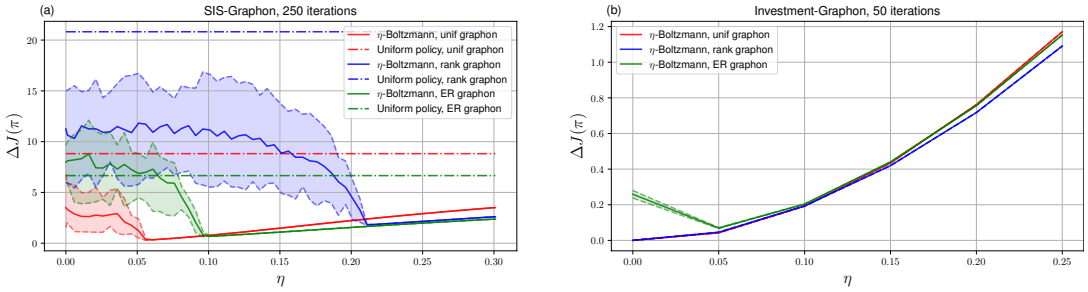


FIGURE B.1: Convergence in exploitability of GMFG algorithms. Final approximate exploitability mean and its minimum / maximum (shaded region) over the last 10 iterations for various temperatures η. We can see convergence for sufficiently high temperatures and choose the lowest temperature such that we still have convergence with low exploitability. Furthermore, compared to the uniformly random policy, our approximate exploitability is significantly lower, indicating a good approximate GMFE. For the Investment-Graphon problem, the approximate exploitability of the uniform policy is not shown, as it is above 30. **(a)**: SIS-Graphon; **(b)**: Investment-Graphon.

B.14 ADDITIONAL EXPERIMENTS

In Figure B.2, we plot investment behavior at quality $x = 0$ as well as expected quality for each α of the approximate equivalence class solution, and similarly in Figure B.3 for the PPO solution with sequential Monte Carlo. Here, for each α we averaged quality over all particles within a distance of 0.05 to α. We can see that PPO achieves qualitatively and quantitatively similar behavior, deviating slightly due to the approximate optimality of the PPO algorithm. To be precise, when evaluating

TABLE B.2: Temperature configurations.

Experiment	η for approximate equivalence classes
SIS-Graphon, W_{unif}	0.101
SIS-Graphon, W_{rank}	0.3
SIS-Graphon, W_{er}	0.101
Investment-Graphon, W_{unif}	0
Investment-Graphon, W_{rank}	0
Investment-Graphon, W_{er}	0.05

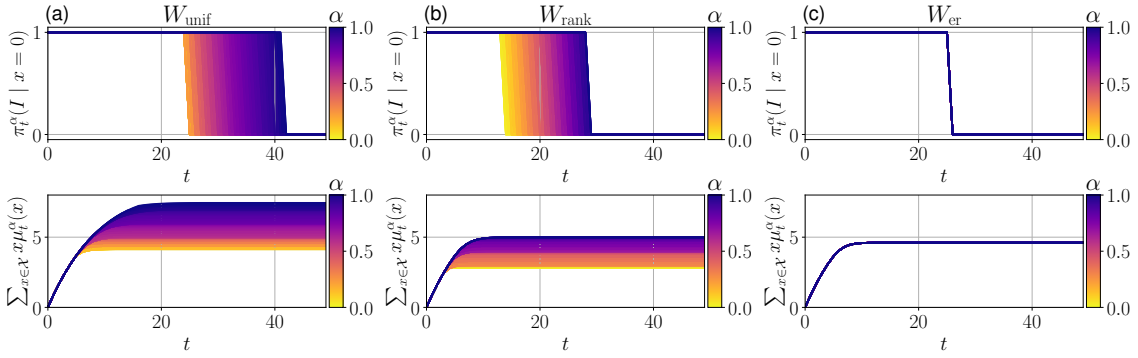


FIGURE B.2: Approximate equivalence classes solution of Investment-Graphon. We plot the probability of investing at state $x = 0$ (top) together with the evolution of average quality (bottom) for $M = 100$. **(a)**: Uniform attachment graphon; **(b)**: Ranked attachment graphon; **(c)**: ER graphon.

exploitability via either solution, we find that the learned policy exploitability remains around $\varepsilon \approx 2$, compared to $\varepsilon > 30$ for the uniform random policy.

In Figure B.4 the equilibrium behavior is shown for the Investment-Graphon problem without softmax policy regularization (except for the ER graphon case), as we find that the problem already converges to a very good equilibrium with low approximate exploitability, see Figure B.1. In this problem, we find that the resulting (deterministic without regularization) policy will let agents invest up to a certain quality, after which any further investment is avoided. The agents with higher connectivity will invest up to a lower quality, as they are in competition with more products.

In Figures B.5 and B.6, we have performed ablations over the number of equivalence classes for the SIS-Graphon problem. As can be observed, the solution obtained by approximate equivalence classes remains stable regardless of the particular number of equivalence classes, showing the stability of discretization approach and supporting Theorem 3.2.5.

Finally, in Figure B.7 we exemplarily show training results of applying state-of-the-art MARL methods such as MAPPO [88] on the finite-agent system with observed, randomized-per-episode graphon indices and W-random graphs. Here, we use the same hyperparameters as shown in Table B.1. As can be seen, due to the non-stationarity of the other agents, a naive application of MARL techniques fails to converge at all.

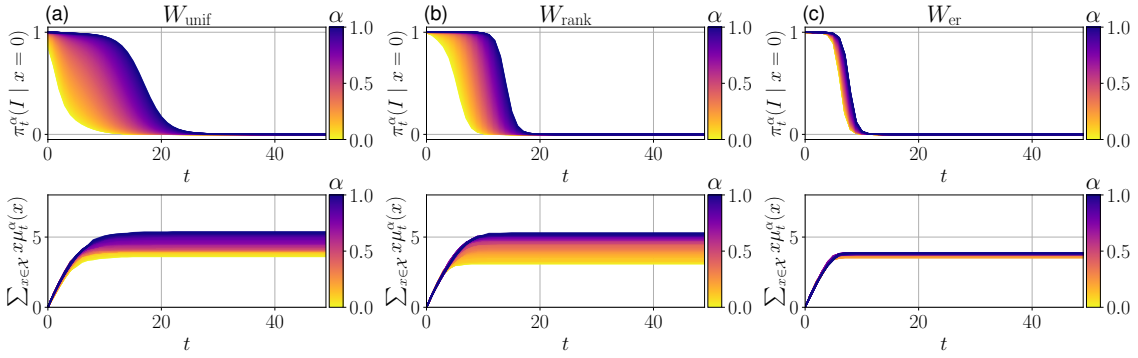


FIGURE B.3: Qualitative behavior of learned PPO equilibrium. The probability of investing at state $x = 0$ (top) together with the evolution of average quality (bottom) for PPO. The solution is similar to Figure B.2, though slightly different due to the approximations stemming from PPO and sequential Monte Carlo. **(a)**: Uniform attachment graphon; **(b)**: Ranked attachment graphon; **(c)**: ER graphon.

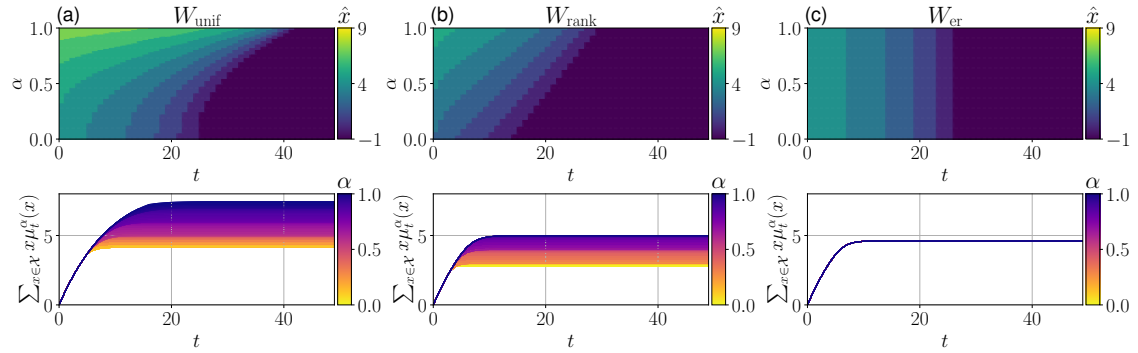


FIGURE B.4: Achieved equilibrium via $M = 100$ approximate equivalence classes in Investment-Graphon. Top: Maximum quality $\hat{x}$ up to which agents will invest ($\pi_t^\alpha(I \mid \hat{x}) > 0.5$), shown for each $\alpha \in \mathcal{I}, t \in \mathcal{T}$. Bottom: Expected quality versus time of each agent $\alpha \in \mathcal{I}$. It can be observed that agents with less connections (higher α) will invest more. **(a)**: Uniform attachment graphon; **(b)**: Ranked attachment graphon; **(c)**: ER graphon.

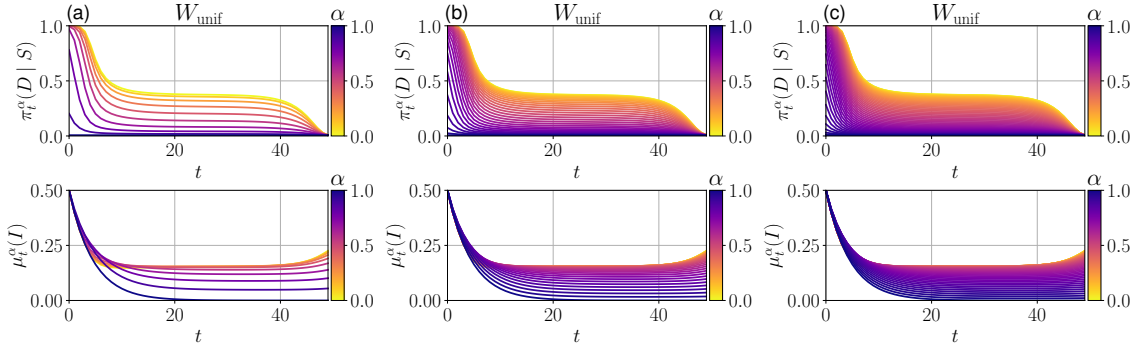


FIGURE B.5: Achieved equilibrium via approximate equivalence classes in SIS-Graphon for uniform attachment graphon, plotted for each representative $\alpha_i \in \mathcal{I}$. Top: Probability of taking precautions when healthy. Bottom: Probability of being infected. **(a)**: $M = 10$; **(b)**: $M = 30$; **(c)**: $M = 50$.

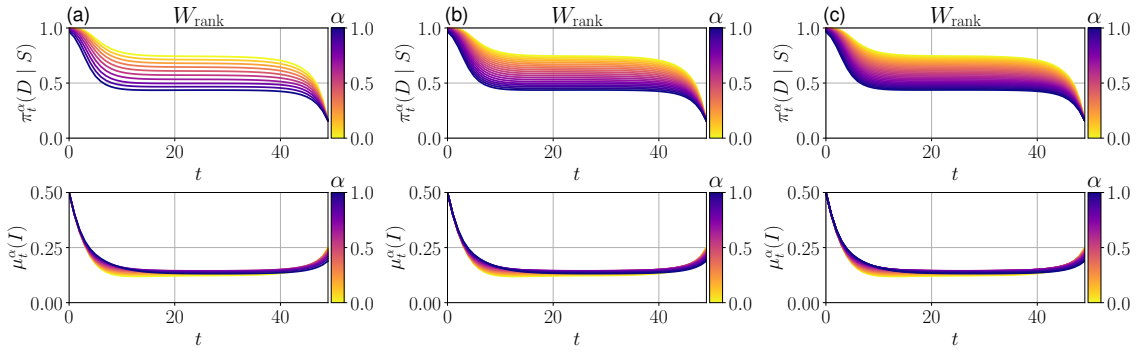


FIGURE B.6: Achieved equilibrium via approximate equivalence classes in SIS-Graphon for ranked attachment graphon, plotted for each representative $\alpha_i \in \mathcal{I}$. Top: Probability of taking precautions when healthy. Bottom: Probability of being infected. **(a)**: $M = 10$; **(b)**: $M = 20$; **(c)**: $M = 30$.

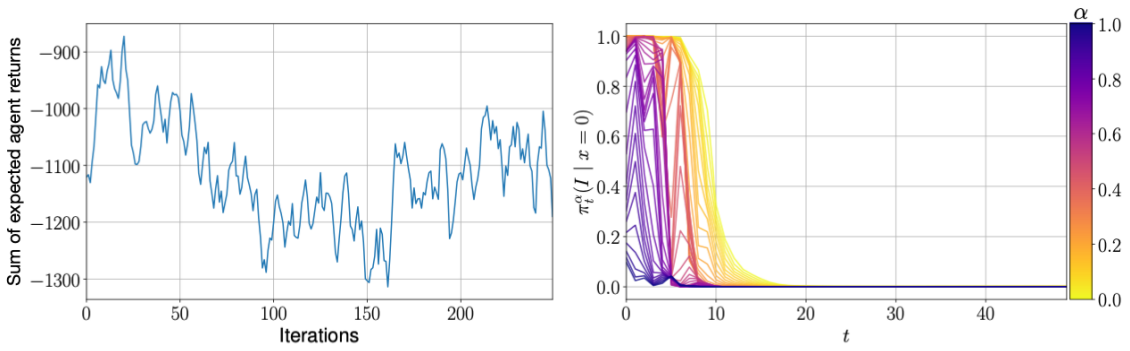


FIGURE B.7: Learning curve and results for direct application application of MAPPO [88]. Left: Sum of expected agent objectives over learning iterations; Right: Final policy probability of taking precautions when healthy.

C

APPENDIX C: SUPPLEMENTARY DETAILS ON SECTION 3.3

c.1 PROOF OF THEOREM 3.3.1

Proof. Under our assumptions, we can verify [24, Assumption 1] for the equivalent standard MFG given by Eq. (3.3.38) as in [7]. By [24, Theorem 3.3] there exists a MFE $(\tilde{\pi}, \tilde{\mu})$ for Eq. (3.3.38). The policy $\tilde{\mu}$ is α-a.e. optimal under the MF $\tilde{\mu}$ by [24, Theorem 3.6]. For all other α, there trivially exists an optimal action, i.e. we can change $\tilde{\pi}$ such that it is optimal for all α. Since the change is on a null set of $\mathcal{I}$, $(\tilde{\pi}, \tilde{\mu})$ remains a MFE. Define the hypergraphon MF policy $\boldsymbol{\pi}$ by $\pi_t^\alpha(u \mid x) = \tilde{\pi}_t(u \mid x, \alpha)$, then $\boldsymbol{\pi}$ is optimal under the hypergraphon MF $\boldsymbol{\mu}$ where $\boldsymbol{\mu} = \Psi(\boldsymbol{\pi})$, since $\mu_t^\alpha(x) = \tilde{\mu}_t(x, \alpha)$ for almost every α. Finally, both $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ are measurable. Therefore, we have proven existence of the HMFE $(\boldsymbol{\pi}, \boldsymbol{\mu})$. □

c.2 PROOF OF THEOREM 3.3.2

In this section, we provide the full proof of Theorem 3.3.2. In contrast to prior work such as [7], we (i) extend existing MF convergence results to n-fold products of the state distributions; and (ii) replace the state distributions by their symmetrized version, in order to obtain convergence results under the generalized cut norm in Eq. (3.3.20). Propagating these changes forward, the rest of the proof is (somewhat) readily generalized and given in the following.

To begin, we introduce some notation to improve readability. Define the D-dimensional neighborhood MF $\nu_W^{\alpha,\boldsymbol{\mu}}$ with d-th component

$$\nu_{W,d}^{\alpha,\boldsymbol{\mu}}(x) \coloneqq \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} W_d(\alpha, \boldsymbol{\beta}) \prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \, \mathrm{d}\boldsymbol{\beta}$$

for all $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{X})^{\mathcal{I}}$, $W \coloneqq (W_1, \dots, W_D) \in \bigtimes_{d=1}^{D} \mathcal{W}_{k_d}$ as well as the transition operator $P_W^{t,\boldsymbol{\pi},\boldsymbol{\mu}} \colon \mathcal{P}(\mathcal{X})^{\mathcal{I}} \to \mathcal{P}(\mathcal{X})^{\mathcal{I}}$ such that

$$\left(\boldsymbol{\mu}' P_W^{t,\boldsymbol{\pi},\boldsymbol{\mu}}\right)^{\alpha} = \sum_{x \in \mathcal{X}} \mu'^{\alpha}(x) \sum_{u \in \mathcal{U}} \pi^{\alpha}(u \mid x) p\left(\cdot \mid x, u, \nu_W^{\alpha,\boldsymbol{\mu}}\right)$$

for all $\mu' \in \mathcal{P}(\mathcal{X})^{\mathcal{I}}$, $\boldsymbol{\pi} \in \mathcal{P}(\mathcal{U})^{\mathcal{X}}$, such that e.g.

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t P_W^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t}.$$

Proof. In the following, consider arbitrary measurable functions $f \colon \mathcal{X} \times \mathcal{I} \to [-M_f, M_f]$, $M_f > 0$ and the telescoping sum

$$\begin{aligned}
&\mathbb{E}\left[\left|\bigotimes^{n} \boldsymbol{\mu}_t^N(f) - \bigotimes^{n} \boldsymbol{\mu}_t(f)\right|\right] \\
&\quad \leq \sum_{i=0}^{n-1} \mathbb{E}\left[\left|\bigotimes^{n-i} \boldsymbol{\mu}_t^N \otimes \bigotimes^{i} \boldsymbol{\mu}_t(f) - \bigotimes^{n-i-1} \boldsymbol{\mu}_t^N \otimes \bigotimes^{i+1} \boldsymbol{\mu}_t(f)\right|\right] \\
&\quad = \sum_{i=0}^{n-1} \mathbb{E}\left[\left|\bigotimes^{n-i-1} \boldsymbol{\mu}_t^N \otimes \left(\boldsymbol{\mu}_t^N - \boldsymbol{\mu}_t\right) \otimes \bigotimes^{i} \boldsymbol{\mu}_t(f)(f)\right|\right] \\
&\quad = \sum_{i=0}^{n-1} \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x_i \in \mathcal{X}} \int_{\mathcal{I}^{[n]\setminus\{i\}}} \sum_{\tilde{x} \in \mathcal{X}^{[n]\setminus\{i\}}} f(x, (\alpha, \boldsymbol{\beta})) \right.\right. \\
&\qquad\qquad \left.\left. \cdot \prod_{j=1}^{n-i-1} \mu^{N,\beta_j}(\tilde{x}_j) \prod_{j=n-i+1}^{n} \mu^{\beta_j}(\tilde{x}_j)\, \mathrm{d}\boldsymbol{\beta} \cdot \left[\mu^{N,\alpha}(x_i) - \mu^{\alpha}(x_i)\right]\, \mathrm{d}\alpha\right|\right] \\
&\quad = \sum_{i=0}^{n-1} \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x_i \in \mathcal{X}} g(x_i, \alpha) \left[\mu^{N,\alpha}(x_i) - \mu^{\alpha}(x_i)\right]\, \mathrm{d}\alpha\right|\right]
\end{aligned}$$

where we defined $g \colon \mathcal{X} \times \mathcal{I} \to [-M_f, M_f]$ as

$$g(x, \alpha) \coloneqq \int_{\mathcal{I}^{[n]\setminus\{i\}}} \sum_{x_{-i} \in \mathcal{X}^{[n]\setminus\{i\}}} f((x, x_{-i}), (\alpha, \boldsymbol{\beta})) \prod_{j=1}^{n-i-1} \mu^{N,\beta_j}(x_j) \prod_{j=n-i+1}^{n} \mu^{\beta_j}(x_j)\, \mathrm{d}\boldsymbol{\beta}.$$

Since g is a measurable function bounded by M_f, due to the prequel it suffices at any time $t \in \mathcal{T}$ to prove Eq. (3.3.42) for $n = 1$, which will imply the statement for all $n \in \mathbb{N}$.

The proof is by induction over t for $n = 1$. At $t = 0$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_0^N(f) - \boldsymbol{\mu}_0(f)\right|\right] \\
&\quad = \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x \in \mathcal{X}} \mu_0^{N,\alpha}(x)\, f(x, \alpha) - \sum_{x \in \mathcal{X}} \mu_0^{\alpha}(x)\, f(x, \alpha)\, \mathrm{d}\alpha\right|\right] \\
&\quad = \mathbb{E}\left[\left|\sum_{i \in [N]} \left(\int_{I_i^N} f(x_0^i, \alpha)\, \mathrm{d}\alpha - \mathbb{E}\left[\int_{I_i^N} f(x_0^i, \alpha)\, \mathrm{d}\alpha\right]\right)\right|\right]
\end{aligned}$$

$$\leq \left(\mathbb{V}\left[\sum_{i\in[N]} \int_{I_i^N} f(x_0^i, \alpha)\, \mathrm{d}\alpha \right] \right)^{\frac{1}{2}}$$
$$= \left(\sum_{i\in[N]} \mathbb{V}\left[\int_{I_i^N} f(x_0^i, \alpha)\, \mathrm{d}\alpha \right] \right)^{\frac{1}{2}} \leq \frac{4M_f}{\sqrt{N}}$$

by i.i.d. $x_0^i \sim \mu_0 = \mu_0^\alpha$ and $\mathbb{V}\left[\int_{I_i^N} f(x_0^i, \alpha)\, \mathrm{d}\alpha\right] \leq \left(\frac{4M_f}{N}\right)^2$.

Assume that Eq. (3.3.42) holds at $t \in \mathcal{T}$. Then at time $t+1$ we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_{t+1}^N(f) - \boldsymbol{\mu}_{t+1}(f)\right|\right] \\
&\quad\leq \mathbb{E}\left[\left|\boldsymbol{\mu}_{t+1}^N(f) - \boldsymbol{\mu}_t^N P_{W^N}^{t,\boldsymbol{\pi}_t^N,\boldsymbol{\mu}_t^N}(f)\right|\right] \\
&\qquad+ \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P_{W^N}^{t,\boldsymbol{\pi}_t^N,\boldsymbol{\mu}_t^N}(f) - \boldsymbol{\mu}_t^N P_{W}^{t,\boldsymbol{\pi}_t^N,\boldsymbol{\mu}_t^N}(f)\right|\right] \\
&\qquad+ \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P_{W}^{t,\boldsymbol{\pi}_t^N,\boldsymbol{\mu}_t^N}(f) - \boldsymbol{\mu}_t^N P_{W}^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t^N}(f)\right|\right] \\
&\qquad+ \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P_{W}^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t^N}(f) - \boldsymbol{\mu}_t^N P_{W}^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t}(f)\right|\right] \\
&\qquad+ \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P_{W}^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t}(f) - \boldsymbol{\mu}_{t+1}(f)\right|\right]
\end{aligned}$$

and in the following we will analyze each term.

For the first term, observe first that by definition,

$$\int_{\mathcal{I}^r<[k_d]\setminus\{1\}} W_d^N(\alpha, \boldsymbol{\beta}) \prod_{j=1}^{k_d-1} \mu_t^{N,\beta_j}\, \mathrm{d}\boldsymbol{\beta} = \frac{1}{N^{k_d-1}} \sum_{\mathbf{m}\in[N]^{k_d-1}} \mathbf{1}_{E^d[H_N]}(\mathbf{m}\cup i)\delta_{\times_{j\neq i} x_t^{m_j}}$$

and therefore

$$\boldsymbol{\mu}_t^N P_{W^N}^{t,\boldsymbol{\pi}_t^N,\boldsymbol{\mu}_t^N}(f) = \mathbb{E}\left[\int_{I_i^N} f(x_{t+1}^i, \alpha)\, \mathrm{d}\alpha \;\middle|\; \mathbf{x}_t\right]$$

such that we again obtain

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_{t+1}^N(f) - \boldsymbol{\mu}_t^N P_{W^N}^{t,\boldsymbol{\pi}_t^N,\boldsymbol{\mu}_t^N}(f)\right|\right] \\
&\quad= \mathbb{E}\left[\left|\sum_{i\in[N]} \left(g(x_{t+1}^i) - \mathbb{E}\left[g(x_{t+1}^i) \mid \mathbf{x}_t\right]\right)\right|\right] \\
&\quad\leq \left(\mathbb{E}\left[\left(\sum_{i\in[N]} \left(g(x_{t+1}^i) - \mathbb{E}\left[g(x_{t+1}^i) \mid \mathbf{x}_t\right]\right)\right)^2\right]\right)^{\frac{1}{2}} \\
&\quad= \left(\sum_{i\in[N]} \mathbb{E}\left[\left(g(x_{t+1}^i) - \mathbb{E}\left[g(x_{t+1}^i) \mid \mathbf{x}_t\right]\right)^2\right]\right)^{\frac{1}{2}} \leq \frac{4M_f}{\sqrt{N}}
\end{aligned}$$

where $g(x) \coloneqq \int_{I_i^N} f(x, \alpha)\, \mathrm{d}\alpha$, $|g| \leq \frac{M_f}{N}$, by using the law of total expectation and conditional independence of $\{x_{t+1}^i\}_{i\in[N]}$ given $\mathbf{x}_t \coloneqq \{x_t^i\}_{i\in[N]}$.

For the second term, first note that we can replace the distributional terms by their symmetrized version: For any $k \in \mathbb{N}$, any $W \in \mathrm{Sym}^{\mathrm{ind}}_{\leq}[k]$ and any step empirical measure or MF $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{X})^{\mathcal{I}}$, we have by symmetry that the associated neighborhood probabilities are invariant to all permutations $\sigma \in \mathrm{Sym}([k-1])$ of states $\mathcal{X}^{k-1}$, i.e. for any $x \in \mathcal{X}^{k-1}$, $\alpha \in \mathcal{I}$

$$\begin{aligned}
&\int_{\mathcal{I}^{r<[k]\setminus\{1\}}} W(\alpha,\beta) \prod_{i=1}^{k-1} \mu^{\beta_i}(x_i)) \,\mathrm{d}\beta \\
&\quad= \frac{1}{(k-1)!} \sum_{\sigma\in\mathrm{Sym}([k-1])} \int_{\mathcal{I}^{r<[k]\setminus\{1\}}} W(\alpha,\beta) \prod_{i=1}^{k-1} \mu^{\beta_i}(x_{\sigma(i)})) \,\mathrm{d}\beta \\
&\quad= \int_{\mathcal{I}^{r<[k]\setminus\{1\}}} W(\alpha,\beta) \underbrace{\frac{1}{(k-1)!} \sum_{\sigma\in\mathrm{Sym}([k-1])} \prod_{i=1}^{k-1} \mu^{\beta_i}(x_{\sigma(i)}))}_{u_1\in\mathrm{Sym}^{\mathrm{ind}}_{\leq}[k-1]} \,\mathrm{d}\beta
\end{aligned}$$

and hence Assumption 3.3.1 implies that

$$\begin{aligned}
&\int_{\mathcal{I}} \left| \int_{\mathcal{I}^{r<[k]\setminus\{1\}}} W(\alpha,\beta) u_1(\beta_{[k]\setminus\{1\}]}) \,\mathrm{d}\beta \right| \mathrm{d}\alpha \\
&\quad\leq \int_{\mathcal{I}^{[k]}} \left| \int_{\mathcal{I}^{r<[k]\setminus[k]}} W(\alpha,\beta) u_1(\alpha_{[k]\setminus\{1\}]}) \,\mathrm{d}\beta \right| \mathrm{d}\alpha \\
&\quad\leq \sup_{\substack{u_1 \colon \mathcal{I}^{r[k-1]}\to\mathcal{I}, \\ u_1\in\mathrm{Sym}^{\mathrm{ind}}_{\leq}[k-1]}} \int_{\mathcal{I}^{[k]}} \left| \int_{\mathcal{I}^{r<[k]\setminus[k]}} W(\alpha,\beta) \,\mathrm{d}\beta u_1(\alpha_{[k]\setminus\{1\}]}) \right| \mathrm{d}\alpha \\
&\quad= \sup_{\substack{u_1 \colon \mathcal{I}^{r[k-1]}\to\mathcal{I}, \\ u_1\in\mathrm{Sym}^{\mathrm{ind}}_{\leq}[k-1]}} \int_{\mathcal{I}^{[k]}} \int_{\mathcal{I}^{r<[k]\setminus[k]}} W(\alpha,\beta) \,\mathrm{d}\beta u_1(\alpha_{[k]\setminus\{1\}]}) \,\mathrm{d}\alpha \\
&\quad\leq \|W\|_{\square^{k-1}} \to 0
\end{aligned}$$

for any $x \in \mathcal{X}^{k-1}$, $\alpha \in \mathcal{I}$ by letting $u_1(\alpha_{r<([k]\setminus\{i\})}) \coloneqq \frac{1}{(k-1)!} \sum_{\sigma\in\mathrm{Sym}([k-1])} \prod_{i=1}^{k-1} \mu^{\beta_i}(x_{\sigma(i)}))$ in Eq. (3.3.20). Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left| \boldsymbol{\mu}^N_t P^{t,\boldsymbol{\pi}^N_t,\boldsymbol{\mu}^N_t}_{W^N}(f) - \boldsymbol{\mu}^N_t P^{t,\boldsymbol{\pi}^N t,\boldsymbol{\mu}^N_t}_{W}(f) \right| \right] \\
&\quad= \mathbb{E}\left[\left| \int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu^{N,\alpha}_t(x) \sum_{u\in\mathcal{U}} \pi^{N,\alpha}_t(u \mid x) \sum_{x'\in\mathcal{X}} f(x',\alpha) \right.\right. \\
&\qquad\qquad \left.\left. \cdot \left[p\left(x' \mid x, u, \nu^{\alpha,\boldsymbol{\mu}^N_t}_{W^N}\right) - p\left(x' \mid x, u, \nu^{\alpha,\boldsymbol{\mu}^N_t}_{W}\right) \right] \mathrm{d}\alpha \right| \right] \\
&\quad\leq |\mathcal{X}| M_f L_p \mathbb{E}\left[\int_{\mathcal{I}} \left\| \nu^{\alpha,\boldsymbol{\mu}^N_t}_{W^N} - \nu^{\alpha,\boldsymbol{\mu}^N_t}_{W} \right\| \mathrm{d}\alpha \right] \\
&\quad\leq |\mathcal{X}| M_f L_p \mathbb{E}\left[\int_{\mathcal{I}} \sum_{d\in[D]} \sum_{x\in\mathcal{X}^{k_d-1}} \left| \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} \left(\prod_{j=1}^{k_d-1} \mu^{N,\beta_j}(x_j) \right) \right.\right. \\
&\qquad\qquad \left.\left. \cdot \left[W^N_d(\alpha,\boldsymbol{\beta}) - W_d(\alpha,\boldsymbol{\beta}) \right] \mathrm{d}\boldsymbol{\beta} \right| \mathrm{d}\alpha \right] \\
&\quad\leq |\mathcal{X}| M_f L_p |\mathcal{X}| \sum_{d\in[D]} \|W^N_d - W_d\|_{\square^{k_d-1}} \to 0
\end{aligned}$$

by Assumption 3.3.1, and at rate $O(1/\sqrt{N})$ if Eq. (3.3.23) converges at rate $O(1/\sqrt{N})$.

For the third term, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P_W^{t,\boldsymbol{\pi}_t^N,\boldsymbol{\mu}_t^N}(f) - \boldsymbol{\mu}_t^N P_W^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t^N}(f)\right|\right] \\
&\quad = \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u\in\mathcal{U}} \left[\pi_t^{N,\alpha}(u \mid x) - \pi_t^{\alpha}(u \mid x)\right]\right.\right. \\
&\qquad\qquad \left.\left. \cdot \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \nu_W^{\alpha,\boldsymbol{\mu}_t^N}\right) f(x',\alpha)\,\mathrm{d}\alpha\right|\right] \\
&\quad \le M_f \mathbb{E}\left[\int_{\mathcal{I}} \left|\pi_t^{N,\alpha}(u \mid x) - \pi_t^{\alpha}(u \mid x)\right| \mathrm{d}\alpha\right] \\
&\quad = M_f \mathbb{E}\left[\sum_{j\in[N]\setminus\{i\}} \int_{I_j^N} \left|\pi_t^{\frac{\lceil N\alpha\rceil}{N}}(u \mid x) - \pi_t^{\alpha}(u \mid x)\right| \mathrm{d}\alpha\right] \\
&\qquad + M_f \mathbb{E}\left[\int_{I_i^N} |\hat{\pi}_t(u \mid x) - \pi_t^{\alpha}(u \mid x)| \,\mathrm{d}\alpha\right] \\
&\quad \le M_f \cdot \frac{L_\pi}{N} + M_f \cdot \frac{2|D_\pi|}{N} + M_f \cdot \frac{2}{N}
\end{aligned}$$

by $\boldsymbol{\pi} \in \boldsymbol{\Pi}_{\mathrm{Lip}}$ with Lipschitz constant L_π and up to D_π discontinuities, where we bound the integrands by 2.

For the fourth term, we find that

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P_W^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t^N}(f) - \boldsymbol{\mu}_t^N P_W^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t}(f)\right|\right] \\
&\quad = \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \mu_t^{N,\alpha}(x) \sum_{u\in\mathcal{U}} \pi_t^{\alpha}(u \mid x) \sum_{x'\in\mathcal{X}} f(x',\alpha)\right.\right. \\
&\qquad\qquad \left.\left. \cdot \left[p\left(x' \mid x, u, \nu_W^{\alpha,\boldsymbol{\mu}_t^N}\right) - p\left(x' \mid x, u, \nu_W^{\alpha,\boldsymbol{\mu}_t}\right)\right] \mathrm{d}\alpha\right|\right] \\
&\quad \le |\mathcal{X}| M_f L_p \mathbb{E}\left[\int_{\mathcal{I}} \left\|\nu_W^{\alpha,\boldsymbol{\mu}_t^N} - \nu_W^{\alpha,\boldsymbol{\mu}_t}\right\| \mathrm{d}\alpha\right] \\
&\quad \le |\mathcal{X}| M_f L_p \mathbb{E}\left[\int_{\mathcal{I}} \sum_{d\in[D]} \sum_{x\in\mathcal{X}^{k_d-1}} \left|\int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} W_d(\alpha,\boldsymbol{\beta})\right.\right. \\
&\qquad\qquad \left.\left. \cdot \left[\prod_{j=1}^{k_d-1} \mu^{N,\beta_j}(x_j) - \prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j)\right] \mathrm{d}\boldsymbol{\beta}\right| \mathrm{d}\alpha\right] \\
&\quad = |\mathcal{X}| M_f L_p \int_{\mathcal{I}} \sum_{d\in[D]} \sum_{x\in\mathcal{X}^{k_d-1}} \mathbb{E}\left[\left|\int_{\mathcal{I}^{[k_d]\setminus\{1\}}} \int W_d(\alpha,\boldsymbol{\beta},\boldsymbol{\zeta})\,\mathrm{d}\boldsymbol{\zeta}\right.\right. \\
&\qquad\qquad \left.\left. \cdot \left[\prod_{j=1}^{k_d-1} \mu^{N,\beta_j}(x_j) - \prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j)\right] \mathrm{d}\boldsymbol{\beta}\right| \mathrm{d}\alpha\right] \\
&\quad = |\mathcal{X}| M_f L_p \int_{\mathcal{I}} \sum_{d\in[D]} \sum_{x\in\mathcal{X}^{k_d-1}} \mathbb{E}\left[\left|\bigotimes^{k_d-1} \boldsymbol{\mu}_t^N(f'_{x,\alpha}) - \bigotimes^{k_d-1} \boldsymbol{\mu}_t(f'_{x,\alpha})\right|\right] \mathrm{d}\alpha \to 0
\end{aligned}$$

at the rate in the induction assumption, by applying the induction assumption Eq. (3.3.42) for $n = k_d - 1$ to the functions

$$f'_{x,\alpha}(x', \beta) = \int_{\mathcal{I}^{r<[k_d]\setminus[k_d]}} W_d(\alpha, \boldsymbol{\beta}, \boldsymbol{\zeta}) \, \mathrm{d}\boldsymbol{\zeta} \cdot \mathbf{1}_{\{x\}}(x')$$

for any $(x, \alpha) \in \mathcal{X}^{k_d-1} \times \mathcal{I}$.

For the fifth term, we analogously obtain

$$\begin{aligned}
&\mathbb{E}\left[\left|\boldsymbol{\mu}_t^N P_W^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t}(f) - \boldsymbol{\mu}_t P_W^{t,\boldsymbol{\pi}_t,\boldsymbol{\mu}_t}(f)\right|\right] \\
&\quad = \mathbb{E}\left[\left|\int_{\mathcal{I}} \sum_{x\in\mathcal{X}} \left[\mu_t^{N,\alpha}(x) - \mu_t^\alpha(x)\right] \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) \cdot \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \nu_W^{\alpha,\boldsymbol{\mu}_t}\right) f(x', \alpha) \, \mathrm{d}\alpha\right|\right] \\
&\quad = \mathbb{E}\left[\left|\boldsymbol{\mu}_t^N(f') - \boldsymbol{\mu}_t(f')\right|\right] \to 0.
\end{aligned}$$

at the rate in the induction assumption, by applying the induction assumption Eq. (3.3.42) to

$$f'(x, \alpha) = \sum_{u\in\mathcal{U}} \pi_t^\alpha(u \mid x) \sum_{x'\in\mathcal{X}} p\left(x' \mid x, u, \nu_W^{\alpha,\boldsymbol{\mu}_t}\right) f(x', \alpha).$$

This concludes the proof by induction. □

C.3 PROOF OF THEOREM 3.3.3

The proof of Theorem 3.3.3 mirrors the proof in [7] apart from propagating the multidimensional convergence results forward, and we give the entire proof for completeness and convenience. Again, we introduce some notation to improve readability. For any $\alpha \in \mathcal{I}$, $d \in [D]$, define maps $\nu_d^\alpha \colon \mathcal{P}(\mathcal{X})^{\mathcal{I}} \to \mathcal{P}(\mathcal{X})$ and $\nu_{N,d}^\alpha \colon \mathcal{P}(\mathcal{X})^{\mathcal{I}} \to \mathcal{P}(\mathcal{X})$ as

$$\begin{aligned}
\nu_d^\alpha(\boldsymbol{\mu})(x) &\coloneqq \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} W_d(\alpha, \boldsymbol{\beta}) \prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \, \mathrm{d}\boldsymbol{\beta}, \\
\nu_{N,d}^\alpha(\boldsymbol{\mu})(x) &\coloneqq \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} W_d^N(\alpha, \boldsymbol{\beta}) \prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \, \mathrm{d}\boldsymbol{\beta}
\end{aligned}$$

with D-dimensional shorthands

$$\begin{aligned}
\nu^\alpha(\boldsymbol{\mu}) &\coloneqq (\nu_d^\alpha(\boldsymbol{\mu}))_{d\in[D]}, \\
\nu_N^\alpha(\boldsymbol{\mu}) &\coloneqq (\nu_{N,d}^\alpha(\boldsymbol{\mu}))_{d\in[D]}
\end{aligned}$$

such that by definition $\nu_t^\alpha = \nu^\alpha(\boldsymbol{\mu}_t)$ and $\nu_t^{N,i} = \nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N)$.

Proof. To begin, we prove (3.3.45) $\implies$ (3.3.46) at any fixed time t. Define the uniform bound M_h and uniform Lipschitz constant L_h of functions in $\mathcal{H}$. For any $h \in \mathcal{H}$ we have

$$\begin{aligned}
&\left|\mathbb{E}\left[h(x_t^i, \nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N))\right] - \mathbb{E}\left[h(\hat{x}_t^{\frac{i}{N}}, \nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| \\
&\quad = \left|\mathbb{E}\left[h(x_t^i, \nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N))\right] - \mathbb{E}\left[h(x_t^i, \nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right|
\end{aligned}$$

$$+\left|\mathbb{E}\left[h(x_t^i,\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]-\mathbb{E}\left[h(x_t^i,\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right|$$
$$+\left|\mathbb{E}\left[h(x_t^i,\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]-\mathbb{E}\left[h(\hat{x}_t^{\frac{i}{N}},\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right|$$

which we will analyze as $N\to\infty$.

For the first term, we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[h(x_t^i,\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N))\right]-\mathbb{E}\left[h(x_t^i,\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right|\\
&\quad\le\mathbb{E}\left[\mathbb{E}\left[\left|h(x_t^i,\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N))-h(x_t^i,\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right|\,\middle|\,x_t^i\right]\right]\\
&\quad\le L_h\mathbb{E}\left[\left\|\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N)-\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t)\right\|\right]\\
&\quad= L_h\mathbb{E}\left[\sum_{d\in[D]}\sum_{x\in\mathcal{X}^{k_d-1}}\left|\int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}}W_d^N(\alpha,\boldsymbol{\beta})\right.\right.\\
&\qquad\left.\left.\cdot\left[\prod_{j=1}^{k_d-1}\mu^{N,\beta_j}(x_j)-\prod_{j=1}^{k_d-1}\mu^{\beta_j}(x_j)\right]\,\mathrm{d}\beta\right|\right]\to 0
\end{aligned}$$

uniformly by applying Theorem 3.3.2 to the functions

$$f'_{N,i,x}(x',\beta)=\int_{\mathcal{I}^{r<[k_d]\setminus[k_d]}}W_d^N(\frac{i}{N},\boldsymbol{\beta},\boldsymbol{\zeta})\,\mathrm{d}\boldsymbol{\zeta}\cdot\mathbf{1}_{\{x\}}(x').$$

For the second term, we analogously have

$$\begin{aligned}
&\left|\mathbb{E}\left[h(x_t^i,\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]-\mathbb{E}\left[h(x_t^i,\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right|\\
&\quad\le L_h\|\nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t)-\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t)\|_1\\
&\quad\le L_h\sum_{d\in[D]}\sum_{x\in\mathcal{X}^{k_d-1}}\left|\int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}}\left(\prod_{j=1}^{k_d-1}\mu^{\beta_j}(x_j)\right)\right.\\
&\qquad\left.\cdot\left[W_d^N(\frac{i}{N},\boldsymbol{\beta})-W_d(\frac{i}{N},\boldsymbol{\beta})\right]\,\mathrm{d}\beta\right|\\
&\quad\le L_h\sum_{d\in[D]}\sum_{x\in\mathcal{X}^{k_d-1}}\left|\int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}}\left(\prod_{j=1}^{k_d-1}\mu^{\beta_j}(x_j)\right)\right.\\
&\qquad\left.\cdot\left[W_d^N(\frac{i}{N},\boldsymbol{\beta})-N\int_{I_i^N}W_d(\alpha,\beta)\,\mathrm{d}\alpha\right]\,\mathrm{d}\beta\right|\\
&\qquad+L_h\sum_{d\in[D]}\sum_{x\in\mathcal{X}^{k_d-1}}\left|\int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}}\left(\prod_{j=1}^{k_d-1}\mu^{\beta_j}(x_j)\right)\right.\\
&\qquad\left.\cdot\left[N\int_{I_i^N}W_d(\alpha,\beta)\,\mathrm{d}\alpha-W_d(\frac{i}{N},\boldsymbol{\beta})\right]\,\mathrm{d}\beta\right|
\end{aligned}$$

where for the former (finite) sum we have

$$\begin{aligned}
&\left| \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} \left(\prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \right) \right. \\
&\qquad \left. \cdot \left[W_d^N(\frac{i}{N}, \boldsymbol{\beta}) - N \int_{I_i^N} W_d(\alpha, \beta) \, \mathrm{d}\alpha \right] \, \mathrm{d}\beta \right| \\
&= \left| N \int_{I_i^N} \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} \left(\prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \right) \right. \\
&\qquad \left. \cdot \left[W_d^N(\alpha, \boldsymbol{\beta}) - W_d(\alpha, \beta) \right] \, \mathrm{d}\beta \, \mathrm{d}\alpha \right| \\
&\leq N \int_{I_i^N} \left| \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} \left(\prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \right) \right. \\
&\qquad \left. \cdot \left[W_d^N(\alpha, \boldsymbol{\beta}) - W_d(\alpha, \beta) \right] \, \mathrm{d}\beta \right| \, \mathrm{d}\alpha \\
&=: I_i^N
\end{aligned}$$

since by definition of the step-hypergraphon, $W_d^N(\frac{i}{N}, \boldsymbol{\beta}) = W_d^N(\alpha, \boldsymbol{\beta})$ over $\alpha \in I_i^N$. Therefore,

$$\frac{1}{N} \sum_{i=1}^{N} I_i^N = \int_{\mathcal{I}^{r<[k_d]}} \left[W_d^N(\beta) - W_d(\beta) \right] \prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \, \mathrm{d}\beta \to 0$$

as in the proof of Theorem 3.3.2 by Assumption 3.3.1. Fix $\varepsilon, \delta > 0$. As N becomes sufficiently large, there must exist $\mathcal{J}_1^N$, $|\mathcal{J}_1^N| \geq \lfloor (1-\delta)N \rfloor$ such that

$$I_i^N < \varepsilon, \quad \forall i \in \mathcal{J}_1^N.$$

We prove this by contradiction: Assume there does not exist such $\mathcal{J}_1^N$, then there exist at least $\lceil pN \rceil$ agents where $I_i^N \geq \varepsilon$. Since $I_i^N \geq 0$, it follows that $\frac{1}{N} \sum_{i=1}^{N} I_i^N \geq \frac{1}{N} \lceil \delta N \rceil \, \varepsilon \geq \varepsilon\delta$, which contradicts the convergence to zero of $\frac{1}{N} \sum_{i=1}^{N} I_i^N$. Repeating the argument for each $d \in [D]$, $x \in \mathcal{X}^{k_d-1}$ bounds the first sum.

For the latter (finite) sum, we have

$$\begin{aligned}
&\left| \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} \left(\prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \right) \right. \\
&\qquad \left. \cdot \left[N \int_{I_i^N} W_d(\alpha, \beta) \, \mathrm{d}\alpha - W_d(\frac{i}{N}, \boldsymbol{\beta}) \right] \, \mathrm{d}\beta \right| \\
&= \left| N \int_{I_i^N} \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} \left(\prod_{j=1}^{k_d-1} \mu^{\beta_j}(x_j) \right) \right. \\
&\qquad \left. \cdot \left[W_d(\alpha, \boldsymbol{\beta}) - W_d(\frac{\lceil N\alpha \rceil}{N}, \beta) \right] \, \mathrm{d}\beta \, \mathrm{d}\alpha \right| \\
&\leq N \int_{I_i^N} \int_{\mathcal{I}^{r<[k_d]\setminus\{1\}}} \left| W_d(\alpha, \boldsymbol{\beta}) - W_d(\frac{\lceil N\alpha \rceil}{N}, \beta) \right| \, \mathrm{d}\beta \, \mathrm{d}\alpha \\
&\leq N \frac{1}{N} \cdot \frac{L_W}{N} = \frac{L_W}{N} \to 0
\end{aligned}$$

by Assumption 3.3.2. Alternatively, under only block-wise Lipschitz W as in Eq. (3.3.35), the same result is obtained by first separating out finitely many i (at most $Q-1$) for which Lipschitzness fails, trivially bounding their terms by $\frac{2(Q-1)}{N}$. For all other i, there exists $k \in \{1, \dots, Q\}$ such that $I_i^N \times I_j \subseteq I_k \times I_j$, i.e. the Lipschitz bound applies.

For the third term, again fix $\varepsilon, \delta > 0$. Then, by our initial assumption of Eq. (3.3.45), for sufficiently large N there exists a set $\mathcal{J}_2^N$, $|\mathcal{J}_2^N| \geq \lfloor (1-\delta)N \rfloor$ such that

$$\left| \mathbb{E}\left[h(x_t^i, \nu^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] - \mathbb{E}\left[h(\hat{x}_t^{\frac{i}{N}}, \nu^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| < \varepsilon, \quad \forall i \in \mathcal{J}_2^N$$

independent of $\hat{\pi} \in \Pi$.

This completes the proof of (3.3.45) $\implies$ (3.3.46) at any time t, since by the prequel, the intersection of all correspondingly chosen, finitely many sets $\mathcal{J}_i^N$ for sufficiently large N has at least $N - \sum_i \lceil \delta_i N \rceil$ elements, which is always larger than $N - \lceil \delta N \rceil$ for any $\delta > 0$ by choosing δ_i sufficiently small.

Finally, we show Eq. (3.3.45) at all times t using the prequel by induction, which will imply Eq. (3.3.46). By definition for $t = 0$, $\hat{x}_t^{\frac{i}{N}} \sim \mu_0$ and $x_t^i \sim \mu_0$ imply

$$\left| \mathbb{E}\left[g(x_0^i) \right] - \mathbb{E}\left[g(\hat{x}_0^{\frac{i}{N}}) \right] \right| = 0.$$

For the induction step, define the uniform bound M_g of functions in $\mathcal{G}$. Observe that

$$\begin{aligned} &\left| \mathbb{E}\left[g(x_{t+1}^i) \right] - \mathbb{E}\left[g(\hat{x}_{t+1}^{\frac{i}{N}}) \right] \right| \\ &\quad = \left| \mathbb{E}\left[l_{N,t}(x_t^i, \nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N)) \right] - \mathbb{E}\left[l_{N,t}(\hat{x}_t^{\frac{i}{N}}, \nu^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| \end{aligned}$$

using the uniformly bounded and Lipschitz functions

$$l_{N,t}(x, \nu) \coloneqq \sum_{u \in \mathcal{U}} \hat{\pi}_t(u \mid x) \sum_{x' \in \mathcal{X}} p(x' \mid x, u, \nu) g(x')$$

with bound M_g and Lipschitz constant $|\mathcal{X}| M_g L_p$. By induction assumption (3.3.45) and (3.3.45) $\implies$ (3.3.46), there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\left| \mathbb{E}\left[l_{N,t}(x_t^i, \nu_N^{\frac{i}{N}}(\boldsymbol{\mu}_t^N)) \right] - \mathbb{E}\left[l_{N,t}(\hat{x}_t^{\frac{i}{N}}, \nu^{\frac{i}{N}}(\boldsymbol{\mu}_t)) \right] \right| < \varepsilon$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}^N$ for some $\mathcal{J}^N \subseteq [N]$ with $|\mathcal{J}^N| \geq \lfloor (1-\delta)N \rfloor$. This concludes the proof by induction. □

C.4 PROOF OF COROLLARY 3.3.1

Proof. The result follows more or less directly from Theorem 3.3.3. Consider first the finite horizon case $\mathcal{T} = \{0, 1, \dots, T-1\}$. Define

$$r^{\hat{\pi}}(x, \nu) \coloneqq \sum_{u \in \mathcal{U}} r(x, u, \nu) \hat{\pi}_t(u \mid x)$$

with uniform bound M_R and Lipschitz constant $|U|L_r$. Therefore, by choosing the maximum over all N' for all finitely many times $t \in \mathcal{T}$ via Theorem 3.3.3, there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\begin{aligned}&\left|J_i^N(\pi^1,\dots,\pi^{i-1},\hat{\pi},\pi^{i+1},\dots,\hat{\pi}) - J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\hat{\pi})\right| \\ &\quad\leq \sum_{t=0}^{T-1}\left|\mathbb{E}\left[r^{\hat{\pi}_t}(x_t^i,\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] - \mathbb{E}\left[r^{\hat{\pi}_t}(\hat{x}_t^{\frac{i}{N}},\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| < \varepsilon.\end{aligned}$$

uniformly over $\hat{\pi} \in \Pi, i \in \mathcal{J}^N$ for some $\mathcal{J}^N \subseteq [N]$ with $|\mathcal{J}^N| \geq \lfloor(1-\delta)N\rfloor$.

For the infinite horizon problem $\mathcal{T} = \mathbb{N}_0$, we first pick some time $T' > \frac{\log\frac{\varepsilon(1-\gamma)}{4M_R}}{\log\gamma}$ such that

$$\begin{aligned}&\sum_{t=0}^{T'-1}\gamma^t\left|\mathbb{E}\left[r_{\hat{\pi}_t}(x_t^i,\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] - \mathbb{E}\left[r_{\hat{\pi}_t}(\hat{x}_t^{\frac{i}{N}},\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| \\ &\qquad+\gamma^{T'}\sum_{t=T'}^{\infty}\gamma^{t-T'}\left|\mathbb{E}\left[r_{\hat{\pi}_t}(x_t^i,\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] - \mathbb{E}\left[r_{\hat{\pi}_t}(\hat{x}_t^{\frac{i}{N}},\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| \\ &\quad< \sum_{t=0}^{T'-1}\gamma^t\left|\mathbb{E}\left[r_{\hat{\pi}_t}(x_t^i,\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right] - \mathbb{E}\left[r_{\hat{\pi}_t}(\hat{x}_t^{\frac{i}{N}},\nu^{\frac{i}{N}}(\boldsymbol{\mu}_t))\right]\right| + \frac{\varepsilon}{2}\end{aligned}$$

and again apply Theorem 3.3.3 to the remaining finite sum. □

C.5 PROOF OF COROLLARY 3.3.2

Proof. The result follows directly from Corollary 3.3.1. Let $\varepsilon, \delta > 0$, then by Corollary 3.3.1 there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\begin{aligned}&\max_{\pi\in\Pi}\left(J_i^N(\pi^1,\dots,\pi^{i-1},\pi,\pi^{i+1},\dots,\pi^N) - J_i^N(\pi^1,\dots,\pi^N)\right) \\ &\quad\leq \max_{\pi\in\Pi}\left(J_i^N(\pi^1,\dots,\pi^{i-1},\pi,\pi^{i+1},\dots,\pi^N) - J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi)\right) \\ &\qquad+\max_{\pi\in\Pi}\left(J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi) - J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi^{\frac{i}{N}})\right) \\ &\qquad+\left(J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi^{\frac{i}{N}}) - J_i^N(\pi^1,\dots,\pi^N)\right) \\ &\quad< \frac{\varepsilon}{2} + 0 + \frac{\varepsilon}{2} = \varepsilon\end{aligned}$$

uniformly over $i \in \mathcal{J}^N$ for some $\mathcal{J}^N \subseteq [N]$ with $|\mathcal{J}^N| \geq \lfloor(1-\delta)N\rfloor$, where by definition of equilibrium optimality we obtained

$$\max_{\pi\in\Pi}\left(J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi) - J_{\frac{i}{N}}^{\boldsymbol{\mu}}(\pi^{\frac{i}{N}})\right) = 0.$$

This concludes the proof. □

C.6 ADDITIONAL EXPERIMENTS

In Figure C.1, we show additional results for the Rumor problem and inverted 3-uniform hypergraphons. There, we find almost inverted results as in Figure 3.12, indicating that the influence of connections from the second layer are more important under the given problem parameters. However, we note that surprisingly, the highest awareness is reached for intermediate α.

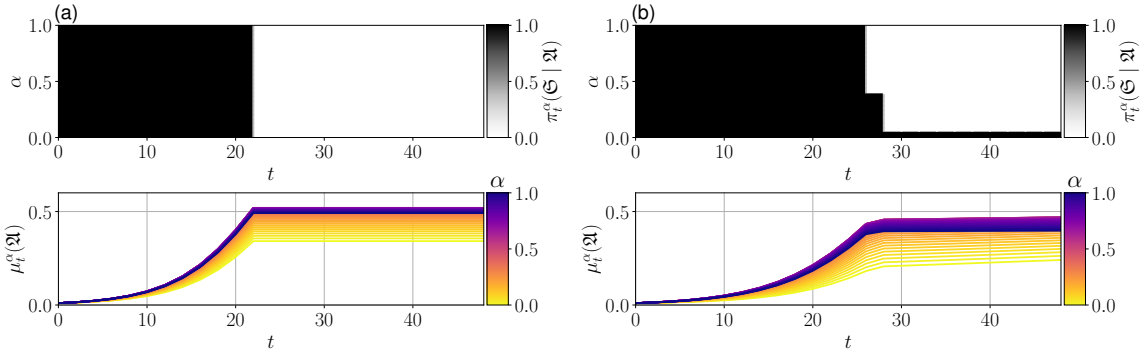


FIGURE C.1: Equilibrium behavior for the Rumor problem. (a): $(W_{\text{rank}}, \hat{W}_{\text{inv-unif}})$; (b): $(W_{\text{unif}}, \hat{W}_{\text{inv-unif}})$.

As an additional example, in the timely SIS problem, we assume that there exists an epidemic that spreads to neighboring nodes according to the classical SIS dynamics, see e.g. [49]. Analogously, we may consider extensions to arbitrary variations of the SIS model such as SIR or SEIR. Each healthy (or susceptible, $\mathfrak{S}$) agent can take costly precautions ($\mathfrak{P}$) to avoid becoming infected ($\mathfrak{I}$), or ignore ($\bar{\mathfrak{P}}$) precautions at no further cost. Since being infected itself is costly, an equilibrium solution must balance the expected cost of infections against the cost of taking precautions.

Formally, we define the state space $\mathcal{X} = \{\mathfrak{S}, \mathfrak{I}\}$ and action space $\mathcal{U} = \{\bar{\mathfrak{P}}, \mathfrak{P}\}$ such that

$$p(\mathfrak{I} \mid \mathfrak{S}, \bar{\mathfrak{P}}, \boldsymbol{\nu}) = \min\left(1, \sum_{d \in [D]} \tau_d \boldsymbol{\nu}_d \left(\mathbf{1}_{\{\mathfrak{I}\}}\right)\right)$$
$$p(\mathfrak{I} \mid \mathfrak{S}, \mathfrak{P}, \cdot) = 0, \quad p(\mathfrak{S} \mid \mathfrak{I}, \cdot, \cdot) = \delta$$

with infection rates $\tau_d > 0$, $\sum_d \tau_d \leq 1$, recovery rate $\delta \in (0, 1)$ and rewards $r(x, u, \cdot) = c_P \mathbf{1}_{\{\mathfrak{P}\}}(u) + c_I \mathbf{1}_{\{\mathfrak{I}\}}(x)$ with infection and precaution costs $c_P > 0$, $c_I > 0$. In our experiments, we will use $\tau_d = 0.8$, $\delta = 0.2$, $c_P = 0.5$, $c_I = 2$, $\mu_0(\mathfrak{I}) = 0.5$ and $\mathcal{T} = \{0, 1, \ldots, 49\}$.

Existing state-of-the-art approaches such as online mirror descent (OMD) [117](and similarly FP, see e.g. [9]) as depicted in Figure C.2 and Figure C.3 for 10 discretization points did not converge to an equilibrium in the considered 2000 iterations, though we expect that the methods will converge when running for significantly more iterations – e.g. 400000 iterations as in [130] – which we could not verify here due to the computational complexity. We expect that existing standard results using monotonicity conditions [22, 117] can be extended to the hypergraphon case in order to guarantee convergence of aforementioned learning algorithms. However, this remains outside the scope of our work. In particular for the ranked-attachment graphon and hypergraphon, the final behavior as seen in Figure C.2 remains with an average final exploitability ΔJ of above 0.25, which is defined as

$$\Delta J(\pi) = \int_{\mathcal{I}} \sup_{\pi^* \in \Pi} J_\alpha^{\Psi(\pi)}(\pi^*) - J_\alpha^{\Psi(\pi)}(\pi) \, \mathrm{d}\alpha$$

and must be zero for an exact equilibrium.

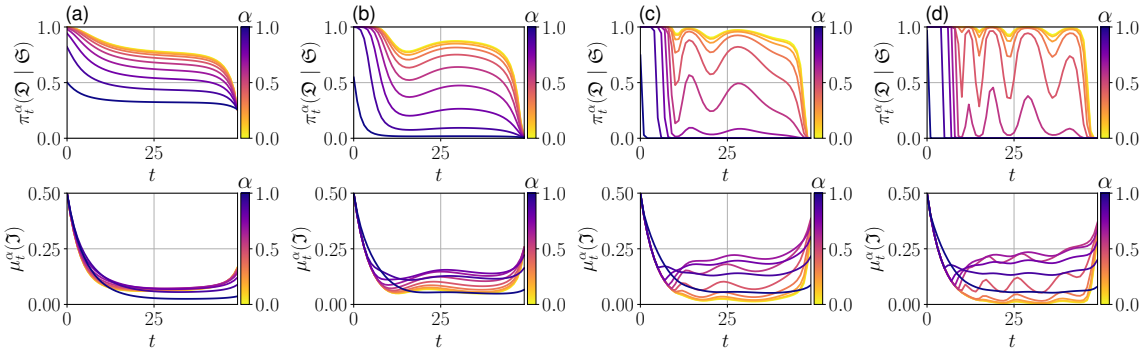


FIGURE C.2: Equilibrium policy and MF for graphons $(W_{\text{unif}}, \hat{W}_{\text{unif}})$ from Figure C.3 at different iterations n. It can be observed that in the SIS problem, the solution oscillates between taking precautions and not taking precautions. (a): $n = 20$; (b): $n = 100$; (c): $n = 500$; (d): $n = 1500$.

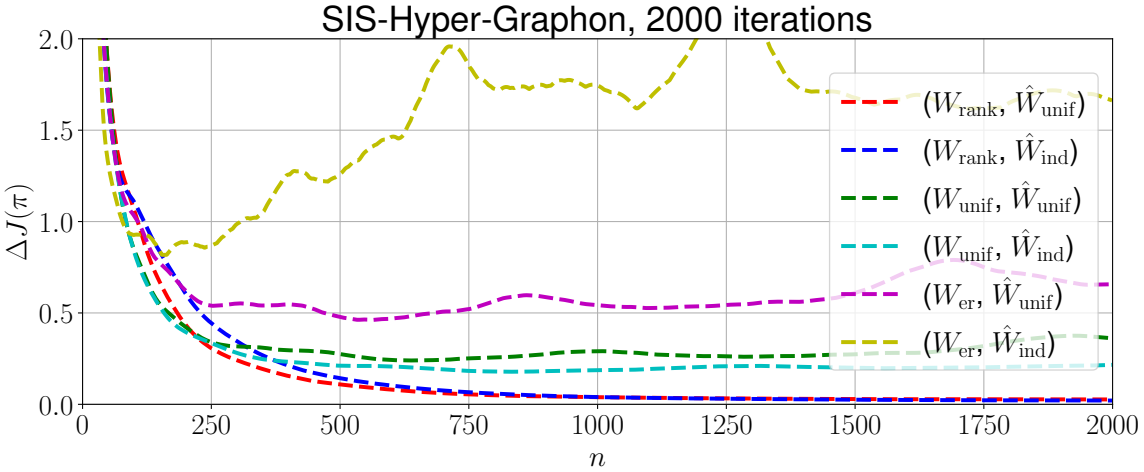


FIGURE C.3: Average exploitability over iterations n of Online Mirror Descent [117] on the SIS problem. It can be observed that for some configurations, the method fails to converge to an equilibrium.

D

APPENDIX D: SUPPLEMENTARY DETAILS ON SECTION 3.4

D.1 CONTINUOUS TIME FICTITIOUS PLAY WITH MAJOR AND MINOR AGENTS

In this section, we prove the FP convergence result for the MFG with a major agent by extending the ideas of [127] to include the major agent.

The aim of this proof is to show the total exploitability is a strong Lyapunov function by showing $\frac{d}{d\tau}\left(\mathcal{E}(\bar{\pi}^{\tau},\bar{\pi}^{0,\tau})+\mathcal{E}^0(\bar{\pi}^{\tau},\bar{\pi}^{0,\tau})\right)\leq-\frac{1}{\tau}\left(\mathcal{E}(\bar{\pi}^{\tau},\bar{\pi}^{0,\tau})+\mathcal{E}^0(\bar{\pi}^{\tau},\bar{\pi}^{0,\tau})\right)$. In the proof we focus on showing $\frac{d}{d\tau}\left(\mathcal{E}(\bar{\pi}^{\tau},\bar{\pi}^{0,\tau})\right)\leq-\frac{1}{\tau}\left(\mathcal{E}(\bar{\pi}^{\tau},\bar{\pi}^{0,\tau})\right)$ first and $\frac{d}{d\tau}\left(\mathcal{E}^0(\bar{\pi}^{\tau},\bar{\pi}^{0,\tau})\right)\leq-\frac{1}{\tau}\left(\mathcal{E}^0(\bar{\pi}^{\tau},\bar{\pi}^{0,\tau})\right)$ second and we combine the results at the end. Before going into the details of the proof of Theorem 3.4.2, we introduce some properties that extend from the ones introduced by [127] where we exchange their common noise formulation Σ_t^0 with state-action histories (i.e., we condition on the noise in order to avoid having to conditioning on the major agent's randomness as common noise).

We first recall a few definitions:

MEAN FIELD. We recall the conditional MF for minor agents, now conditioned on histories, i.e. recursively

$$\begin{aligned}
&\mu^{\pi}_{t+1|x^0_{0:t},u^0_{0:t}}(x') \coloneqq \mathbb{P}_{\pi}(x_{t+1}=x' \mid x^0_{0:t},u^0_{0:t})\\
&= \sum_{x,u\in\mathcal{X}\times\mathcal{U}} \mathbb{P}_{\pi}(x_{t+1}=x' \mid x_t=x,u_t=u,x^0_{0:t},u^0_{0:t})\mathbb{P}_{\pi}(u_t=u,x_t=x \mid x^0_{0:t},u^0_{0:t})\\
&= \sum_{x,u\in\mathcal{X}\times\mathcal{U}} p(x'|x,u,x^0_t,u^0_t)\pi_t(u|x,x^0_{0:t},u^0_{0:t-1})\mu^{\pi}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x).
\end{aligned}$$

while for major agents, we define joint history MFs

$$\begin{aligned}
\mu^{\pi^0}_{t+1}(x^0_{t+1},x^0_{0:t},u^0_{0:t}) &\coloneqq \mathbb{P}_{\pi^0}(x^0_{t+1},x^0_{0:t},u^0_{0:t})\\
&= \mu^{\pi^0}_t(x^0_t,x^0_{0:t-1},u^0_{0:t-1})\pi^0_t(u^0_t \mid x^0_t,x^0_{0:t-1},u^0_{0:t-1})p^0(x^0_{t+1} \mid x^0_t,u^0_t)
\end{aligned}$$

AVERAGED POLICIES. We recall the minor agent FP policy as

$$\bar{\pi}^{\tau}_t(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1}) = \frac{\int_0^{\tau}\pi^{BR,s}(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1})\mu^{\pi^{BR,s}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)ds}{\int_0^{\tau}\mu^{\pi^{BR,s}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)ds} \tag{D.1.1}$$

and similarly for the major agent

$$\bar{\pi}^{0,\tau}_t(u^0_t \mid x^0_t,x^0_{0:t-1},u^0_{0:t-1}) = \frac{\int_0^{\tau}\pi^{0,BR,s}(u^0_t \mid x^0_t,x^0_{0:t-1},u^0_{0:t-1})\mu^{\pi^{0,BR,s}_{0:t-1}}_t(x^0_t,x^0_{0:t-1},u^0_{0:t-1})ds}{\int_0^{\tau}\mu^{\pi^{0,BR,s}_{0:t-1}}_t(x^0_t,x^0_{0:t-1},u^0_{0:t-1})ds}. \tag{D.1.2}$$

AVERAGED MEAN FIELD. We also recall the average FP MF for minor agents

$$\bar{\mu}^{\tau}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t) \coloneqq \frac{1}{\tau}\int_0^{\tau}\mu^{\pi^{BR,s}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)ds. \tag{D.1.3}$$

and for major agents

$$\bar{\mu}^{0,\tau}_t(x^0_t,x^0_{0:t-1},u^0_{0:t-1}) \coloneqq \frac{1}{\tau}\int_0^{\tau}\mu^{\pi^{0,BR,s}_{0:t-1}}_t(x^0_t,x^0_{0:t-1},u^0_{0:t-1})ds \tag{D.1.4}$$

Property D.1.1. *When the game is monotone, we have:*

$$\sum_{x\in\mathcal{X}}\left\langle \nabla_{\mu}\overline{r}(x,x^0,\mu),\frac{d}{d\tau}\mu\right\rangle\frac{d}{d\tau}\mu(x)\leq 0$$

Proof. For all $s\geq 0$ monotonicity condition says that for a fixed $x^0\in\mathcal{X}^0$:

$$\begin{aligned}
&\sum_{x\in\mathcal{X}}(\mu^{\tau}(x)-\mu^{\tau+s}(x))(\overline{r}(x,x^0,\mu^{\tau})-\overline{r}(x,x^0,\mu^{\tau+s}))\leq 0\\
\Rightarrow &\sum_{x\in\mathcal{X}}\frac{\mu^{\tau}(x)-\mu^{\tau+s}(x)}{s}\frac{\overline{r}(x,x^0,\mu^{\tau})-\overline{r}(x,x^0,\mu^{\tau+s})}{s}\leq 0.
\end{aligned}$$

Therefore, the result follows when $s\to 0$. □

Property D.1.2. *The above FP policy $\bar{\pi}^0$ generates the average FP MF $\bar{\mu}$, i.e.*

$$\bar{\mu}_t^{0,s} = \mu_t^{\bar{\pi}_{0:t-1}^{0,s}} \tag{D.1.5}$$

at all times t, τ, and analogously for the minor agents

$$\bar{\mu}_t^{s} = \mu_t^{\bar{\pi}_{0:t-1}^{s}}. \tag{D.1.6}$$

Proof. The joint probabilities of any policy $\pi^{0,BR,s}$ are always given by

$$\begin{aligned}
&\mu_{t+1}^{\pi_{0:t}^{0,BR,s}}(x_{t+1}^0, x_{0:t}^0, u_{0:t}^0) \\
&\quad = p^0(x_{t+1}^0 \mid x_t^0, u_t^0)\mu_t^{\pi_{0:t-1}^{0,BR,s}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\pi_t^{0,BR,s}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)
\end{aligned}$$

and therefore, integrating over all times τ, we have

$$\begin{aligned}
&\frac{1}{\tau}\int_0^\tau \mu_{t+1}^{\pi_{0:t}^{0,BR,s}}(x_{t+1}^0, x_{0:t}^0, u_{0:t}^0)ds \\
&\quad = p^0(x_{t+1}^0 \mid x_t^0, u_t^0)\frac{1}{\tau}\int_0^\tau \mu_t^{\pi_{0:t-1}^{0,BR,s}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\pi_t^{0,BR,s}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)ds.
\end{aligned}$$

Then, using definitions of $\bar{\mu}^{0,s}$ and $\bar{\pi}^{0,s}$, we have by induction starting with $\bar{\mu}^{0,s} = \mu_0^0 = \mu_0^{\bar{\pi}_{0:t}^{0,s}}$,

$$\begin{aligned}
&\bar{\mu}_{t+1}^{0,s}(x_{t+1}^0, x_{0:t}^0, u_{0:t}^0) \\
&\quad = p^0(x_{t+1}^0 \mid x_t^0, u_t^0)\bar{\mu}_t^{0,s}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\bar{\pi}_t^{0,s}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) \\
&\quad = p^0(x_{t+1}^0 \mid x_t^0, u_t^0)\mu_t^{\bar{\pi}_{0:t-1}^{0,s}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\bar{\pi}_t^{0,s}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) \\
&\quad = \mu_{t+1}^{\bar{\pi}_{0:t}^{0,s}}(x_{t+1}^0, x_{0:t}^0, u_{0:t}^0)
\end{aligned}$$

which is the desired result, i.e. the MF generated by $\bar{\pi}^0$ is the same as the average MF of best responses.

For the minor agents, the proof is analogous: The conditional probabilities of any policy $\pi^{BR,s}$ are always given by

$$\begin{aligned}
&\mu_{t+1|x_{0:t}^0, u_{0:t}^0}^{\pi_{0:t}^{BR,s}}(x_{t+1}) \\
&\quad = \mathbb{P}_{\pi^{BR,s}}(x_{t+1}^0 \mid x_{0:t}^0, u_{0:t}^0) \\
&\quad = \sum_{x_t, u_t \in \mathcal{X}\times\mathcal{U}} p(x_{t+1} \mid x_t, u_t, x_t^0, u_t^0)\mu_{t|x_{0:t-1}^0, u_{0:t-1}^0}^{\pi_{0:t}^{BR,s}}(x_t)\pi_t^{BR,s}(u_t|x_t, x_{0:t}^0, u_{0:t-1}^0)
\end{aligned}$$

and therefore, by integrating over all times τ,

$$\begin{aligned}
&\mu_{t+1|x_{0:t}^0, u_{0:t}^0}^{\pi_{0:t}^{BR,s}}(x_{t+1}) \\
&\quad = \sum_{x_t, u_t \in \mathcal{X}\times\mathcal{U}} p(x_{t+1} \mid x_t, u_t, x_t^0, u_t^0)\frac{1}{\tau}\int_0^\tau \mu_{t|x_{0:t-1}^0, u_{0:t-1}^0}^{\pi_{0:t}^{BR,s}}(x_t)\pi_t^{BR,s}(u_t|x_t, x_{0:t}^0, u_{0:t-1}^0)ds
\end{aligned}$$

we obtain

$$\bar{\mu}_{t+1|x_{0:t}^0, u_{0:t}^0}^{s}(x_{t+1})$$

$$= \sum_{x_t,u_t \in \mathcal{X}\times\mathcal{U}} p(x_{t+1} \mid x_t, u_t, x_t^0, u_t^0)\mu_{t|x_{0:t-1}^0,u_{0:t-1}^0}^{\bar{\pi}_{0:t}^s}(x_t)\bar{\pi}_t^s(u_t|x_t, x_{0:t}^0, u_{0:t-1}^0)ds$$
$$= \mu_{t+1|x_{0:t}^0,u_{0:t}^0}^{\bar{\pi}_{0:t}^s}(x_{t+1})$$

which again implies the desired result for minor agents by induction. □

Property D.1.3. *At all times t, τ and state-actions $x_{0:t-1}^0, u_{0:t-1}^0, x_t^0$, we have*

$$\frac{d}{d\tau}\mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) = \frac{1}{\tau}\Big[\mu_t^{\pi_{0:t-1}^{0,BR,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) - \mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\Big] \quad \text{(D.1.7)}$$

$$\mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\frac{d}{d\tau}\bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) \quad \text{(D.1.8)}$$
$$= \mu_t^{\pi_{0:t-1}^{0,BR,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\frac{1}{\tau}\Big[\pi_t^{0,BR,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) - \bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\Big].$$

and also for the minor agents

$$\frac{d}{d\tau}\mu_{t|x_{0:t-1}^0,u_{0:t-1}^0}^{\bar{\pi}_{0:t-1}^{\tau}}(x_t) = \frac{1}{\tau}\Big[\mu_{t|x_{0:t-1}^0,u_{0:t-1}^0}^{\pi_{0:t-1}^{BR,\tau}}(x_t) - \mu_{t|x_{0:t-1}^0,u_{0:t-1}^0}^{\bar{\pi}_{0:t-1}^{\tau}}(x_t)\Big] \quad \text{(D.1.9)}$$

$$\mu_{t|x_{0:t-1}^0,u_{0:t-1}^0}^{\bar{\pi}_{0:t-1}^{\tau}}(x_t)\frac{d}{d\tau}\bar{\pi}_t^{\tau}(u_t \mid x_t, x_{0:t}^0, u_{0:t-1}^0) \quad \text{(D.1.10)}$$
$$= \mu_{t|x_{0:t-1}^0,u_{0:t-1}^0}^{\pi_{0:t-1}^{BR,\tau}}(x_t)\frac{1}{\tau}\Big[\pi_t^{BR,\tau}(u_t \mid x_t, x_{0:t}^0, u_{0:t-1}^0) - \bar{\pi}_t^{\tau}(u_t \mid x_t, x_{0:t}^0, u_{0:t-1}^0)\Big].$$

Proof. The properties are shown together inductively. First, note that Eq. (D.1.7) implies Eq. (D.1.8): Start with Property D.1.2, giving

$$\bar{\mu}_t^{0,s}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\bar{\pi}_t^{0,\tau}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)$$
$$= \frac{1}{\tau}\int_0^\tau \pi^{0,BR,s}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)\mu_t^{\pi_{0:t-1}^{0,BR,s}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)ds$$

which implies by taking the derivative $\frac{d}{d\tau}$ that

$$\frac{d}{d\tau}\bar{\mu}_t^{0,s}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\bar{\pi}_t^{0,\tau}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)$$
$$+ \bar{\mu}_t^{0,s}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\frac{d}{d\tau}\bar{\pi}_t^{0,\tau}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)$$
$$= -\frac{1}{\tau^2}\int_0^\tau \pi^{0,BR,s}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)\mu_t^{\pi_{0:t-1}^{0,BR,s}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)ds$$
$$+ \frac{1}{\tau}\Big[\pi^{0,BR,s}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)\mu_t^{\pi_{0:t-1}^{0,BR,s}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)ds\Big].$$

Applying Eq. (D.1.7) and Property D.1.2 gives

$$\bar{\mu}_t^{0,s}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\frac{d}{d\tau}\bar{\pi}_t^{0,\tau}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)$$
$$= -\frac{1}{\tau^2}\int_0^\tau \pi^{0,BR,s}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)\mu_t^{\pi_{0:t-1}^{0,BR,s}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)ds$$
$$+ \frac{1}{\tau}\Big[\pi^{0,BR,s}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)\mu_t^{\pi_{0:t-1}^{0,BR,s}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)ds\Big]$$
$$- \frac{1}{\tau}\Big[\mu_t^{\pi_{0:t-1}^{0,BR,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) - \mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\Big]\bar{\pi}_t^{0,\tau}(x^0, x_{0:t-1}^0, u_{0:t-1}^0)$$
$$= \mu_t^{\pi_{0:t-1}^{0,BR,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\frac{1}{\tau}\Big[\bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) - \pi_t^{0,BR,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\Big].$$

Now to show Eq. (D.1.7) at all times t, we use induction. At time 0, the property is trivially fulfilled by fixed μ_0^0. Assume Eq. (D.1.7) and therefore Eq. (D.1.8) holds at time t, then for the induction step, at time $t+1$, by Property D.1.2, we have

$$
\begin{aligned}
&\frac{d}{d\tau}\mu_{t+1}^{\bar{\pi}_{0:t}^{0,\tau}}(x_{t+1}^0, x_{0:t}^0, u_{0:t}^0) \\
&= \frac{d}{d\tau}\Big[\mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)p^0(x_{t+1}^0 \mid x_t^0, u_t^0)\Big] \\
&= \frac{d}{d\tau}\mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)p^0(x_{t+1}^0 \mid x_t^0, u_t^0) \\
&\quad + \mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\frac{d}{d\tau}\bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)p^0(x_{t+1}^0 \mid x_t^0, u_t^0) \\
&= \frac{1}{\tau}\Big[\mu_t^{\pi_{0:t-1}^{0,BR,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) - \mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\Big] \\
&\qquad \cdot \bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)p^0(x_{t+1}^0 \mid x_t^0, u_t^0) \\
&\quad + \mu_t^{\pi_{0:t-1}^{0,BR,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) \\
&\qquad \cdot \frac{1}{\tau}\Big[\pi_t^{0,BR,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0) - \bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\Big]p^0(x_{t+1}^0 \mid x_t^0, u_t^0) \\
&= \frac{1}{\tau}\mu_t^{\pi_{0:t-1}^{0,BR,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\pi_t^{0,BR,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)p^0(x_{t+1}^0 \mid x_t^0, u_t^0) \\
&\quad - \frac{1}{\tau}\mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)p^0(x_{t+1}^0 \mid x_t^0, u_t^0) \\
&= \frac{1}{\tau}\Big[\mu_{t+1}^{\pi_{0:t}^{0,BR,\tau}}(x_{t+1}^0, x_{0:t}^0, u_{0:t}^0) - \mu_{t+1}^{\bar{\pi}_{0:t}^{0,\tau}}(x_{t+1}^0, x_{0:t}^0, u_{0:t}^0)\Big]
\end{aligned}
$$

where we used the induction assumption on

$$
\mu_t^{\bar{\pi}_{0:t-1}^{0,\tau}}(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)\frac{d}{d\tau}\bar{\pi}_t^{0,\tau}(u_t^0 \mid x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)
$$

to obtain Eq. (D.1.7).

For the minor agents, the proof is analogous in that Eq. (D.1.9) implies Eq. (D.1.10) by Property D.1.2. Meanwhile, Eq. (D.1.9) follows readily by noting $\bar{\mu}_t^s = \mu_t^{\bar{\pi}_{0:t-1}^s}$ by Property D.1.2 and taking instead the derivative of the definition of $\bar{\mu}_t^s$. □

Proof of Theorem 3.4.2. With the above properties established, we can show the convergence of exploitabilities to zero.

STEP 1: FOCUSING ON THE EXPLOITABILITY OF THE MINOR AGENT. We first start with showing that the exploitability of minor agents, $\mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$ is a strong Lyapunov function. Using the definition of exploitability of the minor agent, we can write:

$$
\begin{aligned}
&\frac{d}{d\tau}\mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) \\
&= \frac{d}{d\tau}\Big[\max_{\pi'} J(\pi', \bar{\pi}^\tau, \bar{\pi}^{0,\tau}) - J(\bar{\pi}^\tau, \bar{\pi}^\tau, \bar{\pi}^{0,\tau})\Big] \\
&= \frac{d}{d\tau}\sum_{t\in\mathcal{T}}\sum_{\substack{x_t\in\mathcal{X}, u_t\in\mathcal{X}, \\ x_{0:t}^0\in\mathcal{X}^{0^{t+1}}, \\ u_{0:t-1}^0\in\mathcal{U}^{0^t}}}\Big[\mathbb{P}_{\pi^{BR,\tau},\bar{\pi}^\tau,\bar{\pi}^{0,\tau}}(x_t, u_t, x_{0:t}^0, u_{0:t-1}^0) - \mathbb{P}_{\bar{\pi}^\tau,\bar{\pi}^\tau,\bar{\pi}^{0,\tau}}(x_t, u_t, x_{0:t}^0, u_{0:t-1}^0)\Big]
\end{aligned}
$$

$$\cdot\, r(x_t, u_t, x_t^0, \mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

and, e.g., at time t for the FP policy

$$\frac{d}{d\tau} \sum_{\substack{x_t\in\mathcal{X},u_t\in\mathcal{X},\\ x^0_{0:t}\in\mathcal{X}^{0^{t+1}},\\ u^0_{0:t-1}\in\mathcal{U}^{0^t}}} \mathbb{P}_{\bar{\pi}^\tau,\bar{\pi}^\tau,\bar{\pi}^{0,\tau}}(x_t,u_t,x^0_{0:t},u^0_{0:t-1}) r(x_t,x_t^0,\mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

$$= \frac{d}{d\tau} \sum_{t\in\mathcal{T}} \sum_{\substack{x_t\in\mathcal{X},u_t\in\mathcal{X},\\ x^0_{0:t}\in\mathcal{X}^{0^{t+1}},\\ u^0_{0:t-1}\in\mathcal{U}^{0^t}}} \mu_t^{\bar{\pi}^{0,\tau}_{0:t-1}}(x_t^0,x^0_{0:t-1},u^0_{0:t-1}) \mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t) \bar{\pi}^\tau_t(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1})$$

$$\cdot\, r(x_t, u_t, x_t^0, \mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

and similarly for the BR policy. By the envelope theorem on $\max_{\pi'} J^0(\bar{\pi}^\tau,\pi')$, the partial derivative with respect to π^{BR} can be dropped. To be precise, here and in [127, Appendix A], continuous differentiability of the objectives with respect to τ is implicitly assumed, which if needed may be guaranteed by introducing a minimal regularization. This is not a problem however, as for example entropy-regularized MFE still achieve arbitrarily good unregularized finite game equilibria, see e.g. [9, Theorem 4]. Therefore, we obtain

$$\frac{d}{d\tau} \sum_{t\in\mathcal{T}} \sum_{\substack{x_t\in\mathcal{X},u_t\in\mathcal{X},\\ x^0_{0:t}\in\mathcal{X}^{0^{t+1}},\\ u^0_{0:t-1}\in\mathcal{U}^{0^t}}} \left[\mathbb{P}_{\pi^{BR,\tau},\bar{\pi}^\tau,\bar{\pi}^{0,\tau}}(x_t,u_t,x^0_{0:t},u^0_{0:t-1}) - \mathbb{P}_{\bar{\pi}^\tau,\bar{\pi}^\tau,\bar{\pi}^{0,\tau}}(x_t,u_t,x^0_{0:t},u^0_{0:t-1})\right]$$

$$\cdot\, r(x_t, u_t, x_t^0, \mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

$$= \sum_{t\in\mathcal{T}} \sum_{\substack{x_t\in\mathcal{X},u_t\in\mathcal{X},\\ x^0_{0:t}\in\mathcal{X}^{0^{t+1}},\\ u^0_{0:t-1}\in\mathcal{U}^{0^t}}} \Bigg[$$

$$\frac{d}{d\tau}\mu_t^{\bar{\pi}^{0,\tau}_{0:t-1}}(x_t^0,x^0_{0:t-1},u^0_{0:t-1}) \mu^{\pi^{BR,\tau}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t) \pi_t^{BR,\tau}(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1}) r(x_t,u_t,x_t^0,\mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

$$+ \mu_t^{\bar{\pi}^{0,\tau}_{0:t-1}}(x_t^0,x^0_{0:t-1},u^0_{0:t-1}) \mu^{\pi^{BR,\tau}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)$$

$$\cdot\, \pi_t^{BR,\tau}(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1}) \frac{d}{d\tau} r(x_t,u_t,x_t^0,\mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

$$- \frac{d}{d\tau}\mu_t^{\bar{\pi}^{0,\tau}_{0:t-1}}(x_t^0,x^0_{0:t-1},u^0_{0:t-1}) \mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t) \bar{\pi}^\tau_t(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1}) r(x_t,u_t,x_t^0,\mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

$$- \mu_t^{\bar{\pi}^{0,\tau}_{0:t-1}}(x_t^0,x^0_{0:t-1},u^0_{0:t-1}) \frac{d}{d\tau}\mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t) \bar{\pi}^\tau_t(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1}) r(x_t,u_t,x_t^0,\mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

$$- \mu_t^{\bar{\pi}^{0,\tau}_{0:t-1}}(x_t^0,x^0_{0:t-1},u^0_{0:t-1}) \mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t) \frac{d}{d\tau}\bar{\pi}^\tau_t(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1}) r(x_t,u_t,x_t^0,\mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})$$

$$- \mu_t^{\bar{\pi}^{0,\tau}_{0:t-1}}(x_t^0,x^0_{0:t-1},u^0_{0:t-1}) \mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t) \bar{\pi}^\tau_t(u_t \mid x_t,x^0_{0:t},u^0_{0:t-1}) \frac{d}{d\tau} r(x_t,u_t,x_t^0,\mu^{\bar{\pi}^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}) \Bigg]$$

where for the first and third term, we use the property from major agent, while for fourth term we use Eq. (D.1.9) in Property D.1.3, and for the fifth we use Eq. (D.1.10) in Property D.1.3.

Therefore, combining the first and third, second and last term, and fourth and fifth terms, we have

$$\frac{d}{d\tau}\mathcal{E}(\bar{\pi}^\tau,\bar{\pi}^{0,\tau})$$

$$
\begin{aligned}
&= \sum_{t\in\mathcal{T}} \sum_{\substack{x_t\in\mathcal{X},u_t\in\mathcal{X},\\ x^0_{0:t}\in\mathcal{X}^{0\,t+1},\\ u^0_{0:t-1}\in\mathcal{U}^{0\,t}}} \Bigg[\\
&\frac{1}{\tau}\Big[\mu_t^{\pi^{0,BR,\tau}_{0:t-1}}(x^0_t,x^0_{0:t-1},u^0_{0:t-1}) - \mu_t^{\bar\pi^{0,\tau}_{0:t-1}}(x^0_t,x^0_{0:t-1},u^0_{0:t-1})\Big] r(x_t,u_t,x^0_t,\mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}) \\
&\quad\cdot\Big[\mu^{\pi^{BR,\tau}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)\pi^{BR,\tau}_t(u_t\mid x_t,x^0_{0:t},u^0_{0:t-1}) - \mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)\bar\pi^\tau_t(u_t\mid x_t,x^0_{0:t},u^0_{0:t-1})\Big] \\
&+ \mu_t^{\bar\pi^{0,\tau}_{0:t-1}}(x^0_t,x^0_{0:t-1},u^0_{0:t-1})\frac{\tau}{\tau}\Big[\mu^{\pi^{BR,\tau}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t) - \mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)\Big] \\
&\quad\cdot\Big\langle\nabla_\mu r(x_t,u_t,x^0_t,\mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}),\frac{d}{d\tau}\mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})\Big\rangle \\
&- \frac{1}{\tau}\mu_t^{\bar\pi^{0,\tau}_{0:t-1}}(x^0_t,x^0_{0:t-1},u^0_{0:t-1}) r(x_t,u_t,x^0_t,\mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}) \\
&\quad\cdot\Big[\mu^{\pi^{BR,\tau}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)\pi^{BR,\tau}_t(u_t\mid x_t,x^0_{0:t},u^0_{0:t-1}) - \mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x_t)\bar\pi^\tau_t(u_t\mid x_t,x^0_{0:t},u^0_{0:t-1})\Big]\Bigg] \\
&= \frac{1}{\tau}\tilde{\mathcal{E}}(\bar\pi^\tau,\pi^{0,BR,\tau},\bar\pi^{0,\tau}) - \frac{1}{\tau}\mathcal{E}(\bar\pi^\tau,\bar\pi^{0,\tau}) \\
&\quad+\sum_{t\in\mathcal{T}}\sum_{\substack{x_t\in\mathcal{X},u_t\in\mathcal{X},\\ x^0_{0:t}\in\mathcal{X}^{0\,t+1},\\ u^0_{0:t-1}\in\mathcal{U}^{0\,t}}} \tau\mu_t^{\bar\pi^{0,\tau}_{0:t-1}}(x^0_t,x^0_{0:t-1},u^0_{0:t-1})\frac{d}{d\tau}\mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}} \\
&\qquad\cdot\Big\langle\nabla_\mu r(x_t,u_t,x^0_t,\mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}),\frac{d}{d\tau}\mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}})\Big\rangle\Bigg] \\
&\quad- \frac{1}{\tau}\mathcal{E}(\bar\pi^\tau,\bar\pi^{0,\tau}) \le -\frac{1}{\tau}\mathcal{E}(\bar\pi^\tau,\bar\pi^{0,\tau})
\end{aligned}
$$

where we use monotonicity and Assumption 3.4.4 to obtain the last inequality.

STEP 2: FOCUSING ON THE EXPLOITABILITY OF THE MAJOR AGENT. Similarly to the case of minor agents, we can write:

$$
\begin{aligned}
&\frac{d}{d\tau}\mathcal{E}^0(\bar\pi^\tau,\bar\pi^{0,\tau}) \\
&= \frac{d}{d\tau}\Big[\max_{\pi'} J^0(\bar\pi^\tau,\pi') - J^0(\bar\pi^\tau,\bar\pi^{0,\tau})\Big] \\
&= \frac{d}{d\tau}\sum_{t\in\mathcal{T}}\sum_{x^0_{0:t}\in\mathcal{X}^{0\,t+1},u^0_{0:t}\in\mathcal{U}^{0\,t+1}}\Big[\mathbb{P}_{\bar\pi^\tau,\pi^{0,BR,\tau}}(x^0_{0:t},u^0_{0:t}) - \mathbb{P}_{\bar\pi^\tau,\bar\pi^{0,\tau}}(x^0_{0:t},u^0_{0:t})\Big] r^0(x^0_t,u^0_t,\mu^{\bar\pi^\tau_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}) \\
&= \frac{d}{d\tau}\sum_{x^0_0\in\mathcal{X}^0,u^0_0\in\mathcal{U}^0}\Big[\mathbb{P}_{\bar\pi^\tau,\pi^{0,BR,\tau}}(x^0_0,u^0_0) - \mathbb{P}_{\bar\pi^\tau,\bar\pi^{0,\tau}}(x^0_0,u^0_0)\Big] r^0(x^0_0,u^0_0,\mu_0) \\
&\quad+\frac{d}{d\tau}\sum_{x^0_{0:1}\in\mathcal{X}^{02},u^0_{0:1}\in\mathcal{U}^{02}}\Big[\mathbb{P}_{\bar\pi^\tau,\pi^{0,BR,\tau}}(x^0_{0:1},u^0_{0:1}) - \mathbb{P}_{\bar\pi^\tau,\bar\pi^{0,\tau}}(x^0_{0:1},u^0_{0:1})\Big] r^0(x^0_1,u^0_1,\mu^{\bar\pi^\tau_0}_{1|x^0_0,u^0_0}) \\
&\quad+\frac{d}{d\tau}\sum_{x^0_{0:2}\in\mathcal{X}^{03},u^0_{0:2}\in\mathcal{U}^{03}}\Big[\mathbb{P}_{\bar\pi^\tau,\pi^{0,BR,\tau}}(x^0_{0:2},u^0_{0:2}) - \mathbb{P}_{\bar\pi^\tau,\bar\pi^{0,\tau}}(x^0_{0:2},u^0_{0:2})\Big] r^0(x^0_2,u^0_2,\mu^{\bar\pi^\tau_{0:1}}_{2|x^0_{0:1},u^0_{0:1}}) \\
&\quad+\ldots
\end{aligned}
$$

Generally, at all times t, we have

$$
\frac{d}{d\tau}\sum_{x^0_{0:t}\in\mathcal{X}^{0\,t+1},u^0_{0:t}\in\mathcal{U}^{0\,t+1}}\mathbb{P}_{\bar\pi^\tau,\bar\pi^{0,\tau}}(x^0_{0:t+1},u^0_{0:t+1})r^0(x^0_{t+1},u^0_{t+1},\mu^{\bar\pi^\tau_{0:t}}_{t+1|x^0_{0:t},u^0_{0:t}})
$$

$$
\begin{aligned}
&= \frac{d}{d\tau} \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \bar{\pi}^{0,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&= \sum_{x^0_{0:t}, u^0_{0:t}} \frac{d}{d\tau} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \bar{\pi}^{0,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&\quad + \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \frac{d}{d\tau} \bar{\pi}^{0,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&\quad + \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \bar{\pi}^{0,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \frac{d}{d\tau} r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}})
\end{aligned}
$$

This leads to the desired result

$$
\begin{aligned}
&\frac{d}{d\tau} \sum_{x^0_{0:t}, u^0_{0:t}} \Big[\mathbb{P}_{\bar{\pi}^{\tau}, \pi^{0,BR,\tau}}(x^0_{0:t+1}, u^0_{0:t+1}) - \mathbb{P}_{\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}}(x^0_{0:t+1}, u^0_{0:t+1}) \Big] r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&= \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\pi^{0,BR,\tau}_{0:t-1}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \pi^{0,BR,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \frac{d}{d\tau} r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&\quad - \sum_{x^0_{0:t}, u^0_{0:t}} \frac{d}{d\tau} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \bar{\pi}^{0,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&\quad - \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \frac{d}{d\tau} \bar{\pi}^{0,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&\quad - \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \bar{\pi}^{0,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \frac{d}{d\tau} r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}})
\end{aligned}
$$

by equating the middle two terms to the exploitability terms at time $t+1$, and analyzing the first and last term as

$$
\begin{aligned}
&\sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\pi^{0,BR,\tau}_{0:t-1}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \pi^{0,BR,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \frac{d}{d\tau} r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&\quad - \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \bar{\pi}^{0,\tau}_{t+1}(u^0_{t+1} \mid x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \frac{d}{d\tau} r^0(x^0_{t+1}, u^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}) \\
&= \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\pi^{0,BR,\tau}_{0:t-1}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \Big\langle \nabla_\mu \bar{r}^0(x^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}), \frac{d}{d\tau} \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}} \Big\rangle \\
&\quad - \sum_{x^0_{0:t}, u^0_{0:t}} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \Big\langle \nabla_\mu \bar{r}^0(x^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}), \frac{d}{d\tau} \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}} \Big\rangle \\
&= \tau \sum_{x^0_{0:t}, u^0_{0:t}} \frac{d}{d\tau} \mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \Big\langle \nabla_\mu \bar{r}^0(x^0_{t+1}, \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}}), \frac{d}{d\tau} \mu^{\bar{\pi}^{\tau}_{0:t}}_{t+1|x^0_{0:t}, u^0_{0:t}} \Big\rangle \leq 0
\end{aligned}
$$

using Property D.1.3, and that the term is non-positive by Assumption 3.4.4.

Therefore, we have

$$
\frac{d}{d\tau} \mathcal{E}^0(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}) \leq -\frac{1}{\tau} \mathcal{E}^0(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}).
$$

Step 3: Combining the results. In Step 1, we showed that $\frac{d}{d\tau} \mathcal{E}(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}) \leq -\frac{1}{\tau} \mathcal{E}(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau})$ and in Step 2, we showed that $\frac{d}{d\tau} \mathcal{E}^0(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}) = -\frac{1}{\tau} \mathcal{E}^0(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau})$. Therefore we can conclude that

$$
\frac{d}{d\tau} \left(\mathcal{E}(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}) + \mathcal{E}^0(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}) \right) \leq -\frac{1}{\tau} \left(\mathcal{E}(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}) + \mathcal{E}^0(\bar{\pi}^{\tau}, \bar{\pi}^{0,\tau}) \right)
$$

which shows the total exploitability of the system is a strong Lyapunov function and therefore it converges with rate $\frac{1}{\tau}$. □

D.2 CONTINUITY OF MF DYNAMICS

In this section, we show continuity of the MF dynamics T_t^π, which will be used in the following proofs.

Lemma D.2.1. *Under Assumptions 3.4.1 and 3.4.3, the transition operator T_t^π is uniformly Lipschitz continuous with constant $L_T \coloneqq |\mathcal{X}|L_p + |\mathcal{X}||\mathcal{U}|L_\Pi + |\mathcal{X}|^2|\mathcal{U}|$.*

Proof of Lemma D.2.1. For any $(x^0, u^0, \mu), (x^{0\prime}, u^{0\prime}, \mu') \in \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X})$, we have

$$
\begin{aligned}
&\left\|T_t^\pi(x^0, u^0, \mu) - T_t^\pi(x^{0\prime}, u^{0\prime}, \mu')\right\| \\
&\quad= \sum_{x^* \in \mathcal{X}} \Bigg| \iint p(x^* \mid x, u, x^0, u^0, \mu)\pi_t(\mathrm{d}u \mid x, x^0, \mu)\mu(\mathrm{d}x) \\
&\qquad\qquad - \iint p(x^* \mid x, u, x^{0\prime}, u^{0\prime}, \mu')\pi_t(\mathrm{d}u \mid x, x^{0\prime}, \mu')\mu'(\mathrm{d}x)\Bigg| \\
&\quad\le \sum_{x^* \in \mathcal{X}} \iint \left|p(x^* \mid x, u, x^0, u^0, \mu) - p(x^* \mid x, u, x^{0\prime}, u^{0\prime}, \mu')\right| \pi_t(\mathrm{d}u \mid x, x^0, \mu)\mu(\mathrm{d}x) \\
&\qquad+ \sum_{x^* \in \mathcal{X}} \int \left| \int p(x^* \mid x, u, x^{0\prime}, u^{0\prime}, \mu') \left(\pi_t(\mathrm{d}u \mid x, x^0, \mu) - \pi_t(\mathrm{d}u \mid x, x^{0\prime}, \mu')\right)\right| \mu(\mathrm{d}x) \\
&\qquad+ \sum_{x^* \in \mathcal{X}} \left| \iint p(x^* \mid x, u, x^{0\prime}, u^{0\prime}, \mu')\pi_t(\mathrm{d}u \mid x, x^{0\prime}, \mu') \left(\mu - \mu'\right)(\mathrm{d}x)\right| \\
&\quad\le \left(|\mathcal{X}|L_p + |\mathcal{X}||\mathcal{U}|L_\Pi + |\mathcal{X}|^2|\mathcal{U}|\right) d((x^0, u^0, \mu), (x^{0\prime}, u^{0\prime}, \mu'))
\end{aligned}
$$

by Assumptions 3.4.1 and 3.4.3, where we have the distances $d((x^0, u^0, \mu), (x^{0\prime}, u^{0\prime}, \mu') = \max(\mathbf{1}_{x^0}(x^{0\prime}), \mathbf{1}_{u^0}(u^{0\prime}), \mathbf{1}_\mu(\mu'))$ as discussed in the main text. □

D.3 APPROXIMATION OF ACTION-VALUE FUNCTIONS

In this section, we give approximation lemmas for the value functions, together with definitions that were omitted in the main text, and are used in some of the following proofs.

For fixed (π, π^0), the true minor action-value function is defined by the Bellman equation

$$
\begin{aligned}
&Q_{\pi,\pi^0}(t, x, u, x^0, \mu) \\
&\quad= \sum_{u^0} \pi_t^0(u^0 \mid x^0, \mu) \Bigg[r(x, u, x^0, u^0, \mu) + \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mu) \\
&\qquad\cdot \sum_{x'} p(x' \mid x, u, x^0, u^0, \mu) \max_{u'} Q_{\pi,\pi^0}(t+1, x', u', x^{0\prime}, T_t^\pi(x^0, u^0, \mu))\Bigg],
\end{aligned}
$$

while its approximated variant follows

$$\hat{Q}_{\pi,\pi^0}(t, x, u, x^0, \mu)$$

$$= \sum_{u^0} \pi_t^0(u^0 \mid x^0, \mathrm{proj}_\delta \mu) \Big[r(x, u, x^0, u^0, \mathrm{proj}_\delta \mu) + \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta \mu)$$
$$\cdot \sum_{x'} p(x' \mid x, u, x^0, u^0, \mathrm{proj}_\delta \mu) \max_{u'} \hat{Q}_{\pi,\pi^0}(t+1, x', u', x^{0\prime}, T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu)) \Big],$$

since we have

$$\hat{Q}_{\pi,\pi^0}(t+1, x', u', x^{0\prime}, \mathrm{proj}_\delta T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu)) = \hat{Q}_{\pi,\pi^0}(t+1, x', u', x^{0\prime}, T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu))$$

by definition.

We can show that the approximate Q functions tend uniformly to the true Q functions as the δ-partition becomes fine. Here, the supremum over policies is over Π, Π^0.

Lemma D.3.1. *Under Assumptions 3.4.1, 3.4.2 and 3.4.3, we have for $\mu, \nu \in \mathcal{P}(\mathcal{X})$ at all times $t \in \mathcal{T}$ that*

$$\sup_{x^0, u^0, \pi, \pi^0} \left| \hat{Q}^0_{\pi,\pi^0}(t, x^0, u^0, \mu) - \hat{Q}^0_{\pi,\pi^0}(t, x^0, u^0, \nu) \right| = \mathcal{O}(\delta + \|\mu - \nu\|), \tag{D.3.11}$$

$$\sup_{x, u, x^0, \pi, \pi^0} \left| \hat{Q}_{\pi,\pi^0}(t, x, u, x^0, \mu) - \hat{Q}_{\pi,\pi^0}(t, x, u, x^0, \nu) \right| = \mathcal{O}(\delta + \|\mu - \nu\|). \tag{D.3.12}$$

Lemma D.3.2. *Under Assumptions 3.4.1, 3.4.2 and 3.4.3, at all times t, the approximate major and minor action-value functions uniformly converge to the true action-value functions,*

$$\sup_{x^0, u^0, \mu, \pi, \pi^0} |\hat{Q}^0_{\pi,\pi^0}(t, x^0, u^0, \mu) - Q^0_{\pi,\pi^0}(t, x^0, u^0, \mu)| = \mathcal{O}(\delta), \tag{D.3.13}$$

$$\sup_{x, u, x^0, \mu, \pi, \pi^0} |\hat{Q}_{\pi,\pi^0}(t, x, u, x^0, \mu) - Q_{\pi,\pi^0}(t, x, u, x^0, \mu)| = \mathcal{O}(\delta). \tag{D.3.14}$$

D.4 PROOF OF LEMMA D.3.1

Proof of Lemma D.3.1. At time $T-1$, we have for any $\delta > 0$ and μ, ν that by Assumption 3.4.2,

$$\begin{aligned}
&\sup_{x^0, u^0, \pi, \pi^0} \left| \hat{Q}^0_{\pi,\pi^0}(T-1, x^0, u^0, \mu) - \hat{Q}^0_{\pi,\pi^0}(T-1, x^0, u^0, \nu) \right| \\
&\quad = \sup_{x^0, u^0, \pi, \pi^0} \left| r^0(x^0, u^0, \mathrm{proj}_\delta \mu) - r^0(x^0, u^0, \mathrm{proj}_\delta \nu) \right| \\
&\quad \leq L_r(2\delta + \|\mu - \nu\|) = \mathcal{O}(\delta + \|\mu - \nu\|)
\end{aligned}$$

by triangle inequality, as the projection of μ, ν can shift μ, ν at most by δ each.

Similarly, for the induction step, assuming Eq. (D.3.11) at time $t+1$, then at time t we have:

$$\begin{aligned}
&\sup_{x^0, u^0, \pi, \pi^0} \left| \hat{Q}^0_{\pi,\pi^0}(t, x^0, u^0, \mu) - \hat{Q}^0_{\pi,\pi^0}(t, x^0, u^0, \nu) \right| \\
&\quad \leq \sup_{x^0, u^0, \pi, \pi^0} \left| r^0(x^0, u^0, \mathrm{proj}_\delta \mu) - r^0(x^0, u^0, \mathrm{proj}_\delta \nu) \right| \\
&\qquad + \sup_{x^0, u^0, \pi, \pi^0} \left| \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta \mu) \max_{u^{0\prime}} \hat{Q}^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu)) \right.
\end{aligned}$$

$$
\begin{aligned}
&\quad - \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta \nu) \max_{u^{0\prime}} \hat{Q}^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T^\pi_t(x^0, u^0, \mathrm{proj}_\delta \nu))\Big| \\
&\leq L_r(2\delta + \|\mu - \nu\|) + Q^0_{\max} |\mathcal{X}^0| L_{p^0}(2\delta + \|\mu - \nu\|) \\
&\quad + 2 \sup_{x^0,u^0,x^{0\prime},u^{0\prime},\pi,\pi^0} \Big| \hat{Q}^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T^\pi_t(x^0, u^0, \mathrm{proj}_\delta \mu)) \\
&\qquad\qquad - Q^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T^\pi_t(x^0, u^0, \mathrm{proj}_\delta \nu))\Big| \\
&= \mathcal{O}(\delta + \|\mu - \nu\|)
\end{aligned}
$$

by Assumptions 3.4.1 and 3.4.2, and induction assumption using

$$
\sup_{x^0,u^0} \left\| T^\pi_t(x^0, u^0, \mathrm{proj}_\delta \mu) - T^\pi_t(x^0, u^0, \mathrm{proj}_\delta \nu) \right\| \leq L_T(2\delta + \|\mu - \nu\|)
$$

by Lemma D.2.1. Here, $Q^0_{\max} \coloneqq T \max r^0$.

The same argument for the minor agent completes the proof for Eq. (D.3.12):

At time $T-1$, we have for any $\delta > 0$ and μ, ν that by Assumption 3.4.2, again

$$
\begin{aligned}
&\sup_{x,u,x^0,\pi,\pi^0} \left| \hat{Q}_{\pi,\pi^0}(T-1, x, u, x^0, \mu) - \hat{Q}_{\pi,\pi^0}(T-1, x, u, x^0, \nu) \right| \\
&\quad = \sup_{x,u,x^0,\pi,\pi^0} \Big| \sum_{u^0} \pi^0_{T-1}(u^0 \mid x^0, \mathrm{proj}_\delta \mu) r(x, u, x^0, u^0, \mathrm{proj}_\delta \mu) \\
&\qquad\qquad - \sum_{u^0} \pi^0_{T-1}(u^0 \mid x^0, \mathrm{proj}_\delta \nu) r(x, u, x^0, u^0, \mathrm{proj}_\delta \nu) \Big| \\
&\quad \leq |\mathcal{U}^0| L_{\Pi^0}(2\delta + \|\mu - \nu\|) \max r^0 + L_r(2\delta + \|\mu - \nu\|) = \mathcal{O}(\delta + \|\mu - \nu\|).
\end{aligned}
$$

For the induction step, assuming Eq. (D.3.12) at time $t+1$, then at time t we have:

$$
\begin{aligned}
&\sup_{x,u,x^0,\pi,\pi^0} \left| \hat{Q}_{\pi,\pi^0}(t, x, u, x^0, \mu) - \hat{Q}_{\pi,\pi^0}(t, x, u, x^0, \nu) \right| \\
&\leq \sup_{x,u,x^0,\pi,\pi^0} \Big| \sum_{u^0} \pi^0_t(u^0 \mid x^0, \mathrm{proj}_\delta \mu) r(x, u, x^0, u^0, \mathrm{proj}_\delta \mu) \\
&\qquad\qquad - \sum_{u^0} \pi^0_t(u^0 \mid x^0, \mathrm{proj}_\delta \nu) r(x, u, x^0, u^0, \mathrm{proj}_\delta \nu) \Big| \\
&\quad + \sup_{x,u,x^0,\pi,\pi^0} \Big| \sum_{u^0} \pi^0_t(u^0 \mid x^0, \mathrm{proj}_\delta \mu) \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta \mu) \\
&\qquad\qquad \cdot \sum_{x'} p(x' \mid x, u, x^0, u^0, \mathrm{proj}_\delta \mu) \max_{u'} \hat{Q}_{\pi,\pi^0}(t+1, x', u', x^{0\prime}, T^\pi_t(x^0, u^0, \mathrm{proj}_\delta \mu)) \\
&\qquad - \sum_{u^0} \pi^0_t(u^0 \mid x^0, \mathrm{proj}_\delta \mu) \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta \nu) \\
&\qquad\qquad \cdot \sum_{x'} p(x' \mid x, u, x^0, u^0, \mathrm{proj}_\delta \mu) \max_{u'} \hat{Q}_{\pi,\pi^0}(t+1, x', u', x^{0\prime}, T^\pi_t(x^0, u^0, \mathrm{proj}_\delta \nu)) \Big| \\
&\leq |\mathcal{U}^0| L_{\Pi^0}(2\delta + \|\mu - \nu\|) \max r^0 + L_r(2\delta + \|\mu - \nu\|) + |\mathcal{U}^0| Q_{\max} L_{\Pi^0}(2\delta + \|\mu - \nu\|)
\end{aligned}
$$

$$
\begin{aligned}
&+ |\mathcal{X}^0| Q_{\max} L_{p^0}(2\delta + \|\mu - \nu\|) + |\mathcal{X}| Q_{\max} L_p(2\delta + \|\mu - \nu\|) \\
&+ 2 \sup_{x^0,u^0,x',u',x^{0\prime},\pi,\pi^0} \Big| \hat{Q}_{\pi,\pi^0}(t+1, x', u', x^{0\prime}, T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu)) \\
&\qquad - Q^0_{\pi,\pi^0}(t+1, x', u', x^{0\prime}, T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \nu)) \Big| \\
&= \mathcal{O}(\delta + \|\mu - \nu\|)
\end{aligned}
$$

by Assumptions 3.4.1, 3.4.2 and 3.4.3, and applying the induction assumption on the last term, where we again use $\sup_{x^0,u^0} \left\| T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu) - T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \nu) \right\| \leq L_T(2\delta + \|\mu - \nu\|)$ by Lemma D.2.1. Here, $Q_{\max} \coloneqq T \max r$. This completes the proof for Eq. (D.3.12). □

D.5 PROOF OF LEMMA D.3.2

Proof of Lemma D.3.2. The proof is by (reverse) induction. At terminal time $t = T - 1$, we have by Assumption 3.4.2

$$
\begin{aligned}
&\sup_{x^0,u^0,\mu,\pi,\pi^0} \left| \hat{Q}^0_{\pi,\pi^0}(T-1, x^0, u^0, \mu) - Q^0_{\pi,\pi^0}(T-1, x^0, u^0, \mu) \right| \\
&\quad = \sup_{x^0,u^0,\mu,\pi,\pi^0} \left| r^0(x^0, u^0, \mathrm{proj}_\delta \mu) - r^0(x^0, u^0, \mu) \right| \leq L_r \delta
\end{aligned}
$$

Assume Eq. (D.3.13) holds at time $t+1$, then at time t we have

$$
\begin{aligned}
&\sup_{x^0,u^0,\mu,\pi,\pi^0} \left| \hat{Q}^0_{\pi,\pi^0}(t, x^0, u^0, \mu) - Q^0_{\pi,\pi^0}(t, x^0, u^0, \mu) \right| \\
&\leq \sup_{x^0,u^0,\mu,\pi,\pi^0} \left| r^0(x^0, u^0, \mathrm{proj}_\delta \mu) - r^0(x^0, u^0, \mu) \right| \\
&\quad + \sup_{x^0,u^0,\mu,\pi,\pi^0} \Bigg| \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta \mu) \max_{u^{0\prime}} \hat{Q}^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu)) \\
&\qquad\qquad - \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mu) \max_{u^{0\prime}} Q^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T_t^\pi(x^0, u^0, \mu)) \Bigg| \\
&\leq L_r \delta + Q^0_{\max} |\mathcal{X}^0| L_{p^0} \delta + \mathcal{O}(\delta) \\
&\quad + 2 \sup_{x^0,u^0,\mu,x^{0\prime},u,\pi,\pi^0} \Big| \hat{Q}^0_{\pi,\pi^0}(t+1, x^{0\prime}, u, T_t^\pi(x^0, u^0, \mu)) \\
&\qquad\qquad - Q^0_{\pi,\pi^0}(t+1, x^{0\prime}, u, T_t^\pi(x^0, u^0, \mu)) \Big| = \mathcal{O}(\delta)
\end{aligned}
$$

by Assumptions 3.4.1 and 3.4.2, the estimate from Lemma D.3.1 with $|T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu) - T_t^\pi(x^0, u^0, \mu)| \leq L_T \delta = \mathcal{O}(\delta)$ by Lemma D.2.1, and the induction assumption for the final term. Here, $Q^0_{\max} \coloneqq T \max r^0$. This completes the proof for Eq. (D.3.13).

For the minor agent in Eq. (D.3.14), we have by the same argument

$$
\begin{aligned}
&\sup_{x,u,x^0,\mu,\pi,\pi^0} \left| \hat{Q}_{\pi,\pi^0}(T-1, x, u, x^0, \mu) - Q_{\pi,\pi^0}(T-1, x, u, x^0, \mu) \right| \\
&\quad \leq \sup_{x,u,x^0,u^0,\mu,\pi,\pi^0} \left| r(x, u, x^0, u^0, \mathrm{proj}_\delta \mu) - r(x, u, x^0, u^0, \mu) \right| \leq L_r \delta
\end{aligned}
$$

at terminal time $T-1$, and then inductively at any time t

$$
\begin{aligned}
&\sup_{x,u,x^0,\mu,\pi,\pi^0} \left|\hat{Q}_{\pi,\pi^0}(t,x,u,x^0,\mu) - Q_{\pi,\pi^0}(t,x,u,x^0,\mu)\right| \\
&\quad\le \sup_{x,u,x^0,u^0,\mu,\pi,\pi^0} \left|r(x,u,x^0,u^0,\mathrm{proj}_\delta\mu) - r(x,u,x^0,u^0,\mu)\right| \\
&\qquad + \sup_{x^0,u^0,\mu,\pi,\pi^0} \Bigg| \sum_{u^0} \pi_t^0(u^0 \mid x^0, \mathrm{proj}_\delta\mu) \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta\mu) \\
&\qquad\qquad \cdot \sum_{x'} p(x' \mid x,u,x^0,u^0,\mathrm{proj}_\delta\mu) \max_{u'} \hat{Q}_{\pi,\pi^0}(t+1,x',u',x^{0\prime},T_t^\pi(x^0,u^0,\mathrm{proj}_\delta\mu)) \\
&\qquad\quad - \sum_{u^0} \pi_t^0(u^0 \mid x^0,\mu) \sum_{x^{0\prime}} p^0(x^{0\prime} \mid x^0,u^0,\mu) \\
&\qquad\qquad \cdot \sum_{x'} p(x' \mid x,u,x^0,u^0,\mu) \max_{u'} Q_{\pi,\pi^0}(t+1,x',u',x^{0\prime},T_t^\pi(x^0,u^0,\mu)) \Bigg| \\
&\quad\le L_r\delta + Q_{\max}|\mathcal{U}^0| L_{\Pi^0}\delta + Q_{\max}|\mathcal{X}^0| L_{p^0}\delta + Q_{\max}|\mathcal{X}| L_p\delta + \mathcal{O}(\delta) \\
&\qquad + \sup_{\mu,x^{0\prime},u,\pi,\pi^0} \Big|\hat{Q}_{\pi,\pi^0}(t+1,x',u',x^{0\prime},T_t^\pi(x^0,u^0,\mu)) \\
&\qquad\qquad - Q_{\pi,\pi^0}(t+1,x',u',x^{0\prime},T_t^\pi(x^0,u^0,\mu))\Big| = \mathcal{O}(\delta)
\end{aligned}
$$

using also Assumption 3.4.3 and $Q_{\max} \coloneqq T \max r$, which completes the proof by induction. □

D.6 PROOF OF THEOREM 3.4.1

Proof of Theorem 3.4.1. For readability, we abbreviate the states and actions at time t as $\mathcal{J}_t \coloneqq (x_t^{0,N}, u_t^{0,N}, x_t^{1,N}, u_t^{1,N}, \ldots, x_t^{N,N}, u_t^{N,N})$, and write $\mathbb{E}_{\mathcal{J}_t}$ for the conditional expectation given $\mathcal{J}_t$. Without loss of generality, we show the statements for families $\mathcal{F}$ that are additionally uniformly bounded by some constant M_f, since the support of $f \in \mathcal{F}$ is compact and we can add any constant to f without changing the difference between expectations in Eq. (3.4.53). We also define the conditional expectation of the empirical MF at time $t+1$ given variables at time t,

$$\hat{\mu}_{t+1}^N \coloneqq T_t^\pi(x_t^{0,N}, u_t^{0,N}, \mu_t^N).$$

We show Eq. (3.4.53) at all times by induction. At time $t=0$, the statement follows from a LLN, see also below. Assuming that Eq. (3.4.53) holds at time t, then at time $t+1$ we have

$$
\begin{aligned}
&\sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(x_{t+1}^{1,N}, u_{t+1}^{1,N}, x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) - f(x_{t+1}, u_{t+1}, x_{t+1}^0, u_{t+1}^0, \mu_{t+1})\right]\right| \\
&\quad\le \sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(x_{t+1}^{1,N}, u_{t+1}^{1,N}, x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) - f(x_{t+1}^{1,N}, u_{t+1}^{1,N}, x_{t+1}^{0,N}, u_{t+1}^{0,N}, \hat{\mu}_{t+1}^N)\right]\right| \qquad \text{(D.6.15)} \\
&\quad + \sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\int f(x_{t+1}^{1,N}, u_{t+1}^{1,N}, x_{t+1}^{0,N}, u^0, \hat{\mu}_{t+1}^N) \pi_{t+1}^0(\mathrm{d}u^0 \mid x_{t+1}^{0,N}, \mu_{t+1}^N)\right]\right. \\
&\qquad\qquad \left. - \mathbb{E}\left[\int f(x_{t+1}^{1,N}, u_{t+1}^{1,N}, x_{t+1}^{0,N}, u^0, \hat{\mu}_{t+1}^N) \pi_{t+1}^0(\mathrm{d}u^0 \mid x_{t+1}^{0,N}, \hat{\mu}_{t+1}^N)\right]\right| \qquad \text{(D.6.16)} \\
&\quad + \sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\iint f(x_{t+1}^{1,N}, u, x_{t+1}^{0,N}, u^0, \hat{\mu}_{t+1}^N) \pi_{t+1}^0(\mathrm{d}u^0 \mid x_{t+1}^{0,N}, \hat{\mu}_{t+1}^N) \hat{\pi}_{t+1}(\mathrm{d}u \mid x_{t+1}^{1,N}, x_{t+1}^{0,N}, \mu_{t+1}^N)\right]\right. \\
&\qquad\qquad \left. - \mathbb{E}\left[\iint f(x_{t+1}^{1,N}, u, x_{t+1}^{0,N}, u^0, \hat{\mu}_{t+1}^N) \pi_{t+1}^0(\mathrm{d}u^0 \mid x_{t+1}^{0,N}, \hat{\mu}_{t+1}^N) \hat{\pi}_{t+1}(\mathrm{d}u \mid x_{t+1}^{1,N}, x_{t+1}^{0,N}, \hat{\mu}_{t+1}^N)\right]\right| \qquad \text{(D.6.17)}
\end{aligned}
$$

$$+ \sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \left| \mathbb{E}\left[\iint f(x_{t+1}^{1,N}, u, x_{t+1}^{0,N}, u^0, \hat{\mu}_{t+1}^N) \pi_{t+1}^0(\mathrm{d}u^0 \mid x_{t+1}^{0,N}, \hat{\mu}_{t+1}^N) \hat{\pi}_{t+1}(\mathrm{d}u \mid x_{t+1}^{1,N}, x_{t+1}^{0,N}, \hat{\mu}_{t+1}^N) \right] \right.$$
$$\left. - \mathbb{E}\left[\iint f(x_{t+1}, u, x_{t+1}^0, u^0, \mu_{t+1}) \pi_{t+1}^0(\mathrm{d}u^0 \mid x_{t+1}^0, \mu_{t+1}) \hat{\pi}_{t+1}(\mathrm{d}u \mid x_{t+1}, x_{t+1}^0, \mu_{t+1}) \right] \right| \tag{D.6.18}$$

The first term (D.6.15) is

$$\begin{aligned}
&\sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \left| \mathbb{E}\left[f(x_{t+1}^{1,N}, u_{t+1}^{1,N}, x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) - f(x_{t+1}^{1,N}, u_{t+1}^{1,N}, x_{t+1}^{0,N}, u_{t+1}^{0,N}, \hat{\mu}_{t+1}^N) \right] \right| \\
&\quad \le \sup_{\hat{\pi},\pi,\pi^0} L_f \mathbb{E}\left[\left\| \mu_{t+1}^N - \hat{\mu}_{t+1}^N \right\| \right] \\
&\quad = \sup_{\hat{\pi},\pi,\pi^0} L_f \mathbb{E}\left[\sum_{x\in\mathcal{X}} \left| \mu_{t+1}^N(x) - \hat{\mu}_{t+1}^N(x) \right| \right] \\
&\quad = \sup_{\hat{\pi},\pi,\pi^0} L_f \sum_{x\in\mathcal{X}} \mathbb{E}\left[\left| \frac{1}{N} \sum_{i=1}^N \mathbf{1}_x(x_{t+1}^{i,N}) - \frac{1}{N} \sum_{i=2}^N \mathbf{1}_x(x_{t+1}^{i,N}) \right| \right] \\
&\qquad + \sup_{\hat{\pi},\pi,\pi^0} L_f \sum_{x\in\mathcal{X}} \mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\left| \frac{1}{N} \sum_{i=2}^N \mathbf{1}_x(x_{t+1}^{i,N}) - \mathbb{E}_{\mathcal{J}_t}\left[\frac{1}{N} \sum_{i=2}^N \mathbf{1}_x(x_{t+1}^{i,N}) \right] \right| \right] \right] \\
&\qquad + \sup_{\hat{\pi},\pi,\pi^0} L_f \sum_{x\in\mathcal{X}} \mathbb{E}\left[\left| \frac{1}{N} \sum_{u\in\mathcal{U}} p(x \mid x_t^{1,N}, u, x_t^{0,N}, u_t^{0,N}, \mu_t^N) \pi_t(u \mid x_t^{1,N}, x_t^{0,N}, \mu_t^N) \right| \right] \\
&\quad \le L_f |\mathcal{X}| \left(\frac{1}{N} + \sqrt{\frac{4}{N}} + \frac{|\mathcal{U}|}{N} \right) = \mathcal{O}(1/\sqrt{N})
\end{aligned}$$

and tends to zero at rate $\mathcal{O}(1/\sqrt{N})$, where the last term is the difference between $\mathbb{E}_{\mathcal{J}_t}\left[\frac{1}{N} \sum_{i=2}^N \mathbf{1}_x(x_{t+1}^{i,N}) \right]$ and $\hat{\mu}_{t+1}^N(x) = \mathbb{E}_{\mathcal{J}_t}\left[\frac{1}{N} \sum_{i=1}^N \mathbf{1}_x(x_{t+1}^{i,N}) \right]$, while the middle term is obtained by tower rule and analyzed by a weak LLN argument, i.e.

$$\begin{aligned}
&\sup_{\hat{\pi},\pi,\pi^0} L_f \sum_{x\in\mathcal{X}} \mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\left| \frac{1}{N} \sum_{i=2}^N \mathbf{1}_x(x_{t+1}^{i,N}) - \mathbb{E}_{\mathcal{J}_t}\left[\frac{1}{N} \sum_{i=2}^N \mathbf{1}_x(x_{t+1}^{i,N}) \right] \right| \right] \right] \\
&\quad = \sup_{\hat{\pi},\pi,\pi^0} L_f \sum_{x\in\mathcal{X}} \mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\left| \frac{1}{N} \sum_{i=2}^N \left(\mathbf{1}_x(x_{t+1}^{i,N}) - \mathbb{E}_{\mathcal{J}_t}\left[\mathbf{1}_x(x_{t+1}^{i,N}) \right] \right) \right| \right] \right] \\
&\quad \le \sup_{\hat{\pi},\pi,\pi^0} L_f \sum_{x\in\mathcal{X}} \mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\left(\frac{1}{N} \sum_{i=2}^N \left(\mathbf{1}_x(x_{t+1}^{i,N}) - \mathbb{E}_{\mathcal{J}_t}\left[\mathbf{1}_x(x_{t+1}^{i,N}) \right] \right) \right)^2 \right] \right]^{1/2} \\
&\quad = \sup_{\hat{\pi},\pi,\pi^0} L_f \sum_{x\in\mathcal{X}} \mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\frac{1}{N} \sum_{i=2}^N \left(\mathbf{1}_x(x_{t+1}^{i,N}) - \mathbb{E}_{\mathcal{J}_t}\left[\mathbf{1}_x(x_{t+1}^{i,N}) \right] \right)^2 \right] \right]^{1/2} \\
&\quad \le L_f |\mathcal{X}| \sqrt{\frac{N-1}{N^2} \cdot 2^2} \le L_f |\mathcal{X}| \sqrt{\frac{4}{N}}
\end{aligned}$$

by conditional independence of $x_{t+1}^{i,N}$ given $\mathcal{J}_t$.

Similarly, for the second term (D.6.16) we obtain

$$\begin{aligned}\sup_{\hat{\pi},\pi,\pi^0}\sup_{f\in\mathcal{F}}\Bigg|&\mathbb{E}\left[\int f(x_{t+1}^{1,N},u_{t+1}^{1,N},x_{t+1}^{0,N},u^0,\hat{\mu}_{t+1}^N)\pi_{t+1}^0(\mathrm{d}u^0\mid x_{t+1}^{0,N},\mu_{t+1}^N)\right]\\&-\mathbb{E}\left[\int f(x_{t+1}^{1,N},u_{t+1}^{1,N},x_{t+1}^{0,N},u^0,\hat{\mu}_{t+1}^N)\pi_{t+1}^0(\mathrm{d}u^0\mid x_{t+1}^{0,N},\hat{\mu}_{t+1}^N)\right]\Bigg|\\&\leq|\mathcal{U}^0|M_fL_{\Pi_0}\mathbb{E}\left[\left\|\mu_{t+1}^N-\hat{\mu}_{t+1}^N\right\|\right]=\mathcal{O}(1/\sqrt{N})\end{aligned}$$

by Assumption 3.4.3, and also for the third term (D.6.17),

$$\begin{aligned}\sup_{\hat{\pi},\pi,\pi^0}\sup_{f\in\mathcal{F}}\Bigg|&\mathbb{E}\left[\iint f(x_{t+1}^{1,N},u,x_{t+1}^{0,N},u^0,\hat{\mu}_{t+1}^N)\pi_{t+1}^0(\mathrm{d}u^0\mid x_{t+1}^{0,N},\hat{\mu}_{t+1}^N)\hat{\pi}_{t+1}(\mathrm{d}u\mid x_{t+1}^{1,N},x_{t+1}^{0,N},\mu_{t+1}^N)\right]\\&-\mathbb{E}\left[\iint f(x_{t+1}^{1,N},u,x_{t+1}^{0,N},u^0,\hat{\mu}_{t+1}^N)\pi_{t+1}^0(\mathrm{d}u^0\mid x_{t+1}^{0,N},\hat{\mu}_{t+1}^N)\hat{\pi}_{t+1}(\mathrm{d}u\mid x_{t+1}^{1,N},x_{t+1}^{0,N},\hat{\mu}_{t+1}^N)\right]\Bigg|\\&\leq|\mathcal{U}|M_fL_{\Pi}\mathbb{E}\left[\left\|\mu_{t+1}^N-\hat{\mu}_{t+1}^N\right\|\right]=\mathcal{O}(1/\sqrt{N}).\end{aligned}$$

For the last term (D.6.18), we have

$$\begin{aligned}&\sup_{\hat{\pi},\pi,\pi^0}\sup_{f\in\mathcal{F}}\Bigg|\mathbb{E}\left[\iint f(x_{t+1}^{1,N},u,x_{t+1}^{0,N},u^0,\hat{\mu}_{t+1}^N)\pi_{t+1}^0(\mathrm{d}u^0\mid x_{t+1}^{0,N},\hat{\mu}_{t+1}^N)\hat{\pi}_{t+1}(\mathrm{d}u\mid x_{t+1}^{1,N},x_{t+1}^{0,N},\hat{\mu}_{t+1}^N)\right]\\&\qquad-\mathbb{E}\left[\iint f(x_{t+1},u_{t+1},x_{t+1}^0,u_{t+1}^0,\mu_{t+1})\pi_{t+1}^0(\mathrm{d}u^0\mid x_{t+1}^0,\mu_{t+1})\hat{\pi}_{t+1}(\mathrm{d}u\mid x_{t+1},x_{t+1}^0,\mu_{t+1})\right]\Bigg|\\&=\sup_{\hat{\pi},\pi,\pi^0}\sup_{f\in\mathcal{F}}\Bigg|\mathbb{E}\left[\iint g_2(x,x^0,x_t^{1,N},u_t^{1,N},x_t^{0,N},u_t^{0,N},\mu_t^N)\right.\\&\qquad\left.p^0(\mathrm{d}x^0\mid x_t^{0,N},u_t^{0,N},\mu_t^N)p(\mathrm{d}x\mid x_t^{1,N},u_t^{1,N},x_t^{0,N},u_t^{0,N},\mu_t^N)\right]\\&\qquad-\mathbb{E}\left[\iiiint g_2(x,x^0,x_t,u_t,x_t^0,u_t^0,\mu_t)p^0(\mathrm{d}x^0\mid x_t^0,u_t^0,\mu_t)p(\mathrm{d}x\mid x_t,u_t,x_t^0,u_t^0,\mu_t)\right]\Bigg|\end{aligned}$$

where we define

$$\begin{aligned}&g_2(x,x^0,x_t,u_t,x_t^0,u_t^0,\mu_t)\\&\quad\coloneqq\iint f(x,u,x^0,u^0,T_t^\pi(x_t^0,u_t^0,\mu_t))\pi_{t+1}^0(\mathrm{d}u^0\mid x^0,T_t^\pi(x_t^0,u_t^0,\mu_t))\hat{\pi}_{t+1}(\mathrm{d}u\mid x,x^0,T_t^\pi(x_t^0,u_t^0,\mu_t)).\end{aligned}$$

We show that the terms inside the expectations are Lipschitz in $(x_t^{1,N},u_t^{1,N},x_t^{0,N},u_t^{0,N},\mu_t^N)$ and $(x_t,u_t,x_t^0,u_t^0,\mu_t)$, which will imply convergence of the second term at rate $\mathcal{O}(1/\sqrt{N})$ by the induction assumption, completing the proof of Eq. (3.4.53).

First, note that T_t^π is Lipschitz with constant L_T by Lemma D.2.1. Therefore, the map $(x,u,x^0,u^0,x_t^0,u_t^0,\mu_t)\mapsto f(x,u,x^0,u^0,T_t^\pi(x_t^0,u_t^0,\mu_t))$ is also Lipschitz with constant L_fL_T. We similarly iteratively obtain Lipschitzness of the maps

$$\begin{aligned}g_1(x,u,x^0,x_t,u_t,x_t^0,u_t^0,\mu_t)&\coloneqq\int f(x,u,x^0,u^0,T_t^\pi(x_t^0,u_t^0,\mu_t))\pi_{t+1}^0(\mathrm{d}u^0\mid x^0,T_t^\pi(x_t^0,u_t^0,\mu_t))\\g_2(x,x^0,x_t,u_t,x_t^0,u_t^0,\mu_t)&\coloneqq\int g_1(x,u,x^0,x_t,u_t,x_t^0,u_t^0,\mu_t)\hat{\pi}_{t+1}(\mathrm{d}u\mid x,x^0,T_t^\pi(x_t^0,u_t^0,\mu_t))\\g_3(x,x_t,u_t,x_t^0,u_t^0,\mu_t)&\coloneqq\int g_2(x,x^0,x_t,u_t,x_t^0,u_t^0,\mu_t)p^0(\mathrm{d}x^0\mid x_t^0,u_t^0,\mu_t)\\g_4(x_t,u_t,x_t^0,u_t^0,\mu_t)&\coloneqq\int g_3(x,x_t,u_t,x_t^0,u_t^0,\mu_t)p(\mathrm{d}x\mid x_t,u_t,x_t^0,u_t^0,\mu_t)\end{aligned}$$

with Lipschitz constants $L_{g_1} = L_f L_T + |\mathcal{U}^0| M_f L_{\Pi^0} L_T$, $L_{g_2} = L_{g_1} + |\mathcal{U}| M_f L_\Pi L_T$, $L_{g_3} = L_{g_2} + |\mathcal{X}^0| M_f L_{p^0}$, $L_{g_4} = L_{g_3} + |\mathcal{X}| M_f L_p$, and finally note that the last term (D.6.18) is equal to

$$\sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \left| \mathbb{E}\left[g_4(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N) - g_4(x_t, u_t, x_t^0, u_t^0, \mu_t) \right] \right| = \mathcal{O}(1/\sqrt{N}),$$

tending to zero by applying the induction assumption to families of L_{g_4}-Lipschitz functions. □

Proof of Corollary 3.4.1. The result follows immediately from Theorem 3.4.1 by noting that $\mathcal{F}^0 \subseteq \mathcal{F}$ when considering functions in $\mathcal{F}^0$ as constant functions over the deviating minor agent's variables in $\mathcal{F}$. □

D.7 PROOF OF COROLLARY 3.4.2

Proof of Corollary 3.4.2. Under (π, π^0), we have for any $\varepsilon > 0$ that there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\begin{aligned}
&\sup_{\hat{\pi}\in\Pi} \left| J_N^1((\hat{\pi}, \pi, \dots, \pi), \pi^0) - J(\hat{\pi}, \pi, \pi^0) \right| \\
&\quad \leq \sup_{\hat{\pi}\in\Pi} \left| \mathbb{E}\left[\sum_{t\in\mathcal{T}} r(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N) \right] - \mathbb{E}\left[\sum_{t\in\mathcal{T}} r(x_t, u_t, x_t^0, u_t^0, \mu_t) \right] \right| \\
&\quad = \sup_{\hat{\pi}\in\Pi} \left| \sum_{t\in\mathcal{T}} \mathbb{E}\left[r(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N) - r(x_t, u_t, x_t^0, u_t^0, \mu_t) \right] \right| \to 0
\end{aligned}$$

by Theorem 3.4.1 and Assumption 3.4.2.

Therefore, using the previous paragraph, for any $\varepsilon > 0$ we also have

$$\begin{aligned}
&\sup_{\hat{\pi}\in\Pi} \left(J_1^N((\hat{\pi}, \dots, \pi), \pi^0) - J_1^N((\pi, \dots, \pi), \pi^0) \right) \\
&\quad \leq \sup_{\hat{\pi}\in\Pi} \left(J_1^N((\hat{\pi}, \pi, \dots, \pi), \pi^0) - J(\hat{\pi}, \pi, \pi^0) \right) \\
&\qquad + \sup_{\hat{\pi}\in\Pi} \left(J(\hat{\pi}, \pi, \pi^0) - J(\pi, \pi, \pi^0) \right) \\
&\qquad + \left(J(\pi, \pi, \pi^0) - J_1^N((\pi, \dots, \pi), \pi^0) \right) \\
&\quad < \frac{\varepsilon}{2} + 0 + \frac{\varepsilon}{2} = \varepsilon
\end{aligned}$$

for N large enough by definition of M3FNE, which is the desired statement for $i = 1$. By symmetry, this applies to all $i \geq 1$.

In the alternate infinite-horizon case with discount $\gamma \in (0, 1)$ and $\mathcal{T} := \mathbb{N}$, we first have for any $\varepsilon > 0$ that there exists $N' \in \mathbb{N}$ such that for all $N > N'$ we have

$$\begin{aligned}
&\sup_{\hat{\pi}\in\Pi} \left| J_N^1((\hat{\pi}, \pi, \dots, \pi), \pi^0) - J(\hat{\pi}, \pi, \pi^0) \right| \\
&\quad \leq \sup_{\hat{\pi}\in\Pi} \left| \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N) \right] - \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, u_t, x_t^0, u_t^0, \mu_t) \right] \right| \\
&\quad = \sup_{\hat{\pi}\in\Pi} \left| \sum_{t=0}^{T} \gamma^t \mathbb{E}\left[r(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N) - r(x_t, u_t, x_t^0, u_t^0, \mu_t) \right] \right| + \frac{\varepsilon}{2} \to 0
\end{aligned}$$

by choosing T large enough, and then applying Theorem 3.4.1.

An analogous argument for the major agent completes the proof, as

$$\begin{aligned}
&\sup_{\hat{\pi}\in\Pi} \left| J_N^0((\hat{\pi}, \pi, \dots, \pi), \pi^0) - J^0(\hat{\pi}, \pi, \pi^0) \right| \\
&\quad \leq \sup_{\hat{\pi}\in\Pi} \left| \mathbb{E}\left[\sum_{t\in\mathcal{T}} r^0(x_t^{0,N}, u_t^{0,N}, \mu_t^N) \right] - \mathbb{E}\left[\sum_{t\in\mathcal{T}} r(x_t, u_t, x_t^0, u_t^0, \mu_t) \right] \right| \\
&\quad = \sup_{\hat{\pi}\in\Pi} \left| \sum_{t\in\mathcal{T}} \mathbb{E}\left[r^0(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - r(x_t, u_t, x_t^0, u_t^0, \mu_t) \right] \right| \to 0
\end{aligned}$$

by Corollary 3.4.1 and Assumption 3.4.2. □

D.8 PROOF OF THEOREM 3.4.3

Proof of Theorem 3.4.3. Observe that the convergence of approximate minor objectives to the true objectives as $\delta \to 0$,

$$\sup_{x,x^0,\mu,\pi,\pi^0} \left| \hat{V}_{\pi,\pi^0}^{\pi}(t, x, x^0, \mu) - V_{\pi,\pi^0}^{\pi}(t, x, x^0, \mu) \right| = \mathcal{O}(\delta) \to 0 \tag{D.8.19}$$

at all times t, follows by the same arguments as in Lemma D.3.2. The only difference is that we estimate one more term from policy evaluation instead of the $\max$ operation, using continuity for π by Assumption 3.4.3.

Therefore, the approximate minor objective converges as desired,

$$\begin{aligned}
\hat{J}(\pi, \pi^0) &\coloneqq \sum_{x,x^0} \mu_0(x)\mu_0^0(x^0)\hat{V}_{\pi,\pi^0}^{\pi}(0, x, x^0, \mu_0) \\
&\to \sum_{x,x^0} \mu_0(x)\mu_0^0(x^0)V_{\pi,\pi^0}^{\pi}(0, x, x^0, \mu_0) \\
&= J(\pi, \pi^0).
\end{aligned}$$

On the other hand, for the approximate exploitability of the minor agent

$$\hat{\mathcal{E}}(\pi, \pi^0) = \sum_{x,x^0} \mu_0(x)\mu_0^0(x^0) \left(\max_{\hat{\pi}'\in\hat{\Pi}} \hat{V}_{\pi,\pi^0}^{\hat{\pi}'}(0, x, x^0, \mu_0)) - \hat{V}_{\pi,\pi^0}^{\pi}(0, x, x^0, \mu_0) \right)$$

and its true exploitability

$$\mathcal{E}(\pi, \pi^0) = \sum_{x,x^0} \mu_0(x)\mu_0^0(x^0) \left(\max_{\pi'\in\Pi} V_{\pi,\pi^0}^{\pi'}(0, x, x^0, \mu_0)) - V_{\pi,\pi^0}^{\pi}(0, x, x^0, \mu_0) \right)$$

we first note that $\max_{\hat{\pi}'\in\hat{\Pi}} \hat{V}_{\pi,\pi^0}^{\hat{\pi}'} = \hat{V}_{\pi,\pi^0}$ and $\max_{\pi'\in\Pi} V_{\pi,\pi^0}^{\pi'} = V_{\pi,\pi^0}$ [69, Theorem 3.2.1 and Condition 3.3.4], where the approximate and true optimal value functions are defined by the maximum of the action-value functions over actions,

$$\hat{V}_{\pi,\pi^0}(t, x, x^0, \mu) \coloneqq \max_u \hat{Q}_{\pi,\pi^0}(t, x, u, x^0, \mu), \quad V_{\pi,\pi^0}(t, x, x^0, \mu) \coloneqq \max_u Q_{\pi,\pi^0}(t, x, u, x^0, \mu).$$

Therefore, we obtain

$$\begin{aligned}&\hat{\mathcal{E}}(\pi,\pi^0)-\mathcal{E}(\pi,\pi^0)\\&\quad=\sum_{x,x^0}\mu_0(x)\mu_0^0(x^0)\left(\hat{V}_{\pi,\pi^0}(0,x,x^0,\mu_0)-V_{\pi,\pi^0}(0,x,x^0,\mu_0)\right)\\&\qquad+\sum_{x,x^0}\mu_0(x)\mu_0^0(x^0)\left(V^{\pi}_{\pi,\pi^0}(0,x,x^0,\mu_0)-\hat{V}^{\pi}_{\pi,\pi^0}(0,x,x^0,\mu_0)\right)=\mathcal{O}(\delta)\to 0\end{aligned}$$

where the first term is estimated by Lemma D.3.2, and similarly the second by Eq. (D.8.19).

Analogous arguments for the major agent complete the proof. □

D.9 ADDITIONAL EXPERIMENTAL DETAILS

In the following, we give a detailed description of the problems considered in evaluation. For initialization of policies, unless mentioned, we use the policy that always picks the first action, in order of definition in the main text. We run our experiments each on a single core of an Intel Xeon Platinum 9242 CPU with 4 GB of memory (RedHat 8.8, and without GPUs). We used around 20 000 CPU core hours. The code is based on Python 3.9 and NumPy 1.23.4 [382], and can be found in the supplementary material.

SIS EPIDEMICS MODEL. Formally, minor agents have states $\mathcal{X} := \{S, I\}$ for susceptible (S) and infected (I), and can choose between actions $\mathcal{U} := \{P, \bar{P}\}$ for prevention (P) and no prevention ($\bar{P}$). The major agent has states $\mathcal{X}^0 := \{H, L\}$ for high (H) and low (L) transmissibility regimes (e.g. from seasonal changes or virus mutations), and actions $\mathcal{U}^0 := \{F, \bar{F}\}$ for forcing (F) preventative actions, or not ($\bar{F}$). The minor dynamics are then given by

$$\begin{aligned}p(I \mid S, \bar{P}, x^0, u^0, \mu_t) &= (0.5 + \mathbf{1}_H(x^0) + \mathbf{1}_{\bar{F}}(u^0))\alpha\mu_t(I)\Delta t,\\ p(I \mid S, P, \ldots) = 0, \quad p(S \mid I, \ldots) &= \beta\Delta t\end{aligned}$$

for transmissibility $\alpha > 0$, recovery rate $\beta > 0$ and step size $\Delta t > 0$. The major dynamics are exogeneous and given by

$$p^0(H \mid L, \ldots) = p^0(L \mid H, \ldots) = \alpha^0\Delta t$$

for rate $\alpha^0 > 0$. Lastly, the reward functions will be set as

$$\begin{aligned}r(x, u, x^0, u^0, \mu) &= -c_I\mathbf{1}_I(x) - c_P\mathbf{1}_P(u)(\mathbf{1}_F(u^0) + 0.5),\\ r^0(x^0, u^0, \mu) &= -c^0_\mu\mu(I) - c^0_F\mathbf{1}_F(u^0)(0.5 - \mu(I)),\end{aligned}$$

i.e. the major agent wants to keep infections low via preventative actions, while the minor agents are interested only in their own infection, trading off between infection and costly prevention. The major government agent has an reputation cost associated with forcing preventative actions that decreases with increasing number of infected agents.

Concretely, as parameters we use $\alpha = 0.8$, $\beta = 0.2$, $\mu_0(I) = 0.2$, $\mu^0_0(H) = 0.5$, $\alpha^0 = 0.4$, $\Delta t = 0.1$, $c_I = 0.75$, $c_P = 0.5$, $c^0_\mu = 2$, $c^0_F = 1$ and a horizon of $T = 300$ for each episode.

BUFFET PROBLEM. Formally, minor agents have states $\mathcal{X} = [L]$ for L buffet locations, and can choose to move to any location $\mathcal{U} = [L]$ with geometric arrival rate, resulting in minor dynamics

$$\begin{aligned} p(n \mid [L] \setminus n, n, \dots) &= \alpha \Delta t, \\ p(n \mid [L] \setminus n, [L] \setminus n, \dots) &= 1 - \alpha \Delta t, \\ p(n \mid n, n, \dots) &= 1. \end{aligned}$$

The major agent state consists of the food state of the foraging locations, and the major agent at any time tries to fill one of the 3 foraging locations such that the locations remain as full as possible, and optionally as equal as possible. Hence, the major agent has states $\mathcal{X}^0 = \{0, \dots, B-1\}^L$ indicating the buffet filling status at each of L locations, and actions $\mathcal{U}^0 = [L]$ for filling up the buffet at a specific location. The major dynamics are such that a filling at location n is gained with probability $\alpha_+^0 \Delta t$ and lost with probability $\alpha_-^0 \mu(n) \Delta t$ whenever the current MF is μ.

Lastly, the rewards are defined as

$$\begin{aligned} r(x, u, x^0, u^0, \mu) &= c_f x_x^0 - c_c \mu(x) - c_u (1 - \mathbf{1}_x(u)), \\ r^0(x^0, u^0, \mu) &= \frac{1}{L} \sum_{i \in [L]} \left(c_f^0 x_i^0 - c_b^0 \left| x_i^0 - \frac{1}{L} \sum_{i \in [L]} x_i^0 \right| \right), \end{aligned}$$

where we have the reward coefficients c_f and c_f^0 for filled buffets, the crowdedness cost c_c, the movement cost c_u, and the imbalance cost c_b^0.

Concretely, as parameters we use $B = 5$, $L = 2$, $\alpha = 0.7$, $\mu_0(0) = 1$, $\mu_0^0 = \text{Unif}$, $\alpha_+^0 = 0.9$, $\alpha_-^0 = 1.0$, $\Delta t = 0.2$, $c_f = 0.75$, $c_c = 0.5$, $c_u = 1.0$, $c_f^0 = 2$, $c_b^0 = 1$ and a horizon of $T = 100$ for each episode.

ADVERTISEMENT DUOPOLY MODEL. The regulator chooses one of the actions $\mathcal{U}^0 = \{0, 1, 2\}$, where 0 denotes average price, 1 denotes low price and 2 denotes high price for advertisement by the second company. The regulator's state is $\mathcal{X}^0 = \{1, 2\}$ where $i,\ i = \{1, 2\}$ denotes the case where Company i is more aggressive in their advertisement. According to the state and the action of the regulator, the company i chooses their advertisement level $a_i(u^0, x^0)$ where a_i is a deterministic function. Similar to the SIS model, major dynamics are not influenced and given by $p^0(1|2, \dots) = p^0(2|1, \dots) = c\Delta t$ where $c > 0$ is an exogenous constant.

Minor agents' state space is $\mathcal{X} = \{1, 2\}$ where $i,\ i = \{1, 2\}$ denotes that they buy product i and they choose one of the actions $\mathcal{U} = \{O, C\}$ where O denotes that they are open to changes and C denotes that they are close to changes.

$$p(i|i^{-1}, x^0, u^0, u) = [a_i(x^0, u^0) - a_{i^{-1}}(x^0, u^0)] \lambda^u \Delta t$$

where λ^u is a coefficient that depends on the control of minor agent with $\lambda^O > \lambda^C$.

The reward functions are given as

$$\begin{aligned} r(x, u, x^0, u^0, \mu) &= \sum_{x \in \mathcal{X}} \mathbf{1}_{\{x=i\}} \left[c_\mu (\mu(i) - \mu(i^{-1})) + c_a a_i(x^0, u^0) \right] - \sum_{u' \in \mathcal{U}} \mathbf{1}_{\{u=u'\}} c_{u'}, \\ r^0(x^0, u^0, \mu) &= -c_m^0 |\mu(1) - \mu(2)| + c_a^0 \mathbf{1}_{\{u^0 \geq 1\}}. \end{aligned}$$

Concretely, as parameters we use $\Delta t = 0.3$, $c = 0.05$, $\mu_0 = \text{Unif}$, $\mu_0^0(1) = 1$, $c_C = 0.75$, $c_O = 1.0$, $c_a = 1.0$, $c_\mu = 1.0$, $c_a^0 = 0.1$, $c_m^0 = 1$, $\lambda^U = 0.2$, $\lambda^O = 1.2$, we let $a_i(u^0, x^0) = k_0 + k_x^0 \mathbf{1}_i(x^0) + k_u^0 \mathbf{1}_i(u^0)$ for $k_0 = 0.2$, $k_x^0 = 0.5$, $k_u^0 = 0.7$, and consider a horizon of $T = 100$ for each episode.

MORE FINITE HORIZON RESULTS Extending the qualitative results in the main text, in Figures D.1 and D.2 we see plausible qualitative equilibrium behavior in the finite horizon case for the Buffet and Advertisement problem. In Buffet, agents begin to move to the other location as the difference in fillings becomes sufficiently large, until the other location reaches a sufficiently high number of other agents. This can be seen both in the visualization of policies, and in the example trajectory plot. Such behavior is plausible, as the instantaneous cost of moving from one location to another must be higher than the perceived future gain from being at the desired location, leading to the observed hysteresis effect. We can also observe the effect of a finite time horizon as $t \to T$, as a potential change in location before the buffet ends is not useful in terms of improving rewards. As expected, the learned policies are symmetric in the locations. Meanwhile, in Advertisement, agents quickly run into an equilibrium that primarily depends on the exogeneous major agent state.

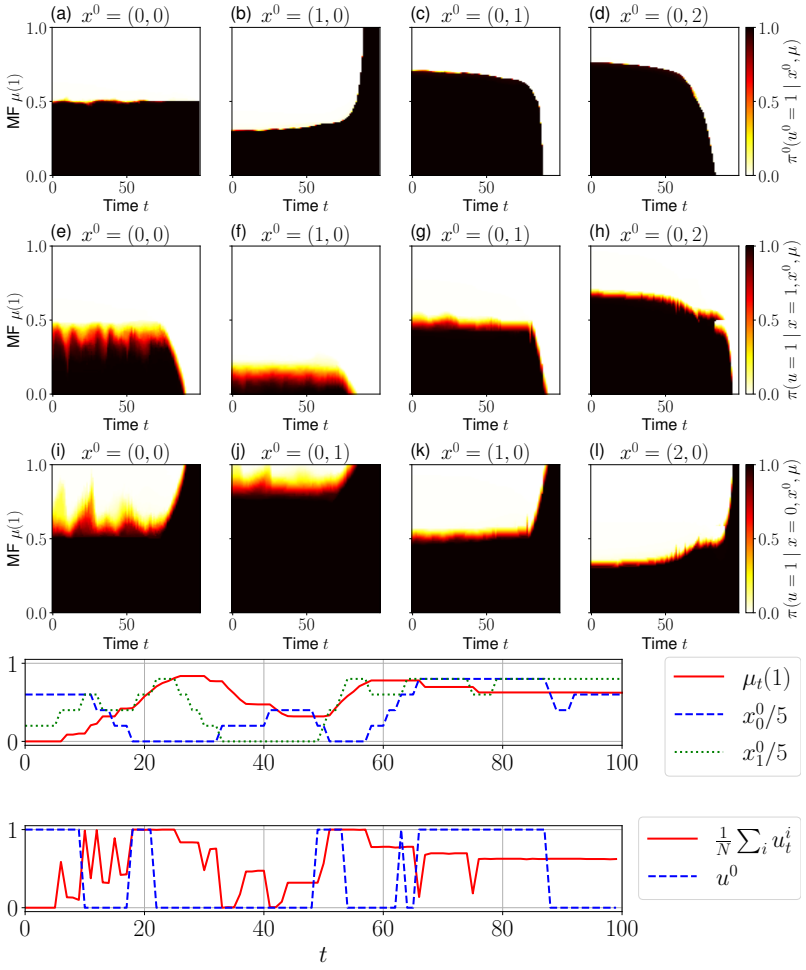


FIGURE D.1: Qualitative behavior in the finite horizon case for Buffet.

Further, as shown in Figure D.3, the behavior of the FP algorithm is consistent regardless of the choice of initialization. This implies robustness against the initialization of the algorithm. And as shown in the main text for the learned policy, we also have for the maximum entropy uniform policy a convergence of objectives over both discretization and number of agents, see Figures D.4 and D.5 respectively. In particular, this maximum entropy policy trivially fulfills Lipschitz conditions as in Assumption 3.4.3, and again verifies Theorems 3.4.1 and 3.4.3.

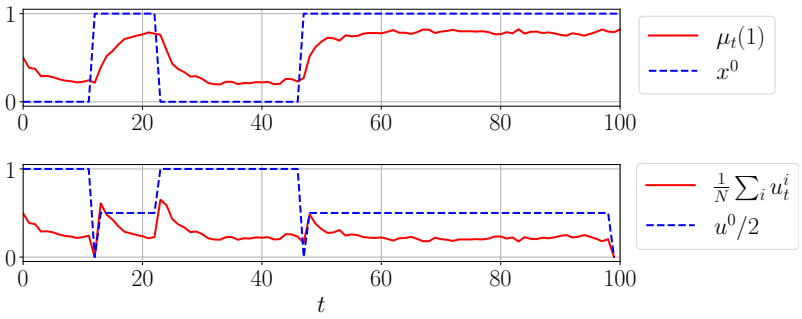


FIGURE D.2: Qualitative behavior in the finite horizon case for Advertisement.

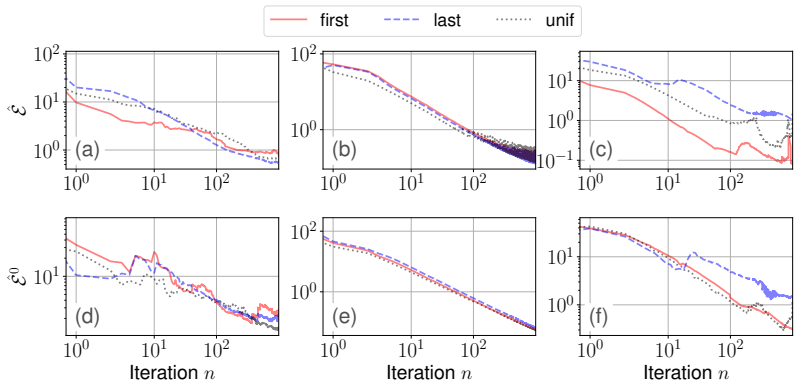


FIGURE D.3: The training curve of FP for various initializations. *first*: initial policy assigns all mass on the first action; *last*: all mass on the last action; *unif*: the uniform maximum entropy policy. Here, actions are ordered as they appear in the problem description. (a-c): Minor exploitability, (d-f): major exploitability, (a, d): SIS, (b, e): Buffet, (c, f): Advertisement.

INFINITE-HORIZON DISCOUNTED RESULTS As discussed in the main text, we can extend our algorithm to the infinite-horizon discounted objective case. We observe similar results and behavior as for the finite-horizon. For the infinite-horizon case, we apply value iteration to compute best responses, stopping value iteration when the maximum TD error over all states is less than 10^{-5}.

In particular, in Figure D.6, we observe the usual non-convergence of FPI, whereas the FP algorithm together with value iteration converges in terms of exploitability. Only sometimes does the FPI converge (here in SIS for $M = 80$), which again motivates the formulation of a FP algorithm. In the second part of the figure, we verify our empirical contribution, i.e. the FP algorithm, which generalizes also to the infinite-horizon discounted objectives.

In Figure D.7,we can see the convergence of objectives over discretization, both for the maximum entropy policy and the FP-learned policy, as well as the stability of the FP algorithm over discretization. The qualitative behavior is similar as the one seen in the main text for the finite horizon case. Further, in Figure D.8, the convergence of objectives and therefore the propagation of chaos over an increasing number of agents is again supported, both for the maximum entropy policy and the FP-learned policy.

Lastly, in Figure D.9, the qualitative behavior of SIS is shown and is comparable to the behavior in Figure 3.20, except for the absence of a transient finite-horizon effect near the end of the problem, due to the stationarity of the optimal policy under the discounted infinite-horizon objective.

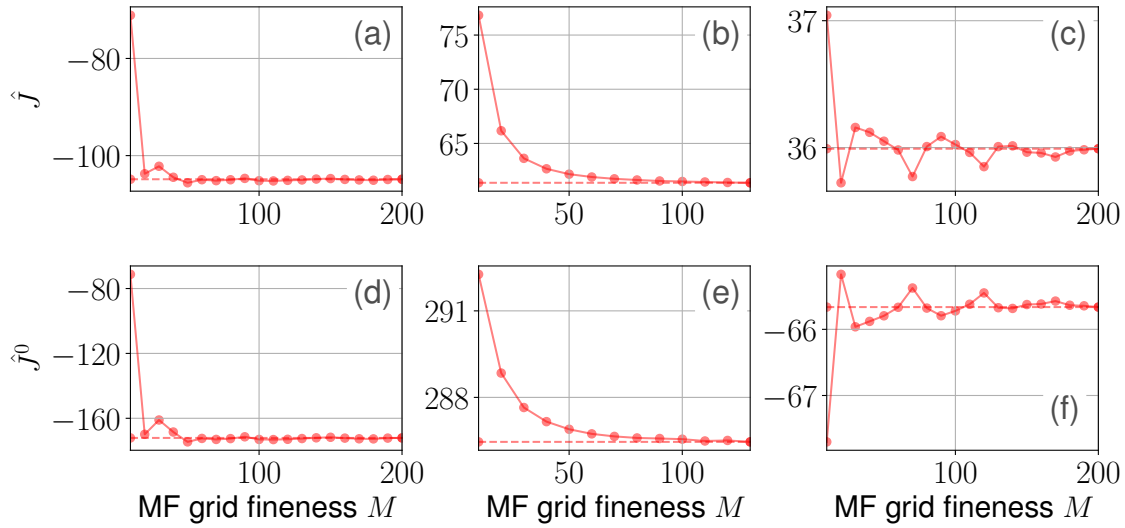


FIGURE D.4: Convergence of discretized objectives in the limit of fine discretization. The approximate objectives of the uniform policy (dashed: right-most entry) quickly converge with finer discretization. (a-c): Minor exploitability, (d-f): major exploitability, (a, d): SIS, (b, e): Buffet, (c, f): Advertisement.

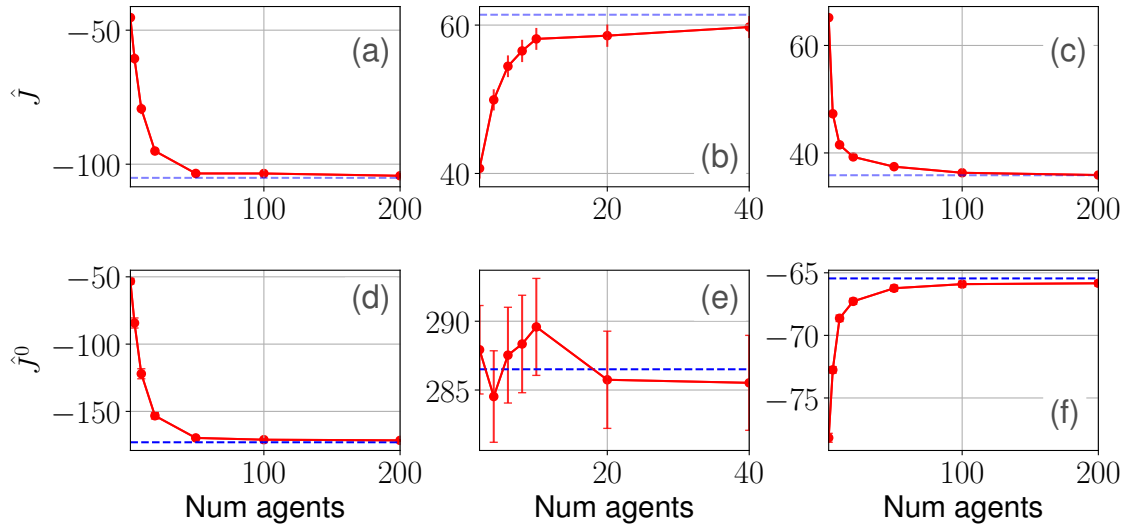


FIGURE D.5: Convergence of finite objectives in the limit. The mean N-agent objective (red) over 1000 (or 5000 in Buffet) episodes, with 95% confidence interval, against MF predictions $\hat{J}$, $\hat{J}^0$ of maximum entropy policy (blue, dashed). (a, d): SIS, (b, e): Buffet, (c, f): Advertisement.

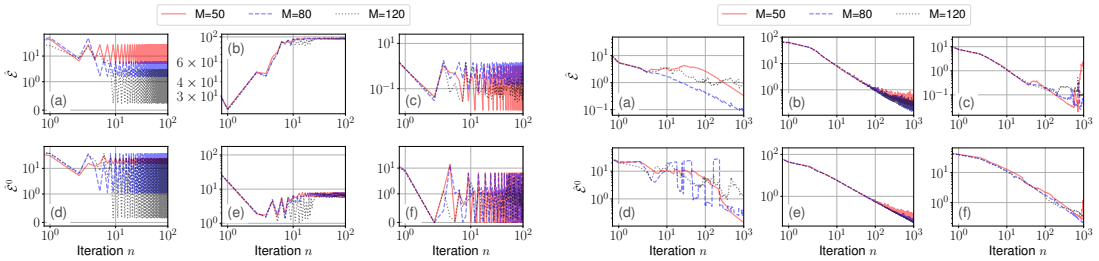


FIGURE D.6: Non-convergence of exploitability in infinite-horizon FPI. Exploitability over iterations of infinite-horizon FPI (left) and FP (right), where the former can run into a limit cycle. (c, f: Advertisement), (a, d: SIS), (b, e: Buffet). (a-c): Minor exploitability, (d-f): major exploitability.

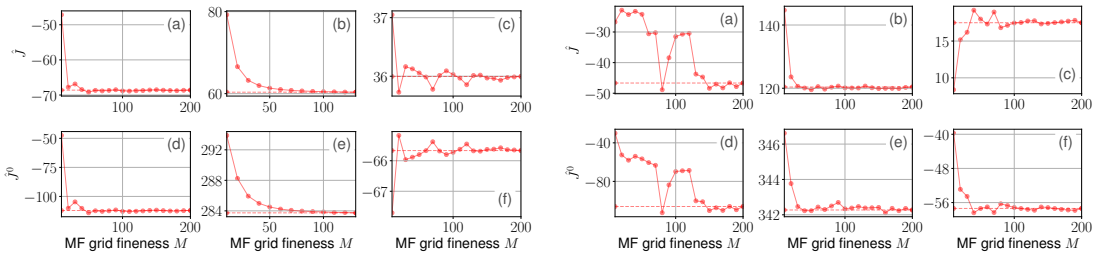


FIGURE D.7: Stability of infinite-horizon FP results under discretization. The infinite-horizon objective J, J^0 of the maximum entropy policy (left) and FP-learned policy (right) under discretization (dashed: right-most entry). The objectives quickly converge with increasing discretization fineness. (a-c): Minor exploitability, (d-f): major exploitability, (a, d): SIS, (b, e): Buffet, (c, f): Advertisement.

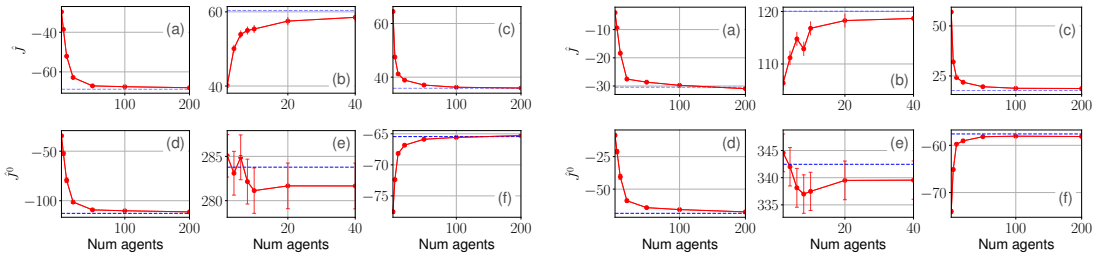


FIGURE D.8: The mean N-agent infinite-horizon objective (red) over 1000 episodes with 95% confidence interval, compared against approximate objectives $\hat{J}$, $\hat{J}^0$ of the maximum entropy policy (left) or FP-learned policy (right) as dashed blue line. (a): SIS, (b): Buffet, (c): Advertisement.

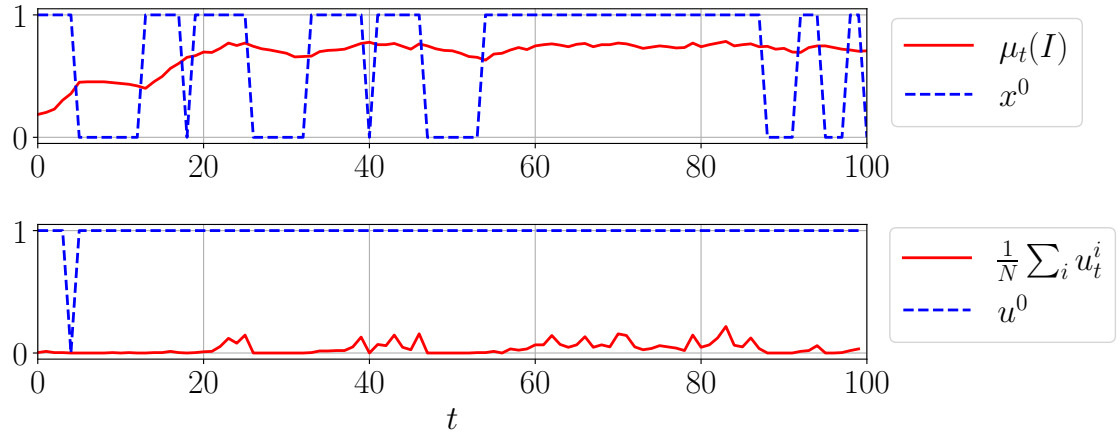


FIGURE D.9: Qualitative equilibrium behavior in infinite-horizon SIS. The resulting equilibrium behavior in infinite-horizon SIS is comparable to the finite-horizon case in Figure 3.20, but without finite horizon effects at the end of the episode.

APPENDIX E: SUPPLEMENTARY DETAILS ON SECTION 4.2

E.1 RELATED WORK

In this section, we provide additional context on related works. Since the introduction of MFGs in continuous and discrete time [22–24], MFGs have been studied in various forms, ranging from partially observed systems [79, 243] over learning-based solutions [9, 95, 97, 117, 118, 127, 367] on graphs [7, 139, 383, 384] to considering correlated equilibria [343, 369, 385].

While many works focus on non-cooperative settings with self-interested agents, this can run counter to the goal of engineering many-agent behavior, e.g., achieving cooperative behavior in swarms of drones. Instead, we focus on the related setting of cooperative MFC [103, 105, 201], see also work on differential [67], static [386], or discrete-time deterministic MFC [387]. For the unfamiliar reader, we point towards many extensive surveys on the topic of MF systems [25, 67, 124].

In general comparison, another well-known line of MF MARL [120, 241, 242, 388] focuses on approximating the influence of other agents on any particular agent by their average actions. Relatedly, some MARL algorithms introduce approximations over agent neighborhoods based on exponential decay [92, 93, 389]. In contrast, MFC assumes dependence on the entire distribution of agents and not, e.g., pairwise terms for each neighbor, per agent.

E.2 DETERMINISTIC MEAN FIELD CONTROL

In the following, we provide proofs that were omitted in the main text. To begin, in this section we recap standard deterministic MFC. Here, our general proof technique is introduced. It generalizes to the M3FC case and allows approximation properties and dynamic programming principles beyond finite spaces and Lipschitz continuity assumptions in compact spaces, for MFC models under simple continuity. In standard MFC, we have the model without major agents,

$$u_t^{i,N} \sim \pi_t(u_t^{i,N} \mid x_t^{i,N}, \mu_t^N), \tag{E.2.1}$$

$$x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, \mu_t^N) \tag{E.2.2}$$

while in the limit, we have the MF evolution

$$\mu_{t+1} = T(\mu_t, \mu_t \otimes \pi_t(\mu_t)) \coloneqq \iint p(\cdot \mid x, u, \mu_t)\pi_t(\mathrm{d}u \mid x, \mu_t)\mu_t(\mathrm{d}x) \tag{E.2.3}$$

and MFC system

$$h_t \sim \hat{\pi}_t(h_t \mid \mu_t), \quad \mu_{t+1} = T(\mu_t, h_t) \tag{E.2.4}$$

with objective $J(\hat{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(\mu_t)\right]$.

DYNAMIC PROGRAMMING AND PROPAGATION OF CHAOS. We may solve the hard finite-agent system (E.2.1) near-optimally by instead solving the MFC MDP, allowing direct application of single-agent RL to the MFC MDP with approximate optimality in large systems. Mild continuity assumptions are required.

Assumption E.2.1. *The transition kernel p and reward r are continuous.*

Assumption E.2.2. *The considered class of policies Π is equi-Lipschitz, i.e. there exists $L_\Pi > 0$ such that for all t and $\pi \in \Pi$, $\pi_t \in \mathcal{P}(\mathcal{U})^{\mathcal{X}\times\mathcal{P}(\mathcal{X})}$ is L_Π-Lipschitz.*

We note that Assumption E.2.1 holds true in studied finite spaces, if each transition matrix entry of P is continuous in the $|\mathcal{X}|$-dimensional MF vector on the simplex (but not necessarily Lipschitz as in [104, 105], the conditions of which we relax for deterministic MFC).

We show a dynamic programming principle [69] to solve for and show existence of a deterministic, stationary optimal policy via the value function V^* as the fixed point of the Bellman equation $V^*(\mu) = \max_{h \in \mathcal{H}(\mu)} r(\mu) + \gamma V^*(T(\mu, h))$.

Theorem E.2.1. *Under Assumption E.2.1, there exists an optimal stationary, deterministic policy $\hat{\pi}$ for Eq.* (E.2.4)*, with $\hat{\pi}(\mu) \in \arg\max_{h \in \mathcal{H}(\mu)} r(\mu) + \gamma V^*(T(\mu, h))$.*

This DPP can be used for computing solutions or to show optimality of stationary policies and existence of an optimum. Next, we show **propagation of chaos** [218]. Here, prior proof techniques [104, 105] are extended by our approach from *finite* to general *compact* spaces.

Theorem E.2.2. *Fix any family of equicontinuous functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{P}(\mathcal{X})}$. Under Assumptions E.2.1 and E.2.2, the empirical MF converges weakly, uniformly over $f \in \mathcal{F}$, $\pi \in \Pi$, $\hat{\pi} = \Phi^{-1}(\pi)$, to the limiting MF at all times $t \in \mathbb{N}$, $\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(\mu_t^N) \right] - \mathbb{E}\left[f(\mu_t) \right] \right| \to 0$.*

Importantly, propagation of chaos allows one to show approximate optimality of MFC policies in the large finite control problem, which is of practical relevance for solving many-agent problems.

Corollary E.2.1. *Under Assumptions E.2.1 and E.2.2, an optimal deterministic MFC policy $\pi^* \in \arg\max_{\hat{\pi}} J(\hat{\pi})$ yields ε-optimal finite-agent policy $\Phi(\pi^*) \in \Pi$, $J^N(\Phi(\pi^*)) \geq \sup_{\pi \in \Pi} J^N(\pi) - \varepsilon$, with $\varepsilon \to 0$ as $N \to \infty$.*

E.3 CONTINUITY OF MEAN FIELD DYNAMICS

First, we find continuity of the MFC dynamics T, which is used in the following proofs.

Lemma E.3.1. *Under Assumption E.2.1, we have $T(\mu_n, \nu_n) \to T(\mu, \nu)$ whenever $(\mu_n, \nu_n) \to (\mu, \nu)$,*

Proof. To show $T(\mu_n, \nu_n) \to T(\mu, \nu)$, consider any Lipschitz and bounded f with Lipschitz constant L_f, then

$$
\begin{aligned}
&\left| \int f \, \mathrm{d}(T(\mu_n, \nu_n) - T(\mu, \nu)) \right| \\
&= \left| \iiint f(x') p(\mathrm{d}x' \mid x, u, \mu_n) \nu_n(\mathrm{d}x, \mathrm{d}u) - \iiint f(x') p(\mathrm{d}x' \mid x, u, \mu) \nu(\mathrm{d}x, \mathrm{d}u) \right| \\
&\leq \iint \left| \int f(x') p(\mathrm{d}x' \mid x, u, \mu_n) - \int f(x') p(\mathrm{d}x' \mid x, u, \mu) \right| \nu_n(\mathrm{d}x, \mathrm{d}u) \\
&\quad + \left| \iiint f(x') p(\mathrm{d}x' \mid x, u, \mu) (\nu_n(\mathrm{d}x, \mathrm{d}u) - \nu(\mathrm{d}x, \mathrm{d}u)) \right| \\
&\leq \sup_{x \in \mathcal{X}, u \in \mathcal{U}} L_f W_1(p(\cdot \mid x, u, \mu_n), p(\cdot \mid x, u, \mu)) \\
&\quad + \left| \iiint f(x') p(\mathrm{d}x' \mid x, u, \mu) (\nu_n(\mathrm{d}x, \mathrm{d}u) - \nu(\mathrm{d}x, \mathrm{d}u)) \right| \to 0
\end{aligned}
$$

for the first term by 1-Lipschitzness of $\frac{f}{L_f}$ and Assumption E.2.1 (with compactness implying the uniform continuity), and for the second by $\nu_n \to \nu$ and from continuity by the same argument of $(x,u) \mapsto \iint f(x')p(\mathrm{d}x' \mid x,u,\mu)$. □

E.4 PROOF OF THEOREM E.2.1

Proof. The MFC MDP fulfills [69], Assumption 4.2.1. Here, we use [69], Condition 3.3.4(b1) instead of (b2), see also alternatively [390].

More specifically, for [69], Assumption 4.2.1(a), the cost function $-r$ is continuous by Assumption E.2.1, therefore also bounded by compactness of $\mathcal{P}(\mathcal{X})$, and finally also inf-compact on the state-action space of the MFC MDP, since for any $\mu \in \mathcal{P}(\mathcal{X})$ the set $\{h \in \mathcal{H}(\mu) \mid -r(\mu) \leq c\}$ is trivially given by $\mathcal{H}(\mu)$ whenever $-r(\mu) \leq c$, and $\emptyset$ otherwise. Here, we show that $\mathcal{H}(\mu) \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{U})$ is a closed subset of the compact space $\mathcal{P}(\mathcal{X} \times \mathcal{U})$ and therefore also compact. Note first that two measures $\mu, \mu' \in \mathcal{P}(\mathcal{X})$ are equal if and only if for all continuous and bounded f we have $\int f \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu'$, see e.g. [215], Theorem 1.3.

Therefore, as $\mathcal{H}(\mu)$ is defined by its first marginal μ, $\mathcal{H}(\mu)$ can be written as an intersection

$$\mathcal{H}(\mu) = \bigcap_{f \in C_b(\mathcal{X})} \left\{ h \in \mathcal{P}(\mathcal{X} \times \mathcal{U}) \;\middle|\; \int f \otimes \mathbf{1} \,\mathrm{d}h = \int f \,\mathrm{d}\mu \right\}$$

of closed sets: Since $h \mapsto \int f \otimes \mathbf{1} \,\mathrm{d}h$ is continuous, its preimage of the closed set $\{\int f \,\mathrm{d}\mu\}$ is closed. Here, $\otimes$ denotes the tensor product of f with the function $\mathbf{1}$ equal one, i.e. $f \otimes \mathbf{1}$ is the map $(x,u) \mapsto f(x)$.

Similarly, for [69], Assumption 4.2.1(b), the transition dynamics T are weakly continuous, as for any $(\mu_n, \nu_n) \to (\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X} \times \mathcal{U})$ we have $T(\mu_n, \nu_n) \to T(\mu, \nu)$ by Lemma E.3.1 and therefore $\int f \,\mathrm{d}\delta_{T(\mu_n,\nu_n)} = f(T(\mu_n,\nu_n)) \to f(T(\mu,\nu)) = \int f \,\mathrm{d}\delta_{T(\mu,\nu)}$ for any continuous and bounded $f \colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$.

Furthermore, the MFC MDP fulfills [69], Assumption 4.2.2 by boundedness of r from Assumption E.2.1. Therefore, the desired statement follows from [69], Theorem 4.2.3. □

E.5 PROOF OF THEOREM E.2.2

Proof. Note that we can also show the slightly stronger L_1 convergence statement with the absolute value inside of the expectation, $\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E}\left[\left|f(\mu_t^N) - f(\mu_t)\right|\right] \to 0$, but since this statement is only true for deterministic MFC, we avoid it here to later extend our proof directly to M3FC.

The statement $\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left|\mathbb{E}\left[f(\mu_t^N)\right] - \mathbb{E}\left[f(\mu_t)\right]\right| \to 0$ is shown inductively over $t \geq 0$. At time $t = 0$, it holds by the weak LLN argument, see also the first term below. Assuming the statement at time t, then for time $t+1$ we have

$$\begin{aligned}
&\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left|\mathbb{E}\left[f(\mu_{t+1}^N) - f(\mu_{t+1})\right]\right| \\
&\quad \leq \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left|\mathbb{E}\left[f(\mu_{t+1}^N) - f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right]\right| && \text{(E.5.5)} \\
&\qquad + \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left|\mathbb{E}\left[f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N))) - f(\mu_{t+1})\right]\right|. && \text{(E.5.6)}
\end{aligned}$$

For the first term (E.5.5), first note that by compactness of $\mathcal{P}(\mathcal{X})$, $\mathcal{F}$ is uniformly equicontinuous, and hence admits a non-decreasing, concave (as in [353], Lemma 6.1) modulus of continuity $\omega_{\mathcal{F}}\colon [0,\infty) \to [0,\infty)$ where $\omega_{\mathcal{F}}(x) \to 0$ as $x \to 0$ and $|f(\mu) - f(\nu)| \leq \omega_{\mathcal{F}}(W_1(\mu,\nu))$ for all $f \in \mathcal{F}$.

We also have uniform equicontinuity of $\mathcal{F}$ with respect to the space $(\mathcal{P}(\mathcal{X}), d_\Sigma)$ instead of $(\mathcal{P}(\mathcal{X}), W_1)$, as the identity map $\mathrm{id}\colon (\mathcal{P}(\mathcal{X}), d_\Sigma) \to (\mathcal{P}(\mathcal{X}), W_1)$ is uniformly continuous (as both d_Σ and W_1 metrize the topology of weak convergence, and $\mathcal{P}(\mathcal{X})$ is compact), and therefore there exists a modulus of continuity $\tilde{\omega}$ for the identity map such that for any $\mu, \nu \in (\mathcal{P}(\mathcal{X}), d_\Sigma)$, by the prequel

$$|f(\mu) - f(\nu)| \leq \omega_{\mathcal{F}}(W_1(\mathrm{id}\,\mu, \mathrm{id}\,\nu)) \leq \omega_{\mathcal{F}}(\tilde{\omega}(d_\Sigma(\mu,\nu)))$$

with $\tilde{\omega}_{\mathcal{F}} \coloneqq \omega_{\mathcal{F}} \circ \tilde{\omega}$, which can be replaced by its least concave majorant (again as in [353], Lemma 6.1).

Therefore, by Jensen's inequality, for Eq. (E.5.5) we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[f(\mu_{t+1}^N) - f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right]\right| \\
&\quad \leq \mathbb{E}\left[\tilde{\omega}_{\mathcal{F}}(d_\Sigma(\mu_{t+1}^N, T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N))))\right] \\
&\quad \leq \tilde{\omega}_{\mathcal{F}}\left(\mathbb{E}\left[d_\Sigma(\mu_{t+1}^N, T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right]\right)
\end{aligned}$$

irrespective of π, f via concavity of $\tilde{\omega}_{\mathcal{F}}$. Introducing for readability $x_t^N \equiv \{x_t^{i,N}\}_{i\in[N]}$, we then obtain

$$\begin{aligned}
&\mathbb{E}\left[d_\Sigma(\mu_{t+1}^N, T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right] \\
&\quad = \sum_{m=1}^{\infty} 2^{-m} \mathbb{E}\left[\left|\int f_m \,\mathrm{d}(\mu_{t+1}^N - T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right|\right] \\
&\quad \leq \sup_{m\geq 1} \mathbb{E}\left[\mathbb{E}_{x_t^N}\left[\left|\int f_m \,\mathrm{d}(\mu_{t+1}^N - T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right|\right]\right],
\end{aligned}$$

and by the following weak LLN argument, for the squared term and any f_m

$$\begin{aligned}
&\mathbb{E}_{x_t^N}\left[\left|\int f_m \,\mathrm{d}(\mu_{t+1}^N - T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right|\right]^2 \\
&\quad = \mathbb{E}_{x_t^N}\left[\left|\frac{1}{N}\sum_{i=1}^N \left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{x_t^N}\left[f_m(x_{t+1}^{i,N})\right]\right)\right|\right]^2 \\
&\quad \leq \mathbb{E}_{x_t^N}\left[\left|\frac{1}{N}\sum_{i=1}^N \left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{x_t^N}\left[f_m(x_{t+1}^{i,N})\right]\right)\right|^2\right] \\
&\quad = \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}_{x_t^N}\left[\left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{x_t^N}\left[f_m(x_{t+1}^{i,N})\right]\right)^2\right] \leq \frac{4}{N} \to 0
\end{aligned}$$

by bounding $|f_m| \leq 1$, as the cross-terms are zero by conditional independence of $x_{t+1}^{i,N}$ given x_t^N. By the prequel, the term (E.5.5) hence converges to zero.

For the second term (E.5.6), we have

$$\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N))) - f(\mu_{t+1})\right]\right|$$

$$
\begin{aligned}
&= \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N))) - f(T(\mu_t, \mu_t \otimes \pi_t(\mu_t)))\right]\right| \\
&\leq \sup_{\pi\in\Pi}\sup_{g\in\mathcal{G}} \left|\mathbb{E}\left[g(\mu_t^N) - g(\mu_t)\right]\right| \to 0
\end{aligned}
$$

by the induction assumption, where we defined $g = f \circ \tilde{T}^{\pi_t}$ from the class $\mathcal{G}$ of equicontinuous functions with modulus of continuity $\omega_{\mathcal{G}} \coloneqq \omega_{\mathcal{F}} \circ \omega_T$, where ω_T denotes the uniform modulus of continuity of $\mu_t \mapsto \tilde{T}^{\pi_t}(\mu_t) \coloneqq T(\mu_t, \mu_t \otimes \pi_t(\mu_t)))$ over all policies π. Here, this equicontinuity of $\{\tilde{T}^{\pi_t}\}_{\pi\in\Pi}$ follows from Lemma E.3.1 and the equicontinuity of functions $\mu_t \mapsto \mu_t \otimes \pi_t(\mu_t)$ due to uniformly Lipschitz Π as we show in the following, completing the proof by induction:

Consider $\mu_n \to \mu \in \mathcal{P}(\mathcal{X})$, then we have

$$
\begin{aligned}
&\sup_{\pi\in\Pi} W_1(\mu_n \otimes \pi_t(\mu_n), \mu \otimes \pi_t(\mu)) \\
&\quad= \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1} \left|\int f' \,\mathrm{d}(\mu_n \otimes \pi_t(\mu_n) - \mu \otimes \pi_t(\mu))\right| \\
&\quad\leq \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1} \left|\iint f'(x,u)(\pi_t(\mathrm{d}u \mid x, \mu_n) - \pi_t(\mathrm{d}u \mid x, \mu))\mu_n(\mathrm{d}x)\right| \\
&\qquad+ \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1} \left|\iint f'(x,u)\pi_t(\mathrm{d}u \mid x, \mu)(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x))\right|
\end{aligned}
$$

where for the first term

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1} \left|\iint f'(x,u)(\pi_t(\mathrm{d}u \mid x, \mu_n) - \pi_t(\mathrm{d}u \mid x, \mu))\mu_n(\mathrm{d}x)\right| \\
&\quad\leq \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1} \int \left|\int f'(x,u)(\pi_t(\mathrm{d}u \mid x, \mu_n) - \pi_t(\mathrm{d}u \mid x, \mu))\right| \mu_n(\mathrm{d}x) \\
&\quad\leq \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\sup_{x\in\mathcal{X}} \left|\int f'(x,u)(\pi_t(\mathrm{d}u \mid x, \mu_n) - \pi_t(\mathrm{d}u \mid x, \mu))\right| \\
&\quad= \sup_{\pi\in\Pi}\sup_{x\in\mathcal{X}} W_1(\pi_t(\cdot \mid x, \mu_n), \pi_t(\cdot \mid x, \mu)) \\
&\quad\leq L_\Pi W_1(\mu_n, \mu) \to 0
\end{aligned}
$$

by Assumption E.2.2, and similarly for the second by first noting 1-Lipschitzness of $x \mapsto \int \frac{f'(x,u)}{L_\Pi+1}\pi_t(\mathrm{d}u \mid x, \mu)$, as for $y \neq x$

$$
\begin{aligned}
&\left|\int \frac{f'(y,u)}{L_\Pi+1}\pi_t(\mathrm{d}u \mid y, \mu) - \int \frac{f'(x,u)}{L_\Pi+1}\pi_t(\mathrm{d}u \mid x, \mu)\right| \\
&\quad\leq \left|\int \frac{f'(y,u) - f'(x,u)}{L_\Pi+1}\pi_t(\mathrm{d}u \mid y, \mu)\right| + \left|\int \frac{f'(x,u)}{L_\Pi+1}(\pi_t(\mathrm{d}u \mid y, \mu) - \pi_t(\mathrm{d}u \mid x, \mu))\right| \\
&\quad\leq \frac{1}{L_\Pi+1} d(y,x) + \frac{1}{L_\Pi+1} W_1(\pi_t(\cdot \mid y, \mu), \pi_t(\cdot \mid x, \mu)) \\
&\quad\leq \left(\frac{1}{L_\Pi+1} + \frac{L_\Pi}{L_\Pi+1}\right) d(x,y)
\end{aligned} \tag{E.5.7}
$$

with $\frac{1}{L_\Pi+1} + \frac{L_\Pi}{L_\Pi+1} = 1 \leq 1$, and therefore again

$$
\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1} \left|\iint f'(x,u)\pi_t(\mathrm{d}u \mid x, \mu)(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x))\right|
$$

$$= \sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} (L_\Pi + 1) \left| \iint \frac{f'(x,u)}{L_\Pi + 1} \pi_t(\mathrm{d}u \mid x, \mu)(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x)) \right|$$
$$\le (L_\Pi + 1) W_1(\mu_n, \mu) \to 0.$$

This completes the proof by induction. □

E.6 PROOF OF COROLLARY E.2.1

Proof. First, we show that from uniform convergence in Theorem E.2.2, the finite-agent objectives converge uniformly to the MFC limit.

Lemma E.6.1. *Under Assumptions E.2.1 and E.2.2, the finite-agent objective converges uniformly to the MFC limit,*

$$\sup_{\pi\in\Pi} \left| J^N(\pi) - J(\Phi^{-1}(\pi)) \right| \to 0. \tag{E.6.8}$$

Proof. For any $\varepsilon > 0$, choose time $T \in \mathbb{N}$ such that $\sum_{t=T}^{\infty} \gamma^t \mathbb{E}\left|\left[r(\mu_t^N) - r(\mu_t)\right]\right| \le \frac{\gamma^T}{1-\gamma} \max_\mu 2|r(\mu)| < \frac{\varepsilon}{2}$. By Theorem E.2.2, $\sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left|\left[r(\mu_t^N) - r(\mu_t)\right]\right| < \frac{\varepsilon}{2}$ for sufficiently large N. The result follows. ■

The approximate optimality of MFC solutions in the finite system follows immediately: By Lemma E.6.1, we have

$$\begin{aligned} J^N(\Phi(\pi^*)) - \sup_{\pi\in\Pi} J^N(\pi) &= \inf_{\pi\in\Pi} (J^N(\pi^*) - J^N(\pi)) \\ &\ge \inf_{\pi\in\Pi} (J^N(\Phi(\pi^*)) - J(\pi^*)) + \inf_{\pi\in\Pi} (J(\pi^*) - J(\Phi^{-1}(\pi))) + \inf_{\pi\in\Pi} (J(\Phi^{-1}(\pi)) - J^N(\pi)) \\ &\ge -\frac{\varepsilon}{2} + 0 - \frac{\varepsilon}{2} = -\varepsilon \end{aligned}$$

for sufficiently large N, where the second term is zero by optimality of π^* in the MFC problem. □

E.7 STOCHASTIC MEAN FIELD CONTROL

For convenience, we also restate the results for MFC with major states, or common noise. We have the finite MFC system with major states

$$u_t^{i,N} \sim \pi_t(u_t^{i,N} \mid x_t^{i,N}, x_t^{0,N}, \mu_t^N), \tag{E.7.9a}$$
$$x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, \mu_t^N), \quad x_{t+1}^{0,N} \sim p^0(x_{t+1}^{0,N} \mid x_t^{0,N}, \mu_t^N) \tag{E.7.9b}$$

and objective $J^N(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^{0,N}, \mu_t^N)\right]$ analogous to Eq. (E.2.1), with the corresponding limiting MFC MDP with major states analogous to Eq. (E.2.4),

$$h_t \sim \hat{\pi}_t(h_t \mid x_t^0, \mu_t), \quad \mu_{t+1} = T(x_t^0, \mu_t, h_t), \quad x_{t+1}^0 \sim p^0(x_{t+1}^0 \mid x_t^0, \mu_t) \tag{E.7.10}$$

with objective $J(\hat{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^0, \mu_t)\right]$, where

$$T(x^0, \mu, h) \coloneqq \iint p(\cdot \mid x, u, x^0, \mu) h(\mathrm{d}x, \mathrm{d}u).$$

Assumption E.7.1. *The transition kernels p, p^0 and rewards r are Lipschitz continuous with constants L_p, L_{p^0}, L_r.*

Assumption E.7.2. *The class of policies Π are equi-Lipschitz, i.e. there exists $L_\Pi > 0$ such that for all t and $\pi \in \Pi$, $\pi_t \in \mathcal{P}(\mathcal{U})^{\mathcal{X}\times\mathcal{P}(\mathcal{X})}$ is L_Π-Lipschitz.*

Theorem E.7.1. *Under Assumption E.7.1, there exists an optimal stationary, deterministic policy $\hat{\pi}$ for the MFC MDP Eq.* (E.7.10) *by choosing $\hat{\pi}(x^0, \mu)$ from the maximizers of $\arg\max_{h\in\mathcal{H}(\mu)} r(x^0, \mu) + \gamma\mathbb{E}_{y^0\sim p^0(y^0|x^0,\mu)} V^*(y^0, T(x^0, \mu, h))$, with V^* the unique fixed point of the Bellman equation $V^*(x^0, \mu) = \max_{h\in\mathcal{H}(\mu)} r(x^0, \mu) + \gamma\mathbb{E}_{y^0\sim p^0(y^0|x^0,\mu)} V^*(y^0, T(x^0, \mu, h))$ (value function).*

Theorem E.7.2. *Fix any family of equi-Lipschitz functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}^0\times\mathcal{P}(\mathcal{X})}$ with shared Lipschitz constant $L_\mathcal{F}$ for all $f \in \mathcal{F}$. Under Assumption E.7.1, the random variable $(x_t^{0,N}, \mu_t^N)$ converges weakly, uniformly over $\mathcal{F}$, Π, to (x_t^0, μ_t) at all times $t \in \mathbb{N}$,*

$$\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(x_t^{0,N}, \mu_t^N) - f(x_t^0, \mu_t)\right]\right| \to 0. \tag{E.7.11}$$

Corollary E.7.1. *Under Assumptions E.7.1 and E.7.2, optimal deterministic MFC policies $\pi^* \in \arg\max_\pi J(\pi)$ result in ε-optimal policies $\Phi(\pi^*)$ in the finite-agent problem with $\varepsilon \to 0$ as $N \to \infty$,*

$$J^N(\Phi(\pi^*)) \geq \sup_{\pi\in\Pi} J^N(\pi) - \varepsilon. \tag{E.7.12}$$

The proofs and interpretation are directly analogous to the M3FC case and the following proofs, by leaving out the major agent actions, or alternatively using the M3FC results with a trivial singleton major action space, $|\mathcal{U}^0| = 1$.

E.8 PROOF OF THEOREM 4.2.1

Proof. The proof is analogous to Appendix E.4 by first showing the continuity of T (proof further below).

Lemma E.8.1. *Under Assumption 4.2.1, for any sequence $(x_n^0, u_n^0, \mu_n, \nu_n) \to (x^0, u^0, \mu, \nu) \in \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}\times\mathcal{U})$, we have $T(x_n^0, u_n^0, \mu_n, \nu_n) \to T(x^0, u^0, \mu, \nu)$.*

For [69], Assumption 4.2.1(a), the cost function $-r$ is continuous by Assumption 4.2.1, therefore also bounded by compactness of $\mathcal{X}^0 \times \mathcal{P}(\mathcal{X})$, and finally also inf-compact on the state-action space of the M3FC MDP, since for any $(x^0, \mu) \in \mathcal{X}^0 \times \mathcal{P}(\mathcal{X})$ the set $\{(h, u^0) \in \mathcal{H}(\mu) \times \mathcal{U}^0 \mid -r(x^0, u^0, \mu) \leq c\}$ is given by $\mathcal{H}(\mu) \times \tilde{r}^{-1}((-\infty, c])$, where we defined $\tilde{r}(u^0) := -r(x^0, u^0, \mu)$. Note that $\mathcal{H}(\mu)$ is compact by the same argument as in Appendix E.4, while $\tilde{r}$ is continuous by Assumption 4.2.1 and therefore its preimage of the closed set $(-\infty, c]$ is compact.

For [69], Assumption 4.2.1(b), consider any continuous and bounded $f\colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. The continuity is uniform by compactness. Hence, $\sup_{x'\in\mathcal{X}^0} |f(x', \mu_n') - f(x', \mu')| \to 0$ as $\mu_n' \to \mu' \in \mathcal{P}(\mathcal{X})$. Thus, whenever $(x_n^0, u_n^0, \mu_n, \nu_n) \to (x^0, u^0, \mu, \nu) \in \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}\times\mathcal{U})$, we have

$$\left|\iint f(x', \mu)\, \delta_{T_n^*}(\mathrm{d}\mu')\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \iint f(x', \mu)\, \delta_{T^*}(\mathrm{d}\mu')\, p^0(\mathrm{d}x' \mid x^0, u^0, \mu)\right|$$

$$
\begin{aligned}
&= \left| \int f(x', T_n^*)\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \int f(x', T^*)\, p^0(\mathrm{d}x' \mid x^0, u^0, \mu) \right| \\
&\leq \left| \int f(x', T_n^*)\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \int f(x', T^*)\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) \right| \\
&\quad + \left| \int f(x', T^*)\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \int f(x', T^*)\, p^0(\mathrm{d}x' \mid x^0, u^0, \mu) \right| \\
&\leq \sup_{x' \in \mathcal{X}^0} \left| f(x', T_n^*) - f(x', T^*) \right| \\
&\quad + \left| \int \tilde{f}(x')\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \int \tilde{f}(x')\, p^0(\mathrm{d}x' \mid x^0, u^0, \mu) \right| \to 0
\end{aligned}
$$

for the first term by the prequel where $T_n^* \coloneqq T(x_n^0, u_n^0, \mu_n, \nu_n) \to T^* \coloneqq T(x^0, u^0, \mu, \nu)$ by Lemma E.8.1, and for the second term by applying Assumption 4.2.1 to $\tilde{f}(x') \coloneqq f(x', T^*)$. This shows weak continuity of the dynamics.

Furthermore, the M3FC MDP fulfills [69], Assumption 4.2.2 by boundedness of r from Assumption 4.2.1. Therefore, the desired statement follows from [69], Theorem 4.2.3. □

E.9 PROOF OF LEMMA E.8.1

Proof. To show $T(x_n^0, u_n^0, \mu_n, \nu_n) \to T(x^0, u^0, \mu, \nu)$, consider any Lipschitz and bounded f with Lipschitz constant L_f, then

$$
\begin{aligned}
&\left| \int f \,\mathrm{d}(T(x_n^0, u_n^0, \mu_n, \nu_n) - T(x^0, u^0, \mu, \nu)) \right| \\
&= \left| \iiint f(x') \left(p(\mathrm{d}x' \mid x, u, x_n^0, u_n^0, \mu_n) \nu_n(\mathrm{d}x, \mathrm{d}u) - p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu) \nu(\mathrm{d}x, \mathrm{d}u) \right) \right| \\
&\leq \iint \left| \int f(x') p(\mathrm{d}x' \mid x, u, x_n^0, u_n^0, \mu_n) - \int f(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu) \right| \nu_n(\mathrm{d}x, \mathrm{d}u) \\
&\quad + \left| \iiint f(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu) (\nu_n(\mathrm{d}x, \mathrm{d}u) - \nu(\mathrm{d}x, \mathrm{d}u)) \right| \\
&\leq \sup_{x \in \mathcal{X}, u \in \mathcal{U}} L_f W_1(p(\cdot \mid x, u, x_n^0, u_n^0, \mu_n), p(\cdot \mid x, u, x^0, u^0, \mu)) \\
&\quad + \left| \iiint f(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu) (\nu_n(\mathrm{d}x, \mathrm{d}u) - \nu(\mathrm{d}x, \mathrm{d}u)) \right| \to 0
\end{aligned}
$$

for the first term by 1-Lipschitzness of $\frac{f}{L_f}$ and Assumption 4.2.1 (with compactness implying the uniform continuity), and for the second by $\nu_n \to \nu$ and continuity of $(x, u) \mapsto \iint f(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)$ by the same argument. □

E.10 PROOF OF THEOREM 4.2.2

Proof. The statement $\sup_{f, \pi, \pi^0} \left| \mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - f(x_t^0, u_t^0, \mu_t) \right] \right|$ is shown inductively over $t \geq 0$. At time $t = 0$, it holds by the weak LLN argument, see also the first term below. Assuming the statement at time t, then for time $t + 1$ we have

$$
\sup_{(\pi, \pi^0) \in \Pi \times \Pi^0} \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) - f(x_{t+1}^0, u_{t+1}^0, \mu_{t+1}) \right] \right|
$$

$$\leq \sup_{\pi,\pi^0}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(x_{t+1}^{0,N},u_{t+1}^{0,N},\mu_{t+1}^N)-f(x_{t+1}^{0,N},u_{t+1}^{0,N},\hat{\mu}_{t+1}^N)\right]\right| \tag{E.10.13}$$

$$+\sup_{\pi,\pi^0}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(x_{t+1}^{0,N},u_{t+1}^{0,N},\hat{\mu}_{t+1}^N)-f(x_{t+1}^{0},u_{t+1}^{0},\mu_{t+1})\right]\right| \tag{E.10.14}$$

where for readability, we again write $\pi_t(x_t^0,\mu_t) \coloneqq \pi_t(\cdot \mid \cdot, x_t^0,\mu_t)$ and introduce the random variable

$$\hat{\mu}_{t+1}^N \coloneqq T(x_t^{0,N},u_t^{0,N},\mu_t^N,\mu_t^N\otimes\pi_t(x_t^{0,N},\mu_t^N)).$$

By compactness of $\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})$, $\mathcal{F}$ is uniformly equicontinuous, and hence admits a non-decreasing, concave (as in [353], Lemma 6.1) modulus of continuity $\omega_{\mathcal{F}}\colon[0,\infty)\to[0,\infty)$ where $\omega_{\mathcal{F}}(x)\to 0$ as $x\to 0$ and $|f(x,u,\mu)-f(x',u',\nu)|\leq\omega_{\mathcal{F}}(d(x,x')+d(u,u')+W_1(\mu,\nu))$ for all $f\in\mathcal{F}$, and analogously there exists such $\tilde{\omega}_{\mathcal{F}}$ with respect to $(\mathcal{P}(\mathcal{X}),d_\Sigma)$ instead of $(\mathcal{P}(\mathcal{X}),W_1)$ as in Appendix E.5.

For the first term (E.10.13), let $x_t^N\equiv\{x_t^{i,N}\}_{i\in[N]}$. Then, by the weak LLN argument,

$$\begin{aligned}
&\sup_{\pi,\pi^0}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(x_{t+1}^{0,N},u_{t+1}^{0,N},\mu_{t+1}^N)-f(x_{t+1}^{0,N},u_{t+1}^{0,N},\hat{\mu}_{t+1}^N)\right]\right|\\
&\quad\leq\sup_{\pi,\pi^0}\mathbb{E}\left[\tilde{\omega}_{\mathcal{F}}(d_\Sigma(\mu_{t+1}^N,\hat{\mu}_{t+1}^N))\right]\\
&\quad\leq\sup_{\pi,\pi^0}\tilde{\omega}_{\mathcal{F}}\left(\sum_{m=1}^{\infty}2^{-m}\mathbb{E}\left[\left|\mu_{t+1}^N(f_m)-\hat{\mu}_{t+1}^N(f_m)\right|\right]\right)\\
&\quad\leq\sup_{\pi,\pi^0}\tilde{\omega}_{\mathcal{F}}\left(\sup_{m\geq 1}\mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\left|\mu_{t+1}^N(f_m)-\hat{\mu}_{t+1}^N(f_m)\right|\right]\right]\right)\\
&\quad=\sup_{\pi,\pi^0}\tilde{\omega}_{\mathcal{F}}\left(\sup_{m\geq 1}\mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(f_m(x_{t+1}^{i,N})-\mathbb{E}_{\mathcal{J}_t}\left[f_m(x_{t+1}^{i,N})\right]\right)\right|\right]\right]\right)\\
&\quad\leq\sup_{\pi,\pi^0}\tilde{\omega}_{\mathcal{F}}\left(\sup_{m\geq 1}\mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(f_m(x_{t+1}^{i,N})-\mathbb{E}_{\mathcal{J}_t}\left[f_m(x_{t+1}^{i,N})\right]\right)\right|^2\right]\right]^{1/2}\right)\\
&\quad=\sup_{\pi,\pi^0}\tilde{\omega}_{\mathcal{F}}\left(\sup_{m\geq 1}\left(\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\mathbb{E}_{\mathcal{J}_t}\left[\left(f_m(x_{t+1}^{i,N})-\mathbb{E}_{\mathcal{J}_t}\left[f_m(x_{t+1}^{i,N})\right]\right)^2\right]\right]\right)^{1/2}\right)\\
&\quad\leq\tilde{\omega}_{\mathcal{F}}\left(\frac{2}{\sqrt{N}}\right)\to 0
\end{aligned}\tag{E.10.15}$$

where $\mathcal{J}_t\coloneqq(x_t^{0,N},u_t^{0,N},x_t^N)$ by bounding $|f_m|\leq 1$, as the cross-terms disappear.

For the second term (E.10.14), by noting $\hat{\mu}_{t+1}^N=T(x_t^{0,N},u_t^{0,N},\mu_t^N,\mu_t^N\otimes\pi_t(x_t^{0,N},\mu_t^N))$, we have

$$\begin{aligned}
&\sup_{\pi,\pi^0}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(x_{t+1}^{0,N},u_{t+1}^{0,N},\hat{\mu}_{t+1}^N)-f(x_{t+1}^{0},u_{t+1}^{0},\mu_{t+1})\right]\right|\\
&\quad=\sup_{\pi,\pi^0}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\iint f(x',u',\hat{\mu}_{t+1}^N)\pi_t^0(\mathrm{d}u'\mid x',\mu_{t+1}^N)p^0(\mathrm{d}x'\mid x_t^{0,N},u_t^{0,N},\mu_t^N)\right.\right.\\
&\qquad\qquad\left.\left.-\iint f(x',u',\mu_{t+1})\pi_t^0(\mathrm{d}u'\mid x',\mu_{t+1})p^0(\mathrm{d}x'\mid x_t^0,u_t^0,\mu_t)\right]\right|
\end{aligned}$$

$$\leq \sup_{\pi,\pi^0} \sup_{f\in\mathcal{F}} \mathbb{E}\left[\sup_{x'}\left|\int f(x',u',\hat{\mu}^N_{t+1})(\pi^0_t(\mathrm{d}u' \mid x',\mu^N_{t+1}) - \pi^0_t(\mathrm{d}u' \mid x',\hat{\mu}^N_{t+1}))\right|\right] \quad \text{(E.10.16)}$$

$$+ \sup_{\pi,\pi^0} \sup_{g\in\mathcal{G}} \left|\mathbb{E}\left[g(x^{0,N}_t, u^{0,N}_t, \mu^N_t) - g(x^0_t, u^0_t, \mu_t)\right]\right| \quad \text{(E.10.17)}$$

and analyze each term separately, where we defined the function $g\colon \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X})$ as

$$g(x^0,u^0,\mu) \coloneqq \iint f(x',u',T^*)\pi^0_t(\mathrm{d}u' \mid x',T^*)p^0(\mathrm{d}x' \mid x^0,u^0,\mu)$$

from the class $\mathcal{G}$ of such functions for any policies π, π^0, where $T^* \coloneqq T(x^0,u^0,\mu,\mu\otimes\pi_t(x^0,\mu))$.

For Eq. (E.10.16), defining a modulus of continuity $\tilde{\omega}_{\Pi^0}$ for Π^0 as for $\mathcal{F}$, we have

$$\begin{aligned}
&\sup_{\pi,\pi^0}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\sup_{x'}\left|\int f(x',u',\hat{\mu}^N_{t+1})(\pi^0_t(\mathrm{d}u' \mid x',\mu^N_{t+1}) - \pi^0_t(\mathrm{d}u' \mid x',\hat{\mu}^N_{t+1}))\right|\right] \\
&\quad\leq \sup_{\pi,\pi^0} \mathbb{E}\left[L_{\mathcal{F}} \sup_{x'} W_1(\pi^0_t(\cdot \mid x',\mu^N_{t+1}), \pi^0_t(\cdot \mid x',\hat{\mu}^N_{t+1}))\right] \\
&\quad\leq \sup_{\pi,\pi^0} \mathbb{E}\left[L_{\mathcal{F}}\tilde{\omega}_{\Pi^0}(d_\Sigma(\mu^N_{t+1},\hat{\mu}^N_{t+1}))\right] \leq L_{\mathcal{F}}\tilde{\omega}_{\Pi^0}\left(\frac{2}{\sqrt{N}}\right) \to 0.
\end{aligned}$$

Lastly, for Eq. (E.10.17), we first note that the class $\mathcal{G}$ of functions is equi-Lipschitz.

Lemma E.10.1. *Under Assumptions 4.2.1 and 4.2.2, the map* $(x^0,u^0,\mu) \mapsto T(x^0,u^0,\mu,\mu\otimes\pi_t(x^0,\mu))$ *is Lipschitz with constant* $L_T \coloneqq (2L_\Pi+1)\cdot(L_p+(L_p+1)L_\Pi+(L_p+L_\Pi+1))$.

Lemma E.10.2. *Under Assumptions 4.2.1 and 4.2.2, for any equi-Lipschitz* $\mathcal{F}$ *with constant* $L_{\mathcal{F}}$*, the function class* $\mathcal{G}$ *is equi-Lipschitz with constant* $L_{\mathcal{G}} \coloneqq (L_{\mathcal{F}}L_T + L_{\mathcal{F}}L_{\Pi^0}L_T + L_{\mathcal{F}}L_\Pi L_{p^0})$.

Therefore, for Eq. (E.10.17), we have

$$\sup_{\pi,\pi^0}\sup_{g\in\mathcal{G}}\left|\mathbb{E}\left[g(x^{0,N}_t,u^{0,N}_t,\mu^N_t) - g(x^0_t,u^0_t,\mu_t)\right]\right| \to 0$$

by the induction assumption over the class $\mathcal{G}$ of equi-Lipschitz functions, completing the proof by induction. The existence of independent optimal π, π^0 follows from Remark 4.2.3. This completes the proof.

For finite minor states, we can quantify the convergence rate more precisely as $\mathcal{O}(1/\sqrt{N})$, since the two metrizations d_Σ and W_1 are then Lipschitz equivalent and the above moduli of continuity simply become a multiplication with the Lipschitz constant, so for convenience we simply use the L_1 distance. The convergence in the first term (E.10.13) is immediate by the weak LLN

$$\begin{aligned}
&\sup_{\pi,\pi^0}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(x^{0,N}_{t+1},u^{0,N}_{t+1},\mu^N_{t+1}) - f(x^{0,N}_{t+1},u^{0,N}_{t+1},\hat{\mu}^N_{t+1})\right]\right| \\
&\quad\leq \sup_{\pi,\pi^0} L_f \mathbb{E}\left[\sum_{x\in\mathcal{X}}\left|\mu^N_{t+1}(x) - \hat{\mu}^N_{t+1}(x)\right|\right] \\
&\quad= \sup_{\pi,\pi^0} L_f \sum_{x\in\mathcal{X}} \mathbb{E}_{x^{0,N}_t,u^{0,N}_t,\mu^N_t}\left[\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \mathbf{1}_x(x^{i,N}_{t+1}) - \mathbb{E}_{x^{0,N}_t,u^{0,N}_t,\mu^N_t}\left[\frac{1}{N}\sum_{i=1}^N \mathbf{1}_x(x^{i,N}_{t+1})\right]\right|\right]\right] \\
&\quad\leq L_f|\mathcal{X}|\sqrt{\frac{4}{N}},
\end{aligned}$$

and for the second term (E.10.14) we again use the induction assumption, completing the proof. □

E.11 PROOF OF LEMMA E.10.1

Proof. First note Lipschitz continuity of $(x^0, \mu) \mapsto \mu \otimes \pi_t(x^0, \mu)$ as in Appendix E.5, as for any $(x^0_*, \mu_*), (x^0, \mu) \in \mathcal{X}^0 \times \mathcal{P}(\mathcal{X})$, then

$$\begin{aligned}
&\sup_{\pi\in\Pi} W_1(\mu_* \otimes \pi_t(x^0_*, \mu_*), \mu \otimes \pi_t(x^0, \mu)) \\
&\quad = \sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \left| \int f' \,\mathrm{d}(\mu_* \otimes \pi_t(x^0_*, \mu_*) - \mu \otimes \pi_t(x^0, \mu)) \right| \\
&\quad \le \sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \left| \iint f'(x,u)(\pi_t(\mathrm{d}u \mid x, x^0_*, \mu_*) - \pi_t(\mathrm{d}u \mid x, x^0, \mu))\mu_*(\mathrm{d}x) \right| \\
&\qquad + \sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \left| \iint f'(x,u)\pi_t(\mathrm{d}u \mid x, x^0, \mu)(\mu_*(\mathrm{d}x) - \mu(\mathrm{d}x)) \right|
\end{aligned}$$

where for the first term

$$\begin{aligned}
&\sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \left| \iint f'(x,u)(\pi_t(\mathrm{d}u \mid x, x^0_*, \mu_*) - \pi_t(\mathrm{d}u \mid x, x^0, \mu))\mu_*(\mathrm{d}x) \right| \\
&\quad \le \sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \int \left| \int f'(x,u)(\pi_t(\mathrm{d}u \mid x, x^0_*, \mu_*) - \pi_t(\mathrm{d}u \mid x, x^0, \mu)) \right| \mu_*(\mathrm{d}x) \\
&\quad \le \sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \sup_{x\in\mathcal{X}} \left| \int f'(x,u)(\pi_t(\mathrm{d}u \mid x, x^0_*, \mu_*) - \pi_t(\mathrm{d}u \mid x, x^0, \mu)) \right| \\
&\quad = \sup_{\pi\in\Pi} \sup_{x\in\mathcal{X}} W_1(\pi_t(\cdot \mid x, x^0_*, \mu_*), \pi_t(\cdot \mid x, x^0, \mu)) \\
&\quad \le L_\Pi d((x^0_*, \mu_*), (x^0, \mu))
\end{aligned}$$

by Assumption 4.2.2, and similarly for the second by noting 1-Lipschitzness of $x \mapsto \int \frac{f'(x,u)}{L_\Pi+1}\pi_t(\mathrm{d}u \mid x, x^0, \mu)$, as before in Eq. (E.5.7), and therefore again

$$\begin{aligned}
&\sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \left| \iint f'(x,u)\pi_t(\mathrm{d}u \mid x, x^0, \mu)(\mu_*(\mathrm{d}x) - \mu(\mathrm{d}x)) \right| \\
&\quad = \sup_{\pi\in\Pi} \sup_{\|f'\|_{\mathrm{Lip}}\le 1} (L_\Pi + 1) \left| \iint \frac{f'(x,u)}{L_\Pi + 1}\pi_t(\mathrm{d}u \mid x, x^0, \mu)(\mu_*(\mathrm{d}x) - \mu(\mathrm{d}x)) \right| \\
&\quad \le (L_\Pi + 1) W_1(\mu_*, \mu).
\end{aligned}$$

Hence, the map $(x^0, u^0, \mu) \mapsto \mu \otimes \pi_t(x^0, \mu)$ is Lipschitz with constant $(2L_\Pi + 1)$.

As a result, the entire map $(x^0, u^0, \mu) \mapsto T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu)$ is Lipschitz, since for any

$$\begin{aligned}
&W_1(T(x^0_*, u^0_*, \mu_*, \mu_* \otimes \pi_t(x^0_*, \mu_*)), T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu)) \\
&\quad = \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \left| \iiint f'(x')p(\mathrm{d}x' \mid x, u, x^0_*, u^0_*, \mu_*)\pi_t(\mathrm{d}u \mid x, x^0_*, \mu_*)\mu_*(\mathrm{d}x) \right. \\
&\qquad\qquad \left. - \iiint f'(x')p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)\pi_t(\mathrm{d}u \mid x, x^0, \mu)\mu(\mathrm{d}x) \right| \\
&\quad \le \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \sup_{(x,u)\in\mathcal{X}\times\mathcal{U}} \left| \int f'(x')(p(\mathrm{d}x' \mid x, u, x^0_*, u^0_*, \mu_*) - p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)) \right|
\end{aligned}$$

$$
\begin{aligned}
&+ \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \sup_{x\in\mathcal{X}} \left| \iint f'(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)(\pi_t(\mathrm{d}u \mid x, x^0_*, \mu_*) - \pi_t(\mathrm{d}u \mid x, x^0, \mu)) \right| \\
&+ \sup_{\|f'\|_{\mathrm{Lip}}\le 1} \left| \iiint f'(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)\pi_t(\mathrm{d}u \mid x, x^0, \mu)(\mu_*(\mathrm{d}x) - \mu(\mathrm{d}x)) \right| \\
\le &\sup_{(x,u)\in\mathcal{X}\times\mathcal{U}} W_1(p(\cdot \mid x, u, x^0_*, u^0_*, \mu_*), p(\cdot \mid x, u, x^0, u^0, \mu)) \\
&+ \sup_{x\in\mathcal{X}} (L_p + 1) W_1(\pi_t(\cdot \mid x, x^0_*, \mu_*), \pi_t(\cdot \mid x, x^0, \mu)) \\
&+ \sup_{(x,u)\in\mathcal{X}\times\mathcal{U}} (L_p + L_\Pi + 1) W_1(\mu_*, \mu) \\
\le &\underbrace{(L_p + (L_p + 1)L_\Pi + (L_p + L_\Pi + 1))}_{L_*} d((x^0_*, u^0_*, \mu_*), (x^0, u^0, \mu))
\end{aligned}
$$

with Lipschitz constant $L_T \coloneqq (2L_\Pi + 1) \cdot L_*$ from Assumptions 4.2.1 and 4.2.2, using the same argument as in Eq. (E.5.7). □

E.12 PROOF OF LEMMA E.10.2

Proof. For any $g \in \mathcal{G}$, for any $(x^0_*, u^0_*, \mu_*), (x^0, u^0, \mu) \in \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X})$, let $T_* \coloneqq T(x^0_*, u^0_*, \mu_*, \mu_* \otimes \pi_t(x^0_*, \mu_*))$ and $T^* \coloneqq T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu))$ for brevity. We have

$$
\begin{aligned}
&\left| g(x^0_*, u^0_*, \mu_*) - g(x^0, u^0, \mu) \right| \\
&= \left| \iint f(x', u', T_*) \pi^0_t(\mathrm{d}u' \mid x', T_*) p^0(\mathrm{d}x' \mid x^0_*, u^0_*, \mu_*) \right. \\
&\qquad \left. - \iint f(x', u', T^*) \pi^0_t(\mathrm{d}u' \mid x', T^*) p^0(\mathrm{d}x' \mid x^0, u^0, \mu) \right| \\
&\le \sup_{x', u'} \left| f(x', u', T_*) - f(x', u', T^*) \right| && \text{(E.12.18)} \\
&\quad + \sup_{x'} \left| \int f(x', u', T^*)(\pi^0_t(\mathrm{d}u' \mid x', T_*) - \pi^0_t(\mathrm{d}u' \mid x', T^*)) \right| && \text{(E.12.19)} \\
&\quad + \left| \iint f(x', u', T^*) \pi^0_t(\mathrm{d}u' \mid x', T^*)(p^0(\mathrm{d}x' \mid x^0_*, u^0_*, \mu_*) - p^0(\mathrm{d}x' \mid x^0, u^0, \mu)) \right|. && \text{(E.12.20)}
\end{aligned}
$$

By Lemma E.10.1, for Eq. (E.12.18) we obtain

$$
\begin{aligned}
&\sup_{x', u'} \left| f(x', u', T(x^0_*, u^0_*, \mu_*, \mu_* \otimes \pi_t(x^0_*, \mu_*))) - f(x', u', T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu))) \right| \\
&\quad \le L_{\mathcal{F}} L_T d((x^0_*, u^0_*, \mu_*), (x^0, u^0, \mu)).
\end{aligned}
$$

Similarly for Eq. (E.12.19), by Assumption 4.2.2 we analogously have

$$
\begin{aligned}
&\sup_{x'} \left| \int f(x', u', T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu))) \right. \\
&\quad \left. (\pi^0_t(\mathrm{d}u' \mid x', T(x^0_*, u^0_*, \mu_*, \mu_* \otimes \pi_t(x^0_*, \mu_*))) - \pi^0_t(\mathrm{d}u' \mid x', T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu)))) \right| \\
&\quad \le L_{\mathcal{F}} W_1(\pi^0_t(\cdot \mid x', T(x^0_*, u^0_*, \mu_*, \mu_* \otimes \pi_t(x^0_*, \mu_*))), \pi^0_t(\cdot \mid x', T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu)))) \\
&\quad \le L_{\mathcal{F}} L_{\Pi^0} L_T d((x^0_*, u^0_*, \mu_*), (x^0, u^0, \mu)).
\end{aligned}
$$

Lastly, for Eq. (E.12.20), as before in Eq. (E.5.7), by Assumptions 4.2.1 and 4.2.2 we have again

$$\begin{aligned}
&\left|\iint f(x', u', T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu)))\pi_t^0(\mathrm{d}u' \mid x', T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu))) \right. \\
&\qquad \left. (p^0(\mathrm{d}x' \mid x_*^0, u_*^0, \mu_*) - p^0(\mathrm{d}x' \mid x^0, u^0, \mu))\right| \\
&\leq L_{\mathcal{F}} L_\Pi W_1(p^0(\cdot \mid x_*^0, u_*^0, \mu_*), p^0(\cdot \mid x^0, u^0, \mu)) \\
&\leq L_{\mathcal{F}} L_\Pi L_{p^0} d((x_*^0, u_*^0, \mu_*), (x^0, u^0, \mu)).
\end{aligned}$$

Therefore, $\mathcal{G}$ is equi-Lipschitz with Lipschitz constant $(L_{\mathcal{F}} L_T + L_{\mathcal{F}} L_{\Pi^0} L_T + L_{\mathcal{F}} L_\Pi L_{p^0})$. □

E.13 PROOF OF COROLLARY 4.2.1

Proof. As in Lemma E.6.1, for any $\varepsilon > 0$, choose time $T \in \mathbb{N}$ such that

$$\sum_{t=T}^{\infty} \gamma^t \left|\mathbb{E}\left[r(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - r(x_t^0, u_t^0, \mu_t)\right]\right| \leq \frac{\gamma^T}{1-\gamma} \max_\mu 2|r(\mu)| < \frac{\varepsilon}{2}.$$

By Theorem 4.2.2,

$$\sum_{t=0}^{T-1} \gamma^t \left|\mathbb{E}\left[r(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - r(x_t^0, u_t^0, \mu_t)\right]\right| < \frac{\varepsilon}{2}$$

for sufficiently large N. Therefore, $\sup_{(\pi,\pi^0)\in\Pi\times\Pi^0} \left|J^N(\pi, \pi^0) - J(\Phi^{-1}(\pi), \pi^0)\right| \to 0$.

As a result, we have

$$\begin{aligned}
J^N(\Phi(\hat{\pi}^*), \pi^{0*}) - \sup_{(\pi,\pi^0)\in\Pi\times\Pi^0} J^N(\pi, \pi^0) &= \inf_{(\pi,\pi^0)\in\Pi\times\Pi^0} (J^N(\Phi(\hat{\pi}^*), \pi^{0*}) - J^N(\pi, \pi^0)) \\
&\geq \inf_{(\pi,\pi^0)\in\Pi\times\Pi^0} (J^N(\Phi(\hat{\pi}^*), \pi^{0*}) - J(\hat{\pi}^*, \pi^{0*})) \\
&\quad + \inf_{(\pi,\pi^0)\in\Pi\times\Pi^0} (J(\hat{\pi}^*, \pi^{0*}) - J(\pi, \pi^0)) \\
&\quad + \inf_{(\pi,\pi^0)\in\Pi\times\Pi^0} (J(\pi, \pi^0) - J^N(\pi, \pi^0)) \\
&\geq -\frac{\varepsilon}{2} + 0 - \frac{\varepsilon}{2} = -\varepsilon
\end{aligned}$$

for sufficiently large N, where the second term is zero by optimality of $(\hat{\pi}^*, \pi^{0*})$ in the M3FC problem. □

E.14 PROOF OF THEOREM 4.2.3

First, for completeness we give the finite M3FC system equations under the assumed Lipschitz parametrization for joint stationary M3FMARL policies $\tilde{\pi}^\theta$ used during centralized training with correlated minor agent actions. Note that deterministic joint policies $\tilde{\pi}^\theta$ (e.g. at convergence, or if using deterministic policy gradients [257]) are equivalent to using separate deterministic minor and major policies in Eq. (4.2.29), see also Remark 4.2.3. The finite M3FC system equations are then given as

$$u_t^{0,N}, \xi_t^N \sim \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N), \quad \pi_t'^N = \Gamma(\xi_t^N), \quad u_t^{i,N} \sim \pi_t'^N(u_t^{i,N} \mid x_t^{i,N}),$$

$$x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N), \quad x_{t+1}^{0,N} \sim p^0(x_{t+1}^{0,N} \mid x_t^{0,N}, u_t^{0,N}, \mu_t^N),$$

as well as the limiting M3FC MDP under such parametrization as

$$\begin{aligned} u_t^0, \xi_t &\sim \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t), \quad \pi_t' = \Gamma(\xi_t), \quad h_t = \mu_t \otimes \pi_t', \\ \mu_{t+1} &= T(x_t^0, u_t^0, \mu_t, h_t), \quad x_{t+1}^0 \sim p^0(x_{t+1}^0 \mid x_t^0, u_t^0, \mu_t). \end{aligned}$$

Then, by [221], the exact policy gradient for the limiting M3FC MDP is given as

$$\nabla_\theta J(\tilde{\pi}^\theta) = \sum_{t=T}^{\infty} \gamma^t \mathbb{E}\left[Q^\theta(x_t^0, \mu_t, u_t^0, \xi_t) \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t) \right]$$

under the action-value function

$$Q^\theta(x^0, \mu, u^0, \xi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^0, u_t^0, \mu_t) \,\middle|\, x_0^0 = x^0, \mu_0 = \mu, u_0^0 = u^0, \xi_0 = \xi \right],$$

while the approximation for the policy gradient on the finite M3FC system is given instead by

$$\widehat{\nabla_\theta J}(\tilde{\pi}^\theta) = \sum_{t=T}^{\infty} \gamma^t \mathbb{E}\left[\widehat{Q}^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \,\middle|\, x_t^{0,N}, \mu_t^N) \right]$$

and the finite-agent action-values

$$\widehat{Q}^\theta(x^0, \mu, u^0, \xi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^{0,N}, u_t^{0,N}, \mu_t^N) \,\middle|\, x_0^{0,N} = x^0, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi \right],$$

which are obtained, e.g., by on-policy samples and using critic estimates. Note that here, the conditional expectations are given by redefining the systems (4.2.29) and (4.2.31) with the values conditioned upon.

We then show that the approximation of the policy gradient is good for large systems, i.e.

$$\left\| \widehat{\nabla_\theta J}(\tilde{\pi}^\theta) - \nabla_\theta J(\hat{\pi}^\theta) \right\| \to 0 \tag{E.14.21}$$

as $N \to \infty$, uniformly over all current policy parameters θ.

Proof of Theorem 4.2.3. We use the following lemmas in the proof of Theorem 4.2.3, for which the proofs are given below.

Proposition E.14.1. *Propagation of chaos holds for the M3FC systems with parameterized actions as in Theorem 4.2.2, i.e. under Assumptions 4.2.1, 4.2.2 and 4.2.3, for any equi-Lipschitz family $\mathcal{F}$, at all times $t \in \mathbb{N}$ uniformly,*

$$\sup_{f, \pi, \pi^0} \left| \mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - f(x_t^0, u_t^0, \mu_t) \right] \right| \to 0. \tag{E.14.22}$$

Proposition E.14.2. *Under Assumptions 4.2.1 and 4.2.2, the approximate action-values converge uniformly, $\widehat{Q}^\theta \to Q^\theta$ as $N \to \infty$.*

As a result, we obtain

$$
\begin{aligned}
&\left\|\widehat{\nabla_\theta J}(\tilde{\pi}^\theta) - \nabla_\theta J(\hat{\pi}^\theta)\right\| \\
&= \Bigg\| \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\Big[\widehat{Q}^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) \\
&\qquad\qquad - Q^\theta(x_t^0, \mu_t, u_t^0, \xi_t) \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t) \Big] \Bigg\| \\
&\leq \left\| \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[\left(\widehat{Q}^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) - Q^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) \right) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) \right] \right\| \\
&\quad + \Bigg\| \sum_{t=T}^{\infty} \gamma^t \mathbb{E}\Big[Q^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) \\
&\qquad\qquad - Q^\theta(x_t^0, \mu_t, u_t^0, \xi_t) \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t) \Big] \Bigg\| \\
&\quad + \Bigg\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\Big[Q^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) \\
&\qquad\qquad - Q^\theta(x_t^0, \mu_t, u_t^0, \xi_t) \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t) \Big] \Bigg\|
\end{aligned}
$$

for any T, such that the first term disappears by Assumption 4.2.3 uniformly bounding $\nabla_\theta \log \tilde{\pi}^\theta$ and Proposition E.14.2. Note that we bounded $\nabla_\theta \log \tilde{\pi}^\theta$ here, but we can also assume bounded gradients $\nabla_\theta \tilde{\pi}^\theta$ instead, e.g. Eq. (E.14.23).

For the second term, we similarly uniformly bound $\nabla_\theta \log \tilde{\pi}^\theta$ by Assumption 4.2.3 and Q by Assumption 4.2.1, then choose T sufficiently large.

Finally, for the last term, we note that we can write the difference as

$$
\begin{aligned}
&\Bigg\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\Big[Q^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) \\
&\qquad\qquad - Q^\theta(x_t^0, \mu_t, u_t^0, \xi_t) \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t) \Big] \Bigg\| \\
&= \Bigg\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\Bigg[\sum_{t'=0}^{\infty} \gamma^t \mathbb{E}\left[r(x_{t'}^{0\prime}, u_{t'}^{0\prime}, \mu_{t'}^\prime) \;\middle|\; x_0^{0\prime} = x_t^{0,N}, \mu_0^\prime = \mu_t^N, u_0^{0\prime} = u_t^{0,N}, \xi_0^\prime = \xi_t^N \right] \\
&\qquad\qquad \cdot \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) \\
&\qquad\quad - \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[r(x_{t'}^{0\prime}, u_{t'}^{0\prime}, \mu_{t'}^\prime) \mid x_0^{0\prime} = x_t^0, \mu_0^\prime = \mu_t, u_0^{0\prime} = u_t^0, \xi_0^\prime = \xi_t \right] \\
&\qquad\qquad \cdot \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t) \Big] \Bigg\| \\
&\leq \Bigg\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\Bigg[\sum_{t'=T'}^{\infty} \gamma^t \mathbb{E}\left[r(x_{t'}^{0\prime}, u_{t'}^{0\prime}, \mu_{t'}^\prime) \;\middle|\; x_0^{0\prime} = x_t^{0,N}, \mu_0^\prime = \mu_t^N, u_0^{0\prime} = u_t^{0,N}, \xi_0^\prime = \xi_t^N \right] \\
&\qquad\qquad \cdot \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) \\
&\qquad\quad - \sum_{t=T'}^{\infty} \gamma^t \mathbb{E}\left[r(x_{t'}^{0\prime}, u_{t'}^{0\prime}, \mu_{t'}^\prime) \mid x_0^{0\prime} = x_t^0, \mu_0^\prime = \mu_t, u_0^{0\prime} = u_t^0, \xi_0^\prime = \xi_t \right] \\
&\qquad\qquad \cdot \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t) \Big] \Bigg\| \\
&\quad + \Bigg\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}\Bigg[\sum_{t'=0}^{T'-1} \gamma^t \mathbb{E}\left[r(x_{t'}^{0\prime}, u_{t'}^{0\prime}, \mu_{t'}^\prime) \;\middle|\; x_0^{0\prime} = x_t^{0,N}, \mu_0^\prime = \mu_t^N, u_0^{0\prime} = u_t^{0,N}, \xi_0^\prime = \xi_t^N \right] \\
&\qquad\qquad \cdot \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) \\
&\qquad\quad - \sum_{t=0}^{T'-1} \gamma^t \mathbb{E}\left[r(x_{t'}^{0\prime}, u_{t'}^{0\prime}, \mu_{t'}^\prime) \mid x_0^{0\prime} = x_t^0, \mu_0^\prime = \mu_t, u_0^{0\prime} = u_t^0, \xi_0^\prime = \xi_t \right]
\end{aligned}
$$

$$\cdot \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t)\big]\big\|$$

where we write the conditional M3FC system and random variables in the inner expectation with a prime, bounding again the former terms by choosing sufficiently large T' and using Assumptions 4.2.1 and 4.2.3, while for the latter terms we use Proposition E.14.1 on the functions

$$f(x^0, \mu) = \iint \mathbb{E}\left[r(x_{t'}^{0\prime}, u_{t'}^{0\prime}, \mu_{t'}') \mid x_0^{0\prime} = x^0, \mu_0 = \mu, u_0^{0\prime} = u^0, \xi_0' = \xi\right] \nabla_\theta \tilde{\pi}^\theta(u^0, \xi \mid x^0, \mu) \mathrm{d}(u^0, \xi) \tag{E.14.23}$$

for all t', which are uniformly Lipschitz by Assumptions 4.2.1 and 4.2.3. This completes the proof. □

E.15 PROOF OF PROPOSITION E.14.1

Proof. The proof is exactly analogous to the proof of Theorem 4.2.2, except that instead of using Lipschitz constants of $x_t^0, u_t^0, \mu_t, h_t \mapsto T(x_t^0, u_t^0, \mu_t, h_t)$, one uses Lipschitz constants of $x_t^0, u_t^0, \mu_t, \xi_t \mapsto T(x_t^0, u_t^0, \mu_t, \mu_t \otimes \Gamma(\xi_t))$ via the additional Assumption 4.2.3 on top of Assumptions 4.2.1 and 4.2.2. □

E.16 PROOF OF PROPOSITION E.14.2

Proof. To show $\widehat{Q}^\theta \to Q^\theta$ as $N \to \infty$ uniformly, it suffices to prove pointwise convergence due to compact support.

Therefore, fix any x^0, μ, u^0, ξ. The convergence follows as in Corollary 4.2.1, from showing at any time t that

$$\sup_{f \in \mathcal{F}} \Big|\mathbb{E}\left[f(x_t^0, u_t^0, \mu_t) \mid x_0^0 = x^0, \mu_0 = \mu, u_0^0 = u^0, \xi_0 = \xi\right] - \mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N) \,\Big|\, x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi\right]\Big| \to 0$$

over any equi-Lipschitz family of functions $\mathcal{F}$, and applying for $f = r$ (using the set $\mathcal{F}$ of L_r-Lipschitz functions) by Assumption 4.2.1.

The statement is shown by considering time $t = 0$, and then by induction for any $t \geq 1$. At time $t = 0$, the statement follows from the weak LLN as in Theorem 4.2.2. For any subsequent times, we similarly have

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} \Big|\mathbb{E}\left[f(x_{t+1}^0, u_{t+1}^0, \mu_{t+1}) \mid x_0^0 = x^0, \mu_0 = \mu, u_0^0 = u^0, \xi_0 = \xi\right] \\
&\qquad - \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) \,\Big|\, x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi\right]\Big| \\
&\leq \sup_{f \in \mathcal{F}} \Big|\mathbb{E}\left[f(x_{t+1}^0, u_{t+1}^0, \mu_{t+1}) \mid x_0^0 = x^0, \mu_0 = \mu, u_0^0 = u^0, \xi_0 = \xi\right] \\
&\quad - \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, T(x_t^{0,N}, u_t^{0,N}, \mu_t^N, \mu_t^N \otimes \Gamma(\xi_t^N))) \,\Big|\, x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi\right]\Big| \\
&+ \sup_{f \in \mathcal{F}} \Big|\mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, T(x_t^{0,N}, u_t^{0,N}, \mu_t^N, \mu_t^N \otimes \Gamma(\xi_t^N))) \,\Big|\, x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi\right] \\
&\quad - \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) \,\Big|\, x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi\right]\Big|.
\end{aligned}$$

As in Theorem 4.2.2, the latter term is bounded by induction assumption, using uniform Lipschitzness of the dynamics, $x_t^0, u_t^0, \mu_t, \xi_t \mapsto T(x_t^0, u_t^0, \mu_t, \mu_t \otimes \Gamma(\xi_t))$ via Assumptions 4.2.2 and 4.2.3, while the former term is bounded as usual by the weak LLN. This completes the proof. □

E.17 EXTENDED MFC OPTIMALITIES

Intuitively, in large MF systems governed by dynamics of the form (4.2.29), almost all information of the joint state $(x_t^{0,N}, x_t^{1,N}, \dots, x_t^{N,N})$ is contained in $(x_t^{0,N}, \mu_t^N)$, while heterogeneous policies should by LLN be replaceable by a shared one. To fully complete the theory of MFC, it is therefore interesting to establish the optimality of the considered MF policies over arbitrary other policies acting on the joint state $(x_t^{0,N}, x_t^{1,N}, \dots, x_t^{N,N})$.

It seems plausible that it would be possible to extend optimality (Corollary 4.2.1) over larger classes of policies in the finite system. In particular, at least for finite state-action spaces, (i) any joint-state policy $\pi(\mathrm{d}u \mid x_t^{0,N}, x_t^{1,N}, \dots, x_t^{N,N})$ might in the limit be replaced by an averaged policy $\bar{\pi}(\mathrm{d}u \mid x^0, \mu) \coloneqq \sum_{x^N \in \mathcal{X}^N \colon \frac{1}{N}\sum_i \delta_{x^{i,N}} = \mu} \pi(\mathrm{d}u \mid x^0, x^N)$ under some exchangeability of agents; (ii) any optimal policy π outputting joint actions for all agents might be replaced by an independent but identical policy for each agent, as in the limit all information is contained in the joint state-action distribution, any of which may be approximated increasingly closely by LLN; and (iii) heterogeneous policies for each minor agent $\pi^1, \dots, \pi^N$ might similarly be replaced by some averaged policy $\bar{\pi}(\pi^1, \dots, \pi^N)$, averaging the action distributions in any specific state over the proportion of agent likelihoods in that state.

Showing such results would allow us to conclude that the policy classes Π are natural and sufficient in MF systems, including MFC and also the competitive MFGs, as more general or heterogeneous policies will not perform much better. A result related to (iii) has been shown for static cases [8, 386] and more recently in MFC and its two-team generalizations [128].

E.18 EXPERIMENTAL DETAILS

In this section, we give lengthy experimental details that were omitted in the main text. Hyperparameters are given in Table E.1.

TABLE E.1: Shared hyperparameter configurations for all algorithms.

Symbol	Name	Value
γ	Discount factor	0.99
λ	GAE lambda	1
β	KL coefficient	0.03
ϵ	Clip parameter	0.2
l_r	Learning rate	0.00005
B_{len}	Training batch size	24000
b_{len}	Mini-batch size	4000
N_{SGD}	Gradient steps per training batch	8

E.18.1 *Problem Details*

In this section, we give details to the problems considered. We omit superscript N if clear from context.

2G. In the 2G problem, we formally let $\mathcal{X} = [-2, 2]^2$, $\mathcal{U} = [-1, 1]^2$, $\mathcal{X}^0 = \{0, 1, \dots 49\}$ according to Eq. (E.7.9). We allow noisy movement of minor agents following the Gaussian law

$$p(x^i_{t+1} \mid x^i_t, u^i_t) = \mathcal{N}\left(x^i_{t+1} \;\middle|\; x^i_t + v_{\max}\frac{u^i_t}{\max(1, \|u^i_t\|_2)}, \operatorname{diag}(\sigma^2, \sigma^2)\right)$$

for some maximum speed $v_{\max} = 0.2$, noise covariance $\sigma^2 = 0.03$ and projecting back actions u with norm larger than 1, with the additional modification that agent positions are clipped back into $\mathcal{X}$ whenever the agents move out of bounds.

We then consider a time-variant mixture of two Gaussians

$$\mu^*_t \coloneqq \frac{1 + \cos(2\pi t/50)}{2}\mathcal{N}\left(\mathbf{e}_1, \operatorname{diag}(\sigma^2_*, \sigma^2_*)\right) + \frac{1 - \cos(2\pi t/50)}{2}\mathcal{N}\left(-\mathbf{e}_1, \operatorname{diag}(\sigma^2_*, \sigma^2_*)\right)$$

for unit vector $\mathbf{e}_1$ and covariance $\sigma^2_* = 0.05$, i.e. we have a period of 50 time steps, and let the major state follow the clock dynamics $p^0(x^0 + 1 \mod 50 \mid x^0, \mu) = 1$.

The goal of minor agents is to minimize the Wasserstein metric $\hat{W}_1$ under the squared Euclidean distance,

$$\hat{W}_1(\mu, \mu') \coloneqq \inf_{\gamma \in \Gamma(\mu, \mu')}\left\{\int \|x - y\|^2_2 \gamma(\mathrm{d}x, \mathrm{d}y)\right\}$$

defined over all couplings $\Gamma(\mu, \mu')$ with first and second marginals μ, μ' (which is strictly speaking not a metric but an optimal transportation cost, since the squared Euclidean distance fails the triangle inequality), between their empirical distribution and the desired mixture of Gaussians

$$r(x^0_t, \mu_t) = -\hat{W}_1(\mu_t, \mu^*_t)$$

which is computed numerically by the empirical distance, sampling 300 samples from μ^*_t.

The initialization of minor agents is uniform, i.e. $\mu_0 = \operatorname{Unif}(\mathcal{X})$, and $x^0_0 = 0$. For sake of simulation, we define the episode length $T = 100$ after which a new episode starts.

FORMATION. The Formation problem is an extension of the 2G problem, where instead $\mathcal{X}^0 = \mathcal{X} \times \mathcal{X}$ and $\mathcal{U}^0 = \mathcal{U}$, the major agent follows the same dynamics as the minor agents, and movements are noise-free, i.e. $\sigma^2 = 0$. The major agent state $x^0_t = (\hat{x}^0_t, x^*_t)$ here contains both the major agent position $\hat{x}^0_t$ and its target position x^*_t. The desired minor agent distribution is centered around the major agent

$$\mu^*_t \coloneqq \mathcal{N}\left(\hat{x}^0_t, \operatorname{diag}(\sigma^2_*, \sigma^2_*)\right)$$

with covariance $\sigma^2_* = 0.3$, and is also observed by agents as in 2G via binning. Additionally, the major agent should follow a random target x^*_t following discretized Ornstein-Uhlenbeck dynamics

$$x^*_{t+1} \sim \mathcal{N}\left(0.95x^*_t, \operatorname{diag}(\sigma^2_{\text{targ}}, \sigma^2_{\text{targ}})\right)$$

with $\sigma^2_{\text{targ}} = 0.02$. Thus, similar to 2G, the reward function becomes

$$r(x^0_t, u^0_t, \mu_t) = -\|\hat{x}^0_t - x^*_t\|_2 - \hat{W}_1(\mu_t, \mu^*_t).$$

The initialization of agents is uniform, while the target starts around zero, i.e. $\mu_0 = \operatorname{Unif}(\mathcal{X})$ and $\mu^0_0 = \operatorname{Unif}(\mathcal{X}) \otimes \mathcal{N}\left(0, \operatorname{diag}(\sigma^2_{\text{targ}}, \sigma^2_{\text{targ}})\right)$. For sake of simulation, we define the episode length $T = 100$ after which a new episode starts.

BEACH BAR PROCESS. In the discrete beach bar process, we consider a discrete torus $\mathcal{X} = \{0, 1, \dots, 4\}^2$, $\mathcal{X}^0 = \mathcal{X} \times \mathcal{X}$ and actions $\mathcal{U} = \mathcal{U}^0 = \{(0,0), (-1,0), (0,-1), (1,0), (0,1)\}$ indicating movement in any of the four cardinal directions. The major agent state $x_t^0 = (\hat{x}_t^0, x_t^*)$ here contains both the major agent position $\hat{x}_t^0$ and its target position x_t^*. In other words, the dynamics follow

$$\hat{x}_{t+1}^0 = \hat{x}_t^0 + u_t^0 \mod (5,5), \quad x_{t+1}^i = x_t^i + u_t^i \mod (5,5).$$

The target position follows a random walk on the torus

$$x_{t+1}^* \sim x_t^* + \epsilon_t \mathrm{Unif}((-1,0),(0,-1),(1,0),(0,1)) \mod (5,5)$$

with walking probability $\epsilon_t \sim \mathrm{Bernoulli}(0.2)$, uniformly in any direction.

The costs are then given by the average toroidal distance d (the L_1 "wrap-around" distance on the torus) between the major agent and its target, the average distance between major and minor agents, and the crowdedness of agents

$$r(x_t^0, u_t^0, \mu_t) = -0.5 d(x_t^0, x_t^*) - 2.5 \int d(x, x_t^0) \mu_t(\mathrm{d}x) - 6.25 \int \mu_t(x) \mu_t(\mathrm{d}x).$$

The initialization of agents is uniform, while the target starts at zero, i.e. $\mu_0 = \mathrm{Unif}(\mathcal{X})$ and $\mu_0^0 = \mathrm{Unif}(\mathcal{X}) \otimes \delta_{(0,0)}$. For sake of simulation, we define the episode length $T = 200$ after which a new episode starts.

For the neural network policy, we use a one-hot encoding of major states as input, i.e. the concatenation of two 5-dimensional one-hot vectors for the major agent position $\hat{x}_t^0$ and its target position x_t^* respectively.

FORAGING. In the Foraging problem, we formally define $\mathcal{X} = [-2,2]^2 \times [0,1], \mathcal{U} = [-1,1]^2 = \mathcal{U}^0$ and $\mathcal{X}^0 = ([-2,2] \times [-2,-1]) \times \bigcup_{n=0}^5 \left([-2,2]^2 \times [0,1.5]\right)^n$. The minor agent states $x_t^i = (\hat{x}_t^i, \tilde{x}_t^i)$ here contain their positions $\hat{x}_t^i \in [-2,2]^2$ and encumbrance (or inversely, free cargo space) $\hat{x}_t^i \in [0,1]$. Meanwhile, the major agent state $x_t^0 = (\hat{x}_t^0, x_t^{\mathrm{env}})$ here contains both the major agent position $\hat{x}_t^0$ restricted to $[-2,2] \times [-2,-1]$, and the current environment state x_t^{env}. Here, the minor and major agents move as in Formation, though with different maximum velocities for minor agents $v_{\max} = 0.3$ and major agent $v_{\max}^0 = 0.1$ respectively.

An additional environmental state consists of up to 5 spatially localized foraging areas, which is not observed by the agents. In each time step, $N_t = \mathrm{Pois}(0.2)$ new foraging areas appear, up to a maximum total number of 5. The location x_t^m of each foraging area $m = 1, \dots, 5$ is sampled uniformly randomly from $\mathrm{Unif}(\mathcal{X})$, while their total initial size L_t^m is sampled from $\mathrm{Unif}([0.5, 1.5])$, making up the environment state $x_t^{\mathrm{env}} = (x_t^m, L_t^m)_m$. At every time step, the foraging areas m are depleted by nearby agents closer than range 0.5,

$$L_{t+1}^m = L_t^m - \Delta L^m(\mu_t),$$
$$\Delta L^m(\mu_t) \coloneqq \min(L_{t+1}^m - L_t^m, \min(0.1, \int (0.5 - \|x - x_t^m\|_2)^+ \mu_t(\mathrm{d}x))$$

where $(\cdot)^+ \coloneqq \max(0, \cdot)$, until they are fully depleted and disappear ($L_{t+1}^m \leq 0$).

Foraging minor agents simulate encumbrance, gaining it from nearby foraging areas and depositing to a nearby major agent, by splitting the foraged amount among all nearby minor agents according to their foraged contribution, and wasting any amount going beyond maximum encumbrance 1,

$$\tilde{x}_{t+1}^i = \begin{cases} \min(1, \tilde{x}_t^i + \Delta L^m(\mu_t) \cdot \frac{(0.5 - \|x - x_t^m\|_2)^+}{\int (0.5 - \|x - x_t^m\|_2)^+ \mu_t(\mathrm{d}x)}) & \text{if} \quad \|x_t^i - x_t^0\|_2 \geq 0.5, \\ 0 \quad \text{else.} \end{cases}$$

The reward at each time step is then given by the according total foraged and then deposited amount by the minor agents, where any clipped amount is wasted.

The initialization of agents is uniform, while the environment starts empty, i.e. $\mu_0 = \mathrm{Unif}(\mathcal{X})$ and $\mu_0^0 = \mathrm{Unif}(\mathcal{X}) \otimes \delta_\emptyset$. For sake of simulation, we define the episode length $T = 200$ after which a new episode starts.

POTENTIAL. Lastly, in Potential we consider minor agents on a continuous one-dimensional torus $\mathcal{X} = [-2, 2]$ (where the points -2 and 2 are identified), actions $\mathcal{U} = [-1, 1]$ and major state $\mathcal{X}^0 = \mathcal{X} \times \mathcal{X}$. The minor agents move as in Foraging (wrapping around the torus instead of clipping), while the major agent follows the gradient of the potential landscape generated by minor agents, with the goal of staying close to its current target. The major agent state $x_t^0 = (\hat{x}_t^0, x_t^*)$ here contains both the major agent position $\hat{x}_t^0$ and its target position x_t^*. For simplicity, here we use a linear repulsive force decreasing from $\frac{1}{N}$ to 0 over a range of 1,

$$\hat{x}_{t+1}^0 = \hat{x}_t^0 + \frac{1}{20} \sum_{x_{\mathrm{off}} \in \{-4,0,4\}} \int (1 - \|\hat{x}_t^0 - x + x_{\mathrm{off}}\|_2)^+ \frac{\hat{x}_t^0 - x + x_{\mathrm{off}}}{\|\hat{x}_t^0 - x + x_{\mathrm{off}}\|_2} \mu_t(\mathrm{d}x) \mod [-2, 2]$$

where we let terms $0/0 = 0$ and use the offset x_{off} to account for the wrap-around on the torus.

The target follows the discretized Ornstein-Uhlenbeck process

$$x_{t+1}^* \sim \mathcal{N}\left(0.99 x_t^*, \mathrm{diag}(\sigma_{\mathrm{targ}}^2, \sigma_{\mathrm{targ}}^2)\right)$$

with covariance $\sigma_{\mathrm{targ}}^2 = 0.005$, and gives rise to the reward function via the toroidal distance between target and major agent

$$r(x_t^0, \mu_t) = -d(\hat{x}_t^0, x_t^*).$$

The initialization of agents is uniform, while the target starts around zero, i.e. $\mu_0 = \mathrm{Unif}(\mathcal{X})$ and $\mu_0^0 = \mathrm{Unif}(\mathcal{X}) \otimes \mathcal{N}\left(0, \mathrm{diag}(\sigma_{\mathrm{targ}}^2, \sigma_{\mathrm{targ}}^2)\right)$. For sake of simulation, we define the episode length $T = 100$ after which a new episode starts. In contrast to $M = 7^2 = 49$ in 2G, Formation and Foraging, here we use $M = 7$ bins for the one-dimensional problem.

E.18.2 *Comparison to M3FA2C*

In Figure E.1 we can see that vanilla M3FA2C typically performs worse than M3FPPO, getting stuck in worse local optima. Here, we used the same hyperparameters as in PPO. This validates our choice of PPO for M3FMARL.

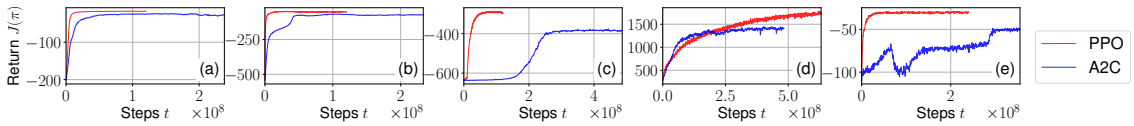


FIGURE E.1: Training curves (mean episode return vs. time steps) of M3FPPO in red, compared to M3FA2C in blue. (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

E.18.3 *Qualitative Results*

In Figure E.2, M3FPPO successfully learns to form mixtures of Gaussians in 2G, and a Gaussian around a moving major agent that tracks its target in Formation. As expected in 2G, the two Gaussians at their sinusoidal peaks $t = 25$ and $t = 50$ are not perfectly tracked, in order to minimize the cost in following time steps, when the other Gaussian reappears. Finally, in Potential the minor agents succeed in pushing the major agent towards its target, while spreading on both sides of the major agent to be able to track any random movement of the target.

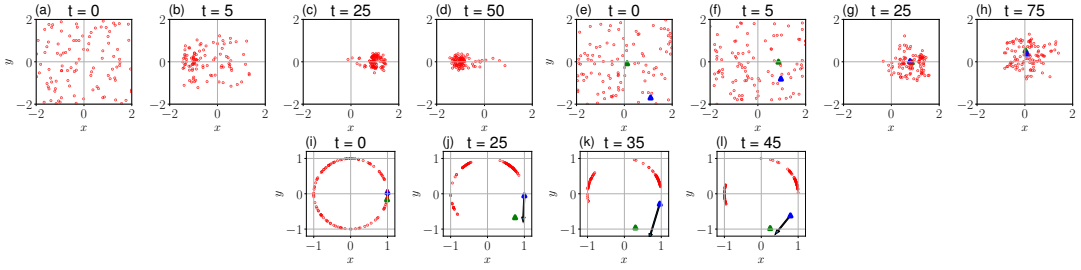


FIGURE E.2: Qualitative visualization of learned M3FC behavior in the 2G (a-d), Formation (e-h) and Potential (i-l) problems. Red: minor agent; blue triangle: major agent; green triangle: major agent target. (i-l): As in (e-h), with arrow for potential gradient (not to scale).

E.18.4 *Training M3FPPO, IPPO and MAPPO on smaller systems*

In Figure 4.10 we verified the training of M3FPPO on small finite system. Comparing to Figures 4.9 and 4.13, for M3FPPO we see little difference between training on a small finite-agent system versus training on a large system and applying the policy on the smaller system. For the chosen hyperparameters, the performance in the Potential problem depends on the initialization. However, M3FPPO compares especially favorably to IPPO in Beach and Foraging, even when directly training on the finite system. This shows that we can either (i) directly apply M3FPPO as a MARL algorithm to small systems, or (ii) train on a fixed system, and transfer the learned behavior to systems of almost arbitrary other sizes.

Analogously, in Figures E.3 and E.4 we show the training results for around a day of IPPO and MAPPO for numbers of agents $N = 5$, $N = 10$ and $N = 20$. As seen in the plot, the results for each number of agents is comparable to the analysis shown in the main text. In particular, transferring M3FPPO or comparing with Figure 4.10, we observe that M3FPPO continues to outperform or match the performance of IPPO and MAPPO, even in the setting with fewer agents.

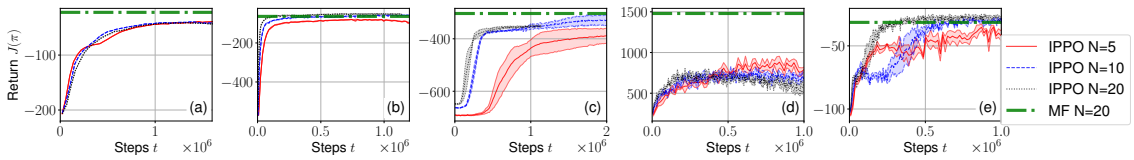


FIGURE E.3: Training curves (mean episode return vs. time steps) of IPPO, trained on the systems with $N \in \{5, 10, 20\}$. (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

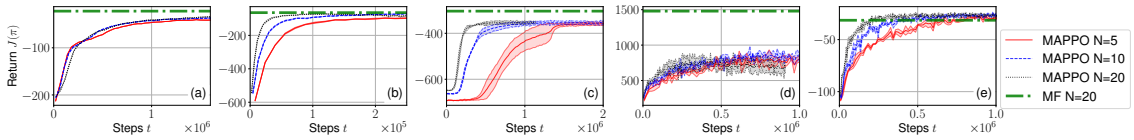


FIGURE E.4: Training curves (mean episode return vs. time steps) of MAPPO, trained on the systems with $N \in \{5, 10, 20\}$. (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

APPENDIX F: SUPPLEMENTARY DETAILS ON SECTION 4.3

F.1 PROOF OF THEOREM 4.3.1

Proof of Theorem 4.3.1. As in the main text, we usually equip $\mathcal{P}(\mathcal{X})$ with the 1-Wasserstein distance. In the proof, however, it is useful to also consider the uniformly equivalent metric $d_\Sigma(\mu,\mu') \coloneqq \sum_{m=1}^{\infty} 2^{-m} \left| \int f_m \,\mathrm{d}(\mu-\mu') \right|$ instead. Here, $(f_m)_{m\geq 1}$ is a fixed sequence of continuous functions $f_m \colon \mathcal{X} \to [-1,1]$, see e.g. [216, Theorem 6.6] for details.

First, let us define the measure $\zeta_t^{\pi,\mu}$ on $\mathcal{X}\times\mathcal{U}$, defined for any measurable set $A\times B \subseteq \mathcal{X}\times\mathcal{U}$ by $\zeta_t^{\pi,\mu}(A\times B) \coloneqq \int_A \int_{\mathcal{Y}} \int_B \pi_t(\mathrm{d}u \mid y) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu_t)\mu_t(\mathrm{d}x)$. For notational convenience we define the MF transition operator $\tilde{T}$ such that

$$\tilde{T}(\mu_t, \zeta_t^{\pi,\mu}) \coloneqq \iint p(\cdot \mid x,u,\mu_t)\zeta_t^{\pi,\mu}(\mathrm{d}x,\mathrm{d}u) = \mu_{t+1}. \tag{F.1.1}$$

Continuity of $\tilde{T}$ follows immediately from Assumption 4.3.1a and [1, Lemma 2.5] which we recall here for convenience.

Proposition F.1.1 ([1], Lemma 2.5)**.** *Under Assumption 4.3.1a, $(\mu_n,\zeta_n) \to (\mu,\zeta)$ implies $\tilde{T}(\mu_n,\zeta_n) \to \tilde{T}(\mu,\zeta)$.*

The rest of the proof is similar to [1, Theorem 2.7] – though we remark that we strengthen the convergence statement from weak convergence to convergence in L_1 uniformly over $f\in\mathcal{F}$ – by showing via induction over t that

$$\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_t^N) - f(\mu_t)\right|\right] \to 0. \tag{F.1.2}$$

Note that the induction start can be verified by a weak LLN argument which is also leveraged in the subsequent induction step. For the induction step we assume that Eq. (F.1.2) holds at time t. At time $t+1$ we have

$$\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_{t+1}^N) - f(\mu_{t+1})\right|\right] \\
&\quad\leq \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f(\mu_{t+1}^N) - f\left(\tilde{T}\left(\mu_t^N, \zeta_t^{\pi,\mu^N}\right)\right)\right|\right] &\text{(F.1.3)}\\
&\qquad+ \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \mathbb{E}\left[\left|f\left(\tilde{T}\left(\mu_t^N, \zeta_t^{\pi,\mu^N}\right)\right) - f(\mu_{t+1})\right|\right]. &\text{(F.1.4)}
\end{aligned}$$

We start by analyzing the first term and recall that a modulus of continuity $\omega_{\mathcal{F}}$ of $\mathcal{F}$ is defined as a function $\omega_{\mathcal{F}} \colon [0,\infty) \to [0,\infty)$ with both $\lim_{x\to 0}\omega_{\mathcal{F}}(x) = 0$ and $|f(\mu)-f(\nu)| \leq \omega_{\mathcal{F}}(W_1(\mu,\nu)), \forall f\in\mathcal{F}$. By [353, Lemma 6.1], such a non-concave and decreasing modulus $\omega_{\mathcal{F}}$ exists for $\mathcal{F}$ because it is uniformly equicontinuous due to the compactness of $\mathcal{P}(\mathcal{X})$. Analogously, we have that $\mathcal{F}$ is uniformly equicontinuous in the space $(\mathcal{P}(\mathcal{X}), d_\Sigma)$ as well. Recalling that $\mathcal{P}(\mathcal{X})$ is compact and the topology of weak convergence is metrized by both d_Σ and W_1, we know that the identity map $\mathrm{id}\colon (\mathcal{P}(\mathcal{X}), d_\Sigma) \to (\mathcal{P}(\mathcal{X}), W_1)$ is uniformly continuous. Leveraging the above findings, we have that for the identity map there exists a modulus of continuity $\tilde{\omega}$ such that

$$|f(\mu)-f(\nu)| \leq \omega_{\mathcal{F}}(W_1(\mathrm{id}\,\mu, \mathrm{id}\,\nu)) \leq \omega_{\mathcal{F}}(\tilde{\omega}(d_\Sigma(\mu,\nu)))$$

holds for all $\mu,\nu\in(\mathcal{P}(\mathcal{X}), d_\Sigma)$. By [353, Lemma 6.1], we can use the least concave majorant of $\tilde{\omega}_{\mathcal{F}} \coloneqq \omega_{\mathcal{F}}\circ\tilde{\omega}$ instead of $\tilde{\omega}_{\mathcal{F}}$ itself. Then, Eq. (F.1.3) can be bounded by

$$\mathbb{E}\left[\left|f(\mu_{t+1}^N) - f\left(\tilde{T}\left(\mu_t^N, \zeta_t^{\pi,\mu^N}\right)\right)\right|\right] \leq \mathbb{E}\left[\tilde{\omega}_{\mathcal{F}}\left(d_\Sigma\left(\mu_{t+1}^N, \tilde{T}\left(\mu_t^N, \zeta_t^{\pi,\mu^N}\right)\right)\right)\right]$$

$$\leq \tilde{\omega}_{\mathcal{F}}\left(\mathbb{E}\left[d_{\Sigma}\left(\mu_{t+1}^{N}, \tilde{T}\left(\mu_{t}^{N}, \zeta_{t}^{\pi,\mu^{N}}\right)\right)\right]\right)$$

irrespective of both π and f by the concavity of $\tilde{\omega}_{\mathcal{F}}$ and Jensen's inequality. For notational convenience, we define $x_t^N \coloneqq (x_t^{i,N})_{i\in[N]}$, and arrive at

$$\begin{aligned}\mathbb{E}\left[d_{\Sigma}\left(\mu_{t+1}^{N}, \tilde{T}\left(\mu_{t}^{N}, \zeta_{t}^{\pi,\mu^{N}}\right)\right)\right] &= \sum_{m=1}^{\infty} 2^{-m}\mathbb{E}\left[\left|\int f_m \,\mathrm{d}\left(\mu_{t+1}^{N} - \tilde{T}\left(\mu_{t}^{N}, \zeta_{t}^{\pi,\mu^{N}}\right)\right)\right|\right] \\ &\leq \sup_{m\geq 1}\mathbb{E}\left[\mathbb{E}\left[\left|\int f_m \,\mathrm{d}\left(\mu_{t+1}^{N} - \tilde{T}\left(\mu_{t}^{N}, \zeta_{t}^{\pi,\mu^{N}}\right)\right)\right| \,\middle|\, x_t^N\right]\right].\end{aligned}$$

Finally, we require the aforementioned weak LLN argument which goes as follows

$$\begin{aligned}&\mathbb{E}\left[\left|\int f_m \,\mathrm{d}\left(\mu_{t+1}^{N} - \tilde{T}\left(\mu_{t}^{N}, \zeta_{t}^{\pi,\mu^{N}}\right)\right)\right| \,\middle|\, x_t^N\right]^2 \\ &\quad= \mathbb{E}\left[\left|\frac{1}{N}\sum_{i\in[N]}\left(f_m(x_{t+1}^i) - \mathbb{E}\left[f_m(x_{t+1}^i) \mid x_t^N\right]\right)\right| \,\middle|\, x_t^N\right]^2 \\ &\quad\leq \mathbb{E}\left[\left|\frac{1}{N}\sum_{i\in[N]}\left(f_m(x_{t+1}^i) - \mathbb{E}\left[f_m(x_{t+1}^i) \mid x_t^N\right]\right)\right|^2 \,\middle|\, x_t^N\right] \\ &\quad= \frac{1}{N^2}\sum_{i\in[N]}\mathbb{E}\left[\left(f_m(x_{t+1}^i) - \mathbb{E}\left[f_m(x_{t+1}^i) \mid x_t^N\right]\right)^2 \,\middle|\, x_t^N\right] \leq \frac{4}{N} \to 0.\end{aligned}$$

Here, we have used that $|f_m| \leq 1$, as well as the conditional independence of x_{t+1}^i given x_t^N. In combination with the above results, the term (F.1.3) thus converges to zero. Moving on to the remaining second term (F.1.4), we note that the induction assumption implies that

$$\begin{aligned}&\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}}\mathbb{E}\left[\left|f\left(\tilde{T}\left(\mu_{t}^{N}, \zeta_{t}^{\pi,\mu^{N}}\right)\right) - f(\mu_{t+1})\right|\right] \\ &\quad= \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}}\mathbb{E}\left[\left|f\left(\tilde{T}\left(\mu_{t}^{N}, \zeta_{t}^{\pi,\mu^{N}}\right)\right) - f\left(\tilde{T}\left(\mu_{t}, \zeta_{t}^{\pi,\mu}\right)\right)\right|\right] \\ &\quad\leq \sup_{\pi\in\Pi}\sup_{g\in\mathcal{G}}\mathbb{E}\left[\left|g(\mu_t^N) - g(\mu_t)\right|\right] \to 0\end{aligned}$$

using the function $g \coloneqq f \circ \tilde{T}_*^{\pi_t}$ which belongs to the class $\mathcal{G}$ of equicontinuous functions with modulus of continuity $\omega_{\mathcal{G}} \coloneqq \omega_{\mathcal{F}} \circ \omega_{\tilde{T}}$. Here, $\omega_{\tilde{T}}$ is the uniform modulus of continuity over all policies π of $\mu_t \mapsto \tilde{T}_*^{\pi_t}(\mu_t) \coloneqq \tilde{T}(\mu_t, \zeta_t^{\pi,\mu})$. The equicontinuity of $\{\tilde{T}_*^{\pi_t}\}_{\pi\in\Pi}$ is a consequence of Proposition F.1.1 as well as the equicontinuity of functions $\mu_t \mapsto \zeta_t^{\pi,\mu}$ which in turn follows from the uniform Lipschitzness of Π. The validation of this claim is provided in the next lines. Note that this also completes the induction and thereby the proof. For a sequence of $\mu_n \to \mu \in \mathcal{P}(\mathcal{X})$ we can write

$$\begin{aligned}&\sup_{\pi\in\Pi} W_1(\zeta_t^{\pi,\mu_n}, \zeta_t^{\pi,\mu}) \\ &\quad\leq \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\left|\iiint f'(x,u)\pi_t(\mathrm{d}u \mid y)(p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu_n) - p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu))\mu_n(\mathrm{d}x)\right| \\ &\qquad+ \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\left|\iiint f'(x,u)\pi_t(\mathrm{d}u \mid y)p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu)(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x))\right|.\end{aligned}$$

Starting with the first term, we apply Assumptions 4.3.1a to 4.3.1b to arrive at

$$\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le 1}\left|\iiint f'(x,u)\pi_t(\mathrm{d}u\mid y)(p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu_n)-p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu))\mu_n(\mathrm{d}x)\right|\\
&\quad\le \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le 1}\int\left|\iint f'(x,u)\pi_t(\mathrm{d}u\mid y)(p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu_n)-p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu))\right|\mu_n(\mathrm{d}x)\\
&\quad\le \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le 1}\sup_{x\in\mathcal{X}}\left|\iint f'(x,u)\pi_t(\mathrm{d}u\mid y)(p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu_n)-p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu))\right|\\
&\quad\le L_\Pi\sup_{x\in\mathcal{X}}W_1(p_{\mathcal{Y}}(\cdot\mid x,\mu_n),p_{\mathcal{Y}}(\cdot\mid x,\mu))\\
&\quad\le L_\Pi L_{p_{\mathcal{Y}}}W_1(\mu_n,\mu)\to 0
\end{aligned}$$

with Lipschitz constant L_Π corresponding to the Lipschitz function $y\mapsto\int f'(x,u)\pi_t(\mathrm{d}u\mid y)$. Alternatively, if $p_{\mathcal{Y}}$ is assumed independent of the MF in Assumption 4.3.1b, the term is zero.

In a similar fashion, we point out the 1-Lipschitzness of $x\mapsto\iint\frac{f'(x,u)}{L_\Pi L_{p_{\mathcal{Y}}}+1}\pi_t(\mathrm{d}u\mid y)p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu)$, as

$$\begin{aligned}
&\left|\iint\frac{f'(z,u)}{L_\Pi L_{p_{\mathcal{Y}}}+1}\pi_t(\mathrm{d}u\mid y)p_{\mathcal{Y}}(\mathrm{d}y\mid z,\mu)-\iint\frac{f'(x,u)}{L_\Pi L_{p_{\mathcal{Y}}}+1}\pi_t(\mathrm{d}u\mid y)p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu)\right|\\
&\quad\le\left|\iint\frac{f'(z,u)-f'(x,u)}{L_\Pi L_{p_{\mathcal{Y}}}+1}\pi_t(\mathrm{d}u\mid y)p_{\mathcal{Y}}(\mathrm{d}y\mid z,\mu)\right|\\
&\qquad+\left|\iint\frac{f'(x,u)}{L_\Pi L_{p_{\mathcal{Y}}}+1}\pi_t(\mathrm{d}u\mid y)(p_{\mathcal{Y}}(\mathrm{d}y\mid z,\mu)-p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu))\right|\\
&\quad\le\frac{1}{L_\Pi L_{p_{\mathcal{Y}}}+1}d(z,x)+\frac{L_\Pi}{L_\Pi L_{p_{\mathcal{Y}}}+1}W_1(p_{\mathcal{Y}}(\mathrm{d}y\mid z,\mu),p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu))\\
&\quad\le\left(\frac{1}{L_\Pi L_{p_{\mathcal{Y}}}+1}+\frac{L_\Pi L_{p_{\mathcal{Y}}}}{L_\Pi L_{p_{\mathcal{Y}}}+1}\right)d(x,y)=d(x,y)
\end{aligned}$$

for $z\neq x$. Alternatively, if the state space is assumed finite in Assumption 4.3.1b, the Lipschitzness follows directly.

This eventually yields the convergence of the second term, i.e.

$$\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le 1}\left|\iiint f'(x,u)\pi_t(\mathrm{d}u\mid y)p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu)(\mu_n(\mathrm{d}x)-\mu(\mathrm{d}x))\right|\\
&\quad=\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le 1}(L_{p_{\mathcal{Y}}}L_\Pi+1)\left|\iint\frac{f'(x,u)}{L_{p_{\mathcal{Y}}}L_\Pi+1}\pi_t(\mathrm{d}u\mid y)p_{\mathcal{Y}}(\mathrm{d}y\mid x,\mu)(\mu_n(\mathrm{d}x)-\mu(\mathrm{d}x))\right|\\
&\quad\le(L_{p_{\mathcal{Y}}}L_\Pi+1)W_1(\mu_n,\mu)\to 0
\end{aligned}$$

and thus completes the proof. □

In the special case of finite states and actions, the approximation rate can also be quantified to $\mathcal{O}(1/\sqrt{N})$ by considering equi-Lipschitz families of functions $\mathcal{F}$ with constant L_f. Then, there is no need to consider the two different metrizations d_Σ and W_1, as they are Lipschitz equivalent, and one can simply use the L_1 distance. The convergence in the first term (F.1.3) is then directly via the weak LLN at rate $\mathcal{O}(1/\sqrt{N})$ by

$$\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}}\mathbb{E}\left[\left|f(\mu_{t+1}^N)-f\left(\tilde{T}\left(\mu_t^N,\zeta_t^{\pi,\mu^N}\right)\right)\right|\right]$$

$$
\begin{aligned}
&\leq \sup_{\pi\in\Pi} L_f \mathbb{E}\left[\sum_{x\in\mathcal{X}} \left|\mu_{t+1}^N(x) - \tilde{T}\left(\mu_t^N, \zeta_t^{\pi,\mu^N}\right)(x)\right|\right] \\
&= \sup_{\pi\in\Pi} L_f \sum_{x\in\mathcal{X}} \mathbb{E}\left[\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N \mathbf{1}_x(x_{t+1}^{i,N}) - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \mathbf{1}_x(x_{t+1}^{i,N}) \,\middle|\, x_t^N\right]\right| \,\middle|\, x_t^N\right]\right] \\
&\leq L_f|\mathcal{X}|\sqrt{\frac{4}{N}},
\end{aligned}
$$

while for the second term (F.1.4) we use the induction assumption, since $\tilde{T}$ is uniformly Lipschitz.

F.2 AGENTS WITH MEMORY AND HISTORY-DEPENDENCE

For agents with bounded memory, we note that such memory can be analyzed by our model by adding the memory state to the usual agent state, and manipulations on the memory either to the actions or transition dynamics.

For example, let $z_t^i \in \mathcal{Q} \coloneqq \{0,1\}^Q$ be the Q-bit memory of an agent at any time. Then, we may consider the new $\mathcal{X}\times\mathcal{Q}$-valued state (x_t^i, z_t^i), which remains compact, and the new $\mathcal{U}\times\mathcal{Q}$-valued actions (u_t^i, w_t^i), where w_t^i is a write action that can arbitrarily rewrite the memory, $z_{t+1}^i = w_t^i$. Theoretical properties are preserved by discreteness of added states and actions.

Analogously, extending transition dynamics to include observations y also allows for description of history-dependent policies. This approach extends to infinite-memory states, by adding observations y also to the transition dynamics, and considering histories for states and observations. Define the observation space of histories $\mathcal{Y}' \coloneqq \bigcup_{i=0}^\infty \mathcal{Y}\times(\mathcal{Y}\times\mathcal{U})^i$, and the according state space $\mathcal{X}' \coloneqq \bigcup_{i=0}^\infty \mathcal{X}\times(\mathcal{Y}\times\mathcal{U})^i$. The new MF μ_t^N, μ_t are thus $\mathcal{P}(\mathcal{X}')$-valued. The new observation-dependent dynamics are then defined by

$$
P'(\cdot \mid x, y, u, \mu) = p(\cdot \mid x_1, u, \operatorname{marg}_1 \mu) \otimes \delta_{(\mathbf{x}_2, y, u)}
$$

where marg_1 maps μ to its first marginal, x_1 is the first component of x, and $\mathbf{x}_2$ is the $(\mathcal{Y}\times\mathcal{U})^t$-valued past history. Here, $(\mathbf{x}_2, y, u)$ defines the new history of an agent, which is observed by

$$
P^{y\prime}(\cdot \mid x, \mu) = p_{\mathcal{Y}}(\cdot \mid x_1, \operatorname{marg}_1 \mu) \otimes \delta_{\mathbf{x}_2}.
$$

Clearly, Lipschitz continuity is preserved. Further, we obtain the MF transition operator

$$
T'(\mu_t, h_t') \coloneqq \iiint P'(\cdot \mid x, y, u, \mu_t) h_t'(\mathrm{d}x, \mathrm{d}y, \mathrm{d}u).
$$

using $\mathcal{X}'\times\mathcal{Y}'\times\mathcal{U}$-valued actions $h_t' = \mu_t \otimes p_{\mathcal{Y}}(\mu_t) \otimes \check{\pi}[h_t']$ for some Lipschitz $\check{\pi}[h_t']\colon \mathcal{Y}' \to \mathcal{P}(\mathcal{U})$. And in particular, the proof of e.g. Theorem 4.3.1 extends to this new case. For example, the weak LLN argument still holds by

$$
\begin{aligned}
&\mathbb{E}\left[d_\Sigma\left(\mu_{t+1}^N, T'\left(\mu_t^N, h_t'\right)\right)\right] \\
&\quad\leq \sup_{m\geq 1} \mathbb{E}\left[\mathbb{E}\left[\left|\int f_m \,\mathrm{d}\left(\mu_{t+1}^N - T'\left(\mu_t^N, h_t'\right)\right)\right| \,\middle|\, x_t^N\right]\right] \\
&\quad\leq \sup_{m\geq 1} \mathbb{E}\left[\left|\frac{1}{N}\sum_{i\in[N]} \left(f_m(x_{t+1}^i, y_0^i, u_0^i, \ldots, y_t^i, u_t^i)\right.\right.\right.
\end{aligned}
$$

$$-\mathbb{E}\left[f_m(x_{t+1}^i, y_0^i, u_0^i, \ldots, y_t^i, u_t^i) \mid x_t^N\right])\Big|^2 \Big| \, x_t^N\Big]^{\frac{1}{2}} \leq \frac{4}{N} \to 0.$$

for appropriate sequences of functions $(f_m)_{m\geq 1}$, $f_m \colon \mathcal{X} \times (\mathcal{Y} \times \mathcal{U})^{t+1} \to [-1, 1]$ [216] and

$$\int f_m \, \mathrm{d}T'\left(\mu_t^N, h_t'\right) = \int f_m(x_{t+1}, y_0, u_0, \ldots, y_t, u_t) P'(\mathrm{d}x_{t+1} \mid x_t, y_t, u_t, \mu_t^N)$$
$$\check{\pi}[h_t'](\mathrm{d}u_t \mid y_t) P^{y\prime}(\mathrm{d}y_t \mid x_t, \mu_t^N) \mu_t^N(\mathrm{d}x_t, \mathrm{d}y_0, \mathrm{d}u_0, \ldots, \mathrm{d}y_{t-1}, \mathrm{d}u_{t-1}).$$

Analogously, we can see that the above is part of a set of equicontinuous functions, and again allows application of the induction assumption, completing the extension.

F.3 PROOF OF COROLLARY 4.3.1

Proof of Corollary 4.3.1. The finite-agent discounted objective converges uniformly over policies to the MFC objective

$$\sup_{\pi \in \Pi} \left|J^N(\pi) - J(\pi)\right| \to 0 \quad \text{as } N \to \infty, \tag{F.3.5}$$

since for any $\varepsilon > 0$, let $T \in \mathcal{T}$ such that $\sum_{t=T}^{\infty} \gamma^t \mathbb{E}\left|\left[r(\mu_t^N) - r(\mu_t)\right]\right| \leq \frac{\gamma^T}{1-\gamma} \max_\mu 2|r(\mu)| < \frac{\varepsilon}{2}$, and further let $\sum_{t=0}^{T-1} \gamma^t \mathbb{E}\left|\left[r(\mu_t^N) - r(\mu_t)\right]\right| < \frac{\varepsilon}{2}$ by Theorem 4.3.1 for sufficiently large N.

Therefore, approximate optimality is obtained by

$$\begin{aligned} J^N(\pi) - \sup_{\pi' \in \Pi} J^N(\pi') &= \inf_{\pi' \in \Pi} (J^N(\pi) - J^N(\pi')) \\ &\geq \inf_{\pi' \in \Pi} (J^N(\pi) - J(\pi)) + \inf_{\pi' \in \Pi} (J(\pi) - J(\pi')) + \inf_{\pi' \in \Pi} (J(\pi') - J^N(\pi')) \\ &\geq -\frac{\varepsilon}{2} + 0 - \frac{\varepsilon}{2} = -\varepsilon \end{aligned}$$

by the optimality of $\pi \in \arg\max_{\pi' \in \Pi} J(\pi')$ and Eq. (F.3.5) for sufficiently large N. □

F.4 PROOF OF PROPOSITION 4.3.1

Proof of Proposition 4.3.1. We begin by showing the first statement. The proof is by showing $\bar{\mu}_t = \mu_t$ at all times $t \in \mathcal{T}$, as it then follows that $\bar{J}(\bar{\pi}) = \sum_{t=0}^{\infty} \gamma^t r(\bar{\mu}_t) = \sum_{t=0}^{\infty} \gamma^t r(\mu_t) = J(\pi)$. At time $t = 0$, we have by definition $\bar{\mu}_0 = \mu_0$. Assume $\bar{\mu}_t = \mu_t$ at time t, then at time $t + 1$, by Eq. (4.3.34) and Eq. (4.3.35), we have

$$\bar{\mu}_{t+1} = \iiint p(x, u, \mu_t) \bar{\pi}_t(\mathrm{d}u \mid y, \bar{\mu}_t) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \bar{\mu}_t) \bar{\mu}_t(\mathrm{d}x) \tag{F.4.6}$$

$$= \iiint p(x, u, \mu_t) \pi_t(\mathrm{d}u \mid y) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu_t) \mu_t(\mathrm{d}x) = \mu_{t+1} \tag{F.4.7}$$

which is the desired statement. An analogous proof for the second statement completes the proof. □

F.5 PROOF OF COROLLARY 4.3.2

Proof of Corollary 4.3.2. Assume $J(\Phi(\bar{\pi})) < \sup_{\pi' \in \Pi} J(\pi')$. Then there exists $\pi' \in \Pi$ such that $J(\Phi(\bar{\pi})) < J(\pi')$. But by Proposition 4.3.1, there exists $\bar{\pi}' \in \bar{\Pi}$ such that $\bar{J}(\bar{\pi}') = J(\pi')$ and hence $\bar{J}(\bar{\pi}) = J(\Phi(\bar{\pi})) < J(\pi') = \bar{J}(\bar{\pi}')$, which contradicts $\bar{\pi} \in \arg\max_{\bar{\pi}' \in \bar{\Pi}} \bar{J}(\bar{\pi}')$. Therefore, $\Phi(\bar{\pi}) \in \arg\max_{\pi' \in \Pi} J(\pi')$. □

F.6 PROOF OF THEOREM 4.3.2

Proof of Theorem 4.3.2. We verify the assumptions in [390]. First, note the (weak) continuity of transition dynamics $\hat{T}$.

Proposition F.6.1. *Under Assumption 4.3.1a, $\hat{T}(\mu_n, h_n) \to \hat{T}(\mu, h)$ for any sequence $(\mu_n, h_n) \to (\mu, h)$ of MFs $\mu_n, \mu \in \mathcal{P}(\mathcal{X})$ and joint distributions $h_n \in \mathcal{H}(\mu_n)$, $h \in \mathcal{H}(\mu)$.*

Proof. The convergence $h_n \to h$ also implies the convergence of its marginal $\int_{\mathcal{Y}} h_n(\cdot, \mathrm{d}y, \cdot) \to \int_{\mathcal{Y}} h(\cdot, \mathrm{d}y, \cdot)$. The proposition then follows immediately from Proposition F.1.1. ■

Furthermore, the reward is continuous and hence bounded by Assumption 4.3.1a. It is inf-compact by

$$\{h \in \mathcal{H}(\mu) \mid -r(\mu) \le c\} = \begin{cases} \mathcal{H}(\mu) & \text{if } -r(\mu) \le c, \\ \emptyset & \text{else,} \end{cases}$$

where $\mathcal{H}(\mu)$ is closed by Appendix E.4, and Lemma F.8.1 if considering equi-Lipschitz policies in Assumption 4.3.1b.

Further, by compactness of $\mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{U})$, $\mathcal{H}(\mu)$ is compact as a closed subset of a compact set.

Lastly, lower semicontinuity of $\mu \mapsto \mathcal{H}(\mu)$ is given, since for any $\mu_n \to \mu$ and $h = \mu \otimes p_{\mathcal{Y}}(\mu) \otimes \check{\pi} \in \mathcal{H}(\mu)$, we can find $h_n \in \mathcal{H}(\mu_n)$: Let $h_n = \mu_n \otimes p_{\mathcal{Y}}(\mu_n) \otimes \check{\pi}$, then

$$\begin{aligned} W_1(h_n, h) &= \sup_{f \in \mathrm{Lip}(1)} \iiint f(x, y, u) \check{\pi}(\mathrm{d}u \mid y) \left(p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu_n)\mu_n(\mathrm{d}x) - p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu)\mu(\mathrm{d}x)\right) \\ &\le \sup_{f \in \mathrm{Lip}(1)} \iiint f(x, y, u) \check{\pi}(\mathrm{d}u \mid y) \left(p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu_n) - p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu)\right) \mu_n(\mathrm{d}x) \\ &\quad + \sup_{f \in \mathrm{Lip}(1)} \iiint f(x, y, u) \check{\pi}(\mathrm{d}u \mid y) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu) \left(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x)\right) \\ &\le \sup_{f \in \mathrm{Lip}(1)} \int \left| \iint f(x, y, u) \check{\pi}(\mathrm{d}u \mid y) \left(p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu_n) - p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu)\right) \right| \mu_n(\mathrm{d}x) \\ &\quad + \sup_{f \in \mathrm{Lip}(1)} \iiint f(x, y, u) \check{\pi}(\mathrm{d}u \mid y) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \mu) \left(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x)\right) \to 0 \end{aligned}$$

since the integrands are Lipschitz by Assumption 4.3.1a and analyzed as in the proof of Theorem 4.3.1.

The proof concludes by [390, Theorem 4.2]. □

F.7 CONVERGENCE LEMMA

Lemma F.7.1. *Assume that (X, d) is a complete metric space and that $(x_n)_{n\in\mathbb{N}}$ is a sequence of elements of X. Then, the convergence condition of the sequence $(x_n)_{n\in\mathbb{N}}$, i.e. that*

$$\exists x \in X : \forall \varepsilon > 0 : \exists N \in \mathbb{N} : \forall n \geq N : d(x, x_n) < \varepsilon \tag{F.7.8}$$

holds, is equivalent to the statement

$$\forall \varepsilon > 0 : \exists x \in X : \exists N \in \mathbb{N} : \forall n \geq N : d(x, x_n) < \varepsilon. \tag{F.7.9}$$

Proof. (F.7.8) $\Rightarrow$ (F.7.9): follows immediately.

(F.7.9) $\Rightarrow$ (F.7.8): Choose some strictly monotonically decreasing, positive sequence of $(\varepsilon_i)_{i\in\mathbb{N}}$ with $\lim_{i\to\infty} \varepsilon_i = 0$. Then, by statement (F.7.9) we can define corresponding sequences $(x_i)_{i\in\mathbb{N}}$ and $(N_i)_{i\in\mathbb{N}}$ such that

$$\forall n \geq N_i : d(x_i, x_n) < \varepsilon_i. \tag{F.7.10}$$

Consider $i, i' \in \mathbb{N}$ and assume w.l.o.g. $i < i'$. We know by the triangle inequality

$$\forall n \geq \max\{N_i, N_{i'}\} : d(x_i, x_{i'}) \leq d(x_i, x_n) + d(x_n, x_{i'}) \leq 2\varepsilon_i. \tag{F.7.11}$$

Thus, the sequence $(x_i)_{i\in\mathbb{N}}$ is Cauchy and therefore converges to some $x \in X$ because (X, d) is a complete metric space by assumption. Specifically, this is equivalent to

$$\exists x \in X : \forall \varepsilon > 0 : \exists I \in \mathbb{N} : \forall i \geq I : d(x, x_i) < \varepsilon. \tag{F.7.12}$$

Finally, statements (F.7.11), (F.7.12), and the triangle inequality yield

$$\exists x \in X : \forall 2\varepsilon > 0 : \exists N \in \mathbb{N} : \forall n \geq N : d(x, x_n) \leq d(x, x_i) + d(x_i, x_n) < 2\varepsilon$$

which implies the desired statement (F.7.8) and concludes the proof. □

F.8 CLOSEDNESS OF JOINT MEASURES UNDER EQUI-LIPSCHITZ KERNELS

Lemma F.8.1. *Let $\mu_{xy} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be arbitrary. For any $h_n = \mu_{xy} \otimes \breve{\pi}_n \to h \in \mathcal{P}(\mathcal{X} \times \mathcal{Y} \times \mathcal{U})$ with L_Π-Lipschitz $\breve{\pi}_n \in \mathcal{P}(\mathcal{U})^{\mathcal{Y}}$, there exists L_Π-Lipschitz $\breve{\pi} \in \mathcal{P}(\mathcal{U})^{\mathcal{Y}}$ such that $h = \mu_{xy} \otimes \breve{\pi}$.*

Proof. For readability, we write $\mu_y \in \mathcal{P}(\mathcal{Y})$ for the second marginal of μ_{xy}. The required $\breve{\pi}$ is constructed as the μ_y-a.e. pointwise limit of $y \mapsto \breve{\pi}_n(y) \in \mathcal{P}(\mathcal{U})$, as $\mathcal{P}(\mathcal{U})$ is sequentially compact under the topology of weak convergence by Prokhorov's theorem [215]. For the proof, we assume Hilbert $\mathcal{Y}$ and finite actions $\mathcal{U}$, making $\mathcal{P}(\mathcal{U})$ Euclidean.

First, **(i)** we show that $\breve{\pi}_n(y)$ must converge for μ_y-a.e. $y \in \mathcal{Y}$ to some arbitrary limit, which we define as $\breve{\pi}(y)$. It then follows by Egorov's theorem (e.g. [253, Lemma 1.38]) that for any $\epsilon > 0$, there exists a measurable set $A \in \mathcal{Y}$ such that $\mu_y(A) < \epsilon$ and $\breve{\pi}_n(y)$ converges uniformly on $\mathcal{Y} \setminus A$. Therefore, we obtain that $\breve{\pi}$ restricted to $\mathcal{Y} \setminus A$ is L_Π-Lipschitz as a uniform limit of L_Π-Lipschitz functions, hence μ_y-a.e. L_Π-Lipschitz. **(ii)** We then extend $\breve{\pi}$ on the entire space $\mathcal{Y}$ to be L_Π-Lipschitz. **(iii)** All that remains is to show that indeed, the extended $\breve{\pi}$ fulfills $h = \mu_y \otimes \breve{\pi}$, which is the desired closedness.

(I) ALMOST-EVERYWHERE CONVERGENCE. To prove the μ_y-a.e. convergence, we perform a proof by contradiction and assume the statement is not true. Then there exists a measurable set $A \subseteq \mathcal{Y}$ with positive measure $\mu_y(A) > 0$ such that for all $y \in A$ the sequence $\breve{\pi}_n(y) \in \mathcal{P}(\mathcal{U})$ does not converge as $n \to \infty$. We show that then, $\mu_y \otimes \breve{\pi}_n$ does not converge to any limiting $\tilde{h} \in \mathcal{P}(\mathcal{Y} \times \mathcal{U})$, which is a contradiction with the premise and completes the proof.

Lemma F.8.2. *There exists $y^* \in A$ such that for any $r > 0$, the set $B_r(y^*) \cap A$ has positive measure.*

Proof of Lemma F.8.2. Consider an open cover $\bigcup_{y \in A} B_r(y)$ of A using balls B_r with radius r, and choose a finite subcover $\{B_r(y_i)\}_{i=1,\dots,K}$ of A by compactness of $\mathcal{Y}$. Then, there exists a ball $B_r(y^*)$ from the finite subcover around a point $y^* \in \mathcal{Y}$ such that $\mu_y(B_r(y^*) \cap A) > 0$, as otherwise $\mu_y(A) = \mu_y(\bigcup_{i=1}^K B_r(y_i) \cap A) \leq \sum_{i=1}^K \mu_y(B_r(y_i) \cap A) = 0$ contradicts $\mu_y(A) > 0$.

By repeating the argument, there must exist $y^* \in A$ for which we have for any $r > 0$ that the ball $B_r(y^*) \cap A$ has positive measure. More precisely, consider a sequence of radii $r_k = 1/k$, $k \geq 1$, and repeatedly choose balls $B_{r_{k+1}} \subseteq B_{r_k}$ from an open cover of $B_{r_k} \cap A$ such that $\mu(B_{r_{k+1}} \cap B_{r_k} \cap A) > 0$, starting with $B_{r_1} \subseteq \mathcal{Y}$ such that $\mu(B_{r_1} \cap A) > 0$. By induction, we thus have for any k that $\mu(B_{r_k} \cap A) > 0$. The sequence $(B_{r_k})_{k \in \mathbb{N}}$ produces a decreasing sequence of compact sets by taking the closure of the balls $\bar{B}_{r_k}$. By Cantor's intersection theorem [391, Theorem 2.36], the intersection is non-empty, $\bigcap_{k \in \mathbb{N}} \bar{B}_{r_k} \neq \emptyset$. Choose arbitrary $y^* \in \bigcap_{k \in \mathbb{N}} \bar{B}_{r_k}$, then for any $r > 0$ we have that $B_{r_k} \subseteq B_r(y^*)$ for some k by $r_k \to 0$. Therefore, $\mu(B_r(y^*) \cap A) \geq \mu(B_{r_k} \cap A) > 0$. ∎

BOUNDING DIFFERENCE TO ASSUMED LIMIT FROM BELOW. Choose y^* according to Lemma F.8.2. By Eq. (F.7.9) in Lemma F.7.1, since $\breve{\pi}_n(y^*) \in \mathcal{P}(\mathcal{U})$ does not converge, there exists $\epsilon > 0$ such that for all $r > 0$, infinitely often (i.o.) in n,

$$W_1\left(\breve{\pi}_n(\cdot \mid y^*), \frac{1}{\mu_y(B_r(y^*))} \int_{B_r(y^*)} \tilde{\pi}(\cdot \mid y) \mu_y(\mathrm{d}y)\right)$$
$$= \frac{1}{2} \sum_{u \in \mathcal{U}} \left| \breve{\pi}_n(u \mid y^*) - \frac{1}{\mu_y(B_r(y^*))} \int_{B_r(y^*)} \tilde{\pi}(u \mid y) \mu_y(\mathrm{d}y) \right| > \epsilon$$

where for finite $\mathcal{U}$, W_1 is equivalent to the total variation norm [392, Theorem 4], which is half the L_1 norm, and $\tilde{\pi}$ is not necessarily Lipschitz and results from disintegration of h into $h = \mu_y \otimes \tilde{\pi}$ [253].

Now fix arbitrary $\epsilon' \in (\frac{\epsilon}{2}, \epsilon)$. Then, by the prequel, we define the non-empty set $\bar{\mathcal{U}}(r) \subseteq \mathcal{U}$ by excluding all actions where the absolute value is less than $\frac{\epsilon - \epsilon'}{|\mathcal{U}|}$, i.e.

$$\bar{\mathcal{U}}(r) \coloneqq \left\{ u \in \mathcal{U} \;\middle|\; \left| \breve{\pi}_n(u \mid y^*) - \frac{1}{\mu_y(B_r(y^*))} \int_{B_r(y^*)} \tilde{\pi}(u \mid y) \mu_y(\mathrm{d}y) \right| \geq \frac{\epsilon - \epsilon'}{|\mathcal{U}|} \right\},$$

such that

$$\frac{1}{2} \sum_{u \in \bar{\mathcal{U}}(r)} \left| \breve{\pi}_n(u \mid y^*) - \frac{1}{\mu_y(B_r(y^*))} \int_{B_r(y^*)} \tilde{\pi}(u \mid y) \mu_y(\mathrm{d}y) \right| > \epsilon' \tag{F.8.13}$$

since we have the bound on the value contributed by excluded actions $u \notin \bar{\mathcal{U}}(r)$

$$\frac{1}{2}\sum_{u\notin\bar{\mathcal{U}}(r)}\left|\check{\pi}_n(u \mid y^*) - \frac{1}{\mu_y(B_r(y^*))}\int_{B_r(y^*)}\tilde{\pi}(u \mid y)\mu_y(\mathrm{d}y)\right| \leq \frac{\epsilon-\epsilon'}{2} < \epsilon - \epsilon'. \tag{F.8.14}$$

By L_Π-Lipschitz $\check{\pi}_n$, we also have for all $y \in B_r(y^*)$ that $W_1(\check{\pi}_n(y), \check{\pi}_n(y^*)) < L_\Pi r$. Hence, in particular if we choose $r = \frac{1}{L_\Pi}\min\left(\epsilon' - \frac{\epsilon}{2}, \frac{\epsilon'}{2}, \frac{\epsilon-\epsilon'}{4|\mathcal{U}|}\right)$, then for all $y \in B_r(y^*)$

$$\frac{1}{2}\sum_{u\in\mathcal{U}}|\check{\pi}_n(u \mid y) - \check{\pi}_n(u \mid y^*)| < \min\left(\epsilon' - \frac{\epsilon}{2}, \frac{\epsilon'}{2}, \frac{\epsilon-\epsilon'}{4|\mathcal{U}|}\right) \tag{F.8.15}$$

and in particular also

$$|\check{\pi}_n(u \mid y) - \check{\pi}_n(u \mid y^*)| < \frac{\epsilon-\epsilon'}{2|\mathcal{U}|}$$

for all actions $u \in \bar{\mathcal{U}}(r)$, such that by definition of $\bar{\mathcal{U}}(r)$, we find that the sign of the value inside the absolute value must not change on the entirety of $y \in B_r(y^*)$, i.e.

$$\begin{aligned}&\operatorname{sgn}\left(\check{\pi}_n(u \mid y) - \frac{1}{\mu_y(B_r(y^*))}\int_{B_r(y^*)}\tilde{\pi}(u \mid y')\mu_y(\mathrm{d}y')\right)\\ &\quad= \operatorname{sgn}\left(\check{\pi}_n(u \mid y^*) - \frac{1}{\mu_y(B_r(y^*))}\int_{B_r(y^*)}\tilde{\pi}(u \mid y')\mu_y(\mathrm{d}y')\right)\end{aligned}$$

which implies, since the signs must match for all y with the term for y^*, by integrating over y

$$\begin{aligned}&\operatorname{sgn}\left(\int_{B_r(y^*)}\check{\pi}_n(u \mid y')\mu_y(\mathrm{d}y') - \int_{B_r(y^*)}\tilde{\pi}(u \mid y')\mu_y(\mathrm{d}y')\right)\\ &\quad= \operatorname{sgn}\left(\int_{B_r(y^*)}\left(\check{\pi}_n(u \mid y^*) - \tilde{\pi}(u \mid y')\right)\mu_y(\mathrm{d}y')\right).\end{aligned} \tag{F.8.16}$$

From the triangle inequality,

$$\begin{aligned}&\frac{1}{2}\sum_{u\in\bar{\mathcal{U}}(r)}\left|\int_{B_r(y^*)}\left(\check{\pi}_n(u \mid y^*) - \tilde{\pi}(u \mid y)\right)\mu_y(\mathrm{d}y)\right|\\ &\quad\leq \frac{1}{2}\sum_{u\in\bar{\mathcal{U}}(r)}\left|\int_{B_r(y^*)}\left(\check{\pi}_n(u \mid y^*) - \check{\pi}_n(u \mid y)\right)\mu_y(\mathrm{d}y)\right|\\ &\qquad+ \frac{1}{2}\sum_{u\in\bar{\mathcal{U}}(r)}\left|\int_{B_r(y^*)}\left(\check{\pi}_n(u \mid y) - \tilde{\pi}(u \mid y)\right)\mu_y(\mathrm{d}y)\right|,\end{aligned}$$

it follows then that for all $y \in B_r(y^*)$ by Eq. (F.8.13) and Eq. (F.8.15), i.o. in n

$$\frac{1}{2}\sum_{u\in\bar{\mathcal{U}}(r)}\left|\int_{B_r(y^*)}\left(\check{\pi}_n(u \mid y) - \tilde{\pi}(u \mid y)\right)\mu_y(\mathrm{d}y)\right|$$

$$
\begin{aligned}
&\geq \frac{1}{2} \sum_{u \in \bar{\mathcal{U}}(r)} \left| \int_{B_r(y^*)} (\check{\pi}_n(u \mid y^*) - \tilde{\pi}(u \mid y))\, \mu_y(\mathrm{d}y) \right| \\
&\quad - \frac{1}{2} \sum_{u \in \bar{\mathcal{U}}(r)} \left| \int_{B_r(y^*)} (\check{\pi}_n(u \mid y^*) - \check{\pi}_n(u \mid y))\, \mu_y(\mathrm{d}y) \right| \\
&> \epsilon' - \frac{\epsilon'}{2} = \frac{\epsilon'}{2}.
\end{aligned} \tag{F.8.17}
$$

PASS TO LIMIT OF LIPSCHITZ FUNCTIONS. Now consider the sequence of m-Lipschitz functions $f_m \colon \mathcal{Y} \times \mathcal{U} \to [0,1]$,

$$
\begin{aligned}
f_m(y,u) = \min\left\{1, \left(1 - (m d(y,y^*) - mr + 1)^+\right)^+\right\} \\
\cdot \operatorname{sgn}\left(\int_{B_r(y^*)} \left(\check{\pi}_n(u \mid y^*) - \tilde{\pi}(u \mid y')\right) \mu_y(\mathrm{d}y') \right),
\end{aligned}
$$

where $(\cdot)^+ = \max(\cdot, 0)$ and sgn is the sign function. Note that $f_m = 0$ for all $y \notin B_r(y^*)$. Further, as $m \to \infty$,

$$
f_m(y,u) \uparrow \mathbf{1}_{B_r(y^*)}(y) \operatorname{sgn}\left(\int_{B_r(y^*)} \left(\check{\pi}_n(u \mid y^*) - \tilde{\pi}(u \mid y')\right) \mu_y(\mathrm{d}y') \right).
$$

Then, by the prequel, we have by monotone convergence, as $m \to \infty$,

$$
\begin{aligned}
&\iint f_m(y,u)(\check{\pi}_n(\mathrm{d}u \mid y) - \tilde{\pi}(\mathrm{d}u \mid y))\mu_y(\mathrm{d}y) \\
&\quad = \int_{B_r(y^*)} \sum_{u \in \mathcal{U}} f_m(y,u)(\check{\pi}_n(u \mid y) - \tilde{\pi}(u \mid y))\mu_y(\mathrm{d}y) \\
&\to \int_{B_r(y^*)} \sum_{u \in \mathcal{U}} \operatorname{sgn}\left(\int_{B_r(y^*)} \left(\check{\pi}_n(u \mid y^*) - \tilde{\pi}(u \mid y')\right) \mu_y(\mathrm{d}y') \right) (\check{\pi}_n(u \mid y) - \tilde{\pi}(u \mid y))\mu_y(\mathrm{d}y) \\
&\quad = \sum_{u \in \bar{\mathcal{U}}(r)} \left| \int_{B_r(y^*)} (\check{\pi}_n(u \mid y) - \tilde{\pi}(u \mid y))\, \mu_y(\mathrm{d}y) \right| \\
&+ \sum_{u \notin \bar{\mathcal{U}}(r)} \operatorname{sgn}\left(\int_{B_r(y^*)} \left(\check{\pi}_n(u \mid y^*) - \tilde{\pi}(u \mid y')\right) \mu_y(\mathrm{d}y') \right) \int_{B_r(y^*)} (\check{\pi}_n(u \mid y) - \tilde{\pi}(u \mid y))\mu_y(\mathrm{d}y) \\
&\quad \geq \sum_{u \in \bar{\mathcal{U}}(r)} \left| \int_{B_r(y^*)} (\check{\pi}_n(u \mid y) - \tilde{\pi}(u \mid y))\, \mu_y(\mathrm{d}y) \right| \\
&\qquad - \sum_{u \notin \bar{\mathcal{U}}(r)} \left| \int_{B_r(y^*)} (\check{\pi}_n(u \mid y) - \check{\pi}_n(u \mid y^*))\mu_y(\mathrm{d}y) \right| \\
&\qquad - \sum_{u \notin \bar{\mathcal{U}}(r)} \left| \int_{B_r(y^*)} (\check{\pi}_n(u \mid y^*) - \tilde{\pi}(u \mid y))\mu_y(\mathrm{d}y) \right| \\
&\quad > \frac{\epsilon'}{2} \cdot 2\mu_y(B_r(y^*)) - \frac{2\epsilon' - \epsilon}{4} \cdot 2\mu_y(B_r(y^*)) - \frac{\epsilon - \epsilon'}{2} \cdot 2\mu_y(B_r(y^*))
\end{aligned}
$$

$$= \frac{1}{2}\left(\epsilon' - \frac{\epsilon}{2}\right)\mu_y(B_r(y^*)) > 0$$

i.o. in n, for the first term by Eq. (F.8.16) and Eq. (F.8.17), second by Eq. (F.8.15) and third by Eq. (F.8.14), noting that $\epsilon' > \frac{\epsilon}{2}$.

Hence, we may choose m^* such that e.g.

$$\iint f_{m^*}(y,u)(\check{\pi}_n(\mathrm{d}u \mid y) - \tilde{\pi}(\mathrm{d}u \mid y))\mu_y(\mathrm{d}y) > \frac{1}{4}\left(\epsilon' - \frac{\epsilon}{2}\right)\mu_y(B_r(y^*)).$$

LOWER BOUND. Finally, by noting that $\frac{1}{m^*} f_{m^*} \in \mathrm{Lip}(1)$ and applying the Kantorovich-Rubinstein duality, we have

$$\begin{aligned} W_1(\mu_y \otimes \check{\pi}_n, \mu_y \otimes \tilde{\pi}) &= \sup_{f \in \mathrm{Lip}(1)} \iint f(y,u)(\check{\pi}_n(\mathrm{d}u \mid y) - \tilde{\pi}(\mathrm{d}u \mid y))\mu_y(\mathrm{d}y) \\ &\geq \iint \frac{1}{m^*} f_{m^*}(y,u)(\check{\pi}_n(\mathrm{d}u \mid y) - \tilde{\pi}(\mathrm{d}u \mid y))\mu_y(\mathrm{d}y) \\ &> \frac{1}{m^*}\frac{1}{4}\left(\epsilon' - \frac{\epsilon}{2}\right)\mu_y(B_r(y^*)) > 0 \end{aligned}$$

i.o. in n, and therefore $\mu_y \otimes \check{\pi}_n \not\to \mu_y \otimes \tilde{\pi}$. But $\tilde{h} = \mu_y \otimes \tilde{\pi}$ was assumed to be the limit of $\mu_y \otimes \check{\pi}_n$, leading to a contradiction. Hence, μ_y-a.e. convergence must hold.

(II) LIPSCHITZ EXTENSION OF LOWER-LEVEL POLICIES. For finite actions, note that $\mathcal{P}(\mathcal{U})$ is (Lipschitz) equivalent to a subset of the Hilbert space $\mathbb{R}^{|\mathcal{U}|}$. Therefore, by the Kirszbraun-Valentine theorem (see e.g. [393, Theorem 4.2.3]), we can modify $\check{\pi}$ to be L_Π-Lipschitz not only μ_y-a.e., but on full $\mathcal{Y}$.

(III) EQUALITY OF LIMITS. We show that for any $\epsilon > 0$, $W_1(h, \mu_y \otimes \check{\pi}) < \epsilon$, which implies $W_1(h, \mu_y \otimes \check{\pi}) = 0$ and therefore $h = \mu_y \otimes \check{\pi}$. First, note that by the triangle inequality, we have

$$W_1(h, \mu_y \otimes \check{\pi}) \leq W_1(h, \mu_y \otimes \check{\pi}_n) + W_1(\mu_y \otimes \check{\pi}_n, \mu_y \otimes \check{\pi})$$

and thus by $\mu_y \otimes \check{\pi}_n \to h$ for sufficiently large n, it suffices to show $W_1(\mu_y \otimes \check{\pi}_n, \mu_y \otimes \check{\pi}) < \epsilon$.

By the prequel, we choose a measurable set $A \subseteq \mathcal{Y}$ such that $\mu_y(A) < \frac{\epsilon}{2\,\mathrm{diam}(\mathcal{U})}$ and $\check{\pi}_n(y)$ converges uniformly on $\mathcal{Y} \setminus A$. Now by uniform convergence, we choose n sufficiently large such that $W_1(\check{\pi}_n(y), \check{\pi}(y)) < \frac{\epsilon}{2}$ on $\mathcal{Y} \setminus A$. By Kantorovich-Rubinstein duality, we have

$$\begin{aligned} W_1(\mu_y \otimes \check{\pi}_n, \mu_y \otimes \check{\pi}) &= \sup_{f \in \mathrm{Lip}(1)} \iint f(y,u)(\check{\pi}_n(\mathrm{d}u \mid y) - \check{\pi}(\mathrm{d}u \mid y))\mu_y(\mathrm{d}y) \\ &\leq \int \left(\sup_{f \in \mathrm{Lip}(1)} \int f(y,u)(\check{\pi}_n(\mathrm{d}u \mid y) - \check{\pi}(\mathrm{d}u \mid y)) \right) \mu_y(\mathrm{d}y) \\ &= \int W_1(\check{\pi}_n(y), \check{\pi}(y))\mu_y(\mathrm{d}y) \\ &= \int_A W_1(\check{\pi}_n(y), \check{\pi}(y))\mu_y(\mathrm{d}y) + \int_{\mathcal{Y}\setminus A} W_1(\check{\pi}_n(y), \check{\pi}(y))\mu_y(\mathrm{d}y) \\ &< \frac{\epsilon}{2\,\mathrm{diam}(\mathcal{U})}\,\mathrm{diam}(\mathcal{U}) + \left(1 - \frac{\epsilon}{2\,\mathrm{diam}(\mathcal{U})}\right)\frac{\epsilon}{2} < \epsilon. \end{aligned}$$

This completes the proof. □

F.9 PROOF OF PROPOSITION 4.3.2

Proof of Proposition 4.3.2. The proof is similar to the proof of Proposition 4.3.1 by induction. We begin by showing the first statement. We show $\bar{\mu}_t = \hat{\mu}_t$ at all times $t \in \mathcal{T}$, as it then follows that $\bar{J}(\bar{\pi}) = \sum_{t=0}^{\infty} \gamma^t r(\bar{\mu}_t) = \sum_{t=0}^{\infty} \gamma^t r(\hat{\mu}_t) = \hat{J}(\hat{\pi})$ under deterministic $\hat{\pi} \in \hat{\Pi}$. At time $t = 0$, we have by definition $\bar{\mu}_0 = \mu_0 = \hat{\mu}_0$. Assume $\bar{\mu}_t = \hat{\mu}_t$ at time t, then at time $t+1$, we have

$$\hat{\mu}_{t+1} = \hat{T}(\hat{\mu}_t, h_t) = \iiint p(x, u, \hat{\mu}_t) \check{\pi}[h_t](\mathrm{d}u \mid y) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}_t) \hat{\mu}_t(\mathrm{d}x) \tag{F.9.18}$$

$$= \iiint p(x, u, \mu_t) \bar{\pi}_t(\mathrm{d}u \mid y, \bar{\mu}_t) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \bar{\mu}_t) \bar{\mu}_t(\mathrm{d}x) = \bar{\mu}_{t+1} \tag{F.9.19}$$

by definition of $\bar{\pi}_t(\nu) = \check{\pi}[h_t]$, which is the desired statement. An analogous proof for the second statement in the opposite direction completes the proof. □

F.10 PROOF OF COROLLARY 4.3.3

Proof of Corollary 4.3.3. As in the proof of Corollary 4.3.2, we first show Dec-POMFC optimality of $\Phi(\Psi(\hat{\pi}))$. Assume $J(\Phi(\Psi(\hat{\pi}))) < \sup_{\pi' \in \Pi} J(\pi')$. Then there exists $\pi' \in \Pi$ such that $J(\Phi(\Psi(\hat{\pi}))) < J(\pi')$. But by Proposition 4.3.1, there exists $\bar{\pi}' \in \bar{\Pi}$ such that $\bar{J}(\bar{\pi}') = J(\pi')$. Further, by Proposition 4.3.2, there exists $\hat{\pi}' \in \hat{\Pi}$ such that $\bar{J}(\bar{\pi}') = \hat{J}(\hat{\pi}')$. Thus, $\hat{J}(\hat{\pi}) = \bar{J}(\bar{\pi}) = J(\Phi(\Psi(\hat{\pi}))) < J(\pi') = \bar{J}(\bar{\pi}') = \hat{J}(\hat{\pi}')$, which contradicts $\hat{\pi} \in \arg\max_{\hat{\pi}'} \hat{J}(\hat{\pi}')$. Therefore, $\Phi(\Psi(\hat{\pi})) \in \arg\max_{\pi' \in \Pi} J(\pi')$. Hence, $\Phi(\Psi(\hat{\pi}))$ fulfills the conditions of Corollary 4.3.1, completing the proof. □

F.11 PROOF OF PROPOSITION 4.3.3

Proof of Proposition 4.3.3. First, note that

$$\begin{aligned} |\nabla_y \kappa(y_b, y)| &= \exp\left(\frac{-\|y_b - y\|^2}{2\sigma^2}\right) \frac{|\langle y_b - y, y \rangle|}{2\sigma^2} \\ &\leq \frac{1}{2\sigma^2} \operatorname{diam}(\mathcal{Y}) \max_{y \in \mathcal{Y}} \|y\| \end{aligned}$$

for diameter $\operatorname{diam}(\mathcal{Y}) < \infty$ by compactness of $\mathcal{Y}$, which is equal one for discrete spaces. Further,

$$\left| \sum_{b' \in [M_{\mathcal{Y}}]} \kappa(y_{b'}, y) \right| = \sum_{b' \in [M_{\mathcal{Y}}]} \kappa(y_{b'}, y) \geq M_{\mathcal{Y}} \exp\left(-\frac{\operatorname{diam}(\mathcal{Y})^2}{2\sigma^2}\right)$$

and $|\kappa(y_b, y)| \leq 1$.

Hence, the RBF kernel $y \mapsto \kappa(y_b, y) p_b = \exp(\frac{-\|y_b - y\|^2}{2\sigma^2})$ with parameter $\sigma^2 > 0$ on $\mathcal{Y}$ is Lipschitz for any $b \in [M_{\mathcal{Y}}]$, since for any $y, y' \in \mathcal{Y}$,

$$\begin{aligned} &\left|\nabla_y \left(Z^{-1}(y) \kappa(y_b, y)\right)\right| \\ &= \left| \frac{\nabla_y \kappa(y_b, y) \sum_{b' \in [M_{\mathcal{Y}}]} \kappa(y_{b'}, y) + \sum_{b' \in [M_{\mathcal{Y}}]} \nabla_y \kappa(y_{b'}, y) \kappa(y_b, y)}{\left(\sum_{b' \in [M_{\mathcal{Y}}]} \kappa(y_{b'}, y)\right)^2} \right| \end{aligned}$$

$$\leq \frac{1}{M_{\mathcal{Y}}^2 \exp^2\left(-\frac{\operatorname{diam}(\mathcal{Y})^2}{2\sigma^2}\right)}\left(\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})\max_{y\in\mathcal{Y}}\|y\|M_{\mathcal{Y}} + M_{\mathcal{Y}}\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})\max_{y\in\mathcal{Y}}\|y\|\right)$$
$$= \frac{\operatorname{diam}(\mathcal{Y})\max_{y\in\mathcal{Y}}\|y\|}{\sigma^2 M_{\mathcal{Y}}\exp^2\left(-\frac{\operatorname{diam}(\mathcal{Y})^2}{2\sigma^2}\right)}$$

for any $b \in [M_{\mathcal{Y}}]$. Hence, by noting that the following supremum is invariant to addition of constants,

$$\begin{aligned}
&W_1\left(Z^{-1}(y)\sum_{b\in[M_{\mathcal{Y}}]}\kappa(y_b,y)p_b, Z^{-1}(y')\sum_{b\in[M_{\mathcal{Y}}]}\kappa(y_b,y')p_b\right)\\
&= \sup_{f\in\mathrm{Lip}(1)}\int f\left(Z^{-1}(y)\sum_{b\in[M_{\mathcal{Y}}]}\kappa(y_b,y)\mathrm{d}p_b - Z^{-1}(y')\sum_{b\in[M_{\mathcal{Y}}]}\kappa(y_b,y')\mathrm{d}p_b\right)\\
&= \sup_{f\in\mathrm{Lip}(1),|f|\leq\frac{1}{2}\operatorname{diam}(\mathcal{U})}\int f\left(Z^{-1}(y)\sum_{b\in[M_{\mathcal{Y}}]}\kappa(y_b,y) - Z^{-1}(y')\sum_{b\in[M_{\mathcal{Y}}]}\kappa(y_b,y')\right)\mathrm{d}p_b\\
&\leq \sum_{b\in[M_{\mathcal{Y}}]}\left|Z^{-1}(y)\kappa(y_b,y) - Z^{-1}(y')\kappa(y_b,y')\right| \sup_{f\in\mathrm{Lip}(1),|f|\leq\frac{1}{2}\operatorname{diam}(\mathcal{U})}\int f\mathrm{d}p_b\\
&\leq M_{\mathcal{Y}}\frac{\operatorname{diam}(\mathcal{Y})\max_{y\in\mathcal{Y}}\|y\|}{\sigma^2 M_{\mathcal{Y}}\exp^2\left(-\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})^2\right)}\|y-y'\|\cdot\frac{1}{2}\operatorname{diam}(\mathcal{U}).
\end{aligned}$$

which is L_Π-Lipschitz if

$$\begin{aligned}
&\frac{\operatorname{diam}(\mathcal{Y})\operatorname{diam}(\mathcal{U})\max_{y\in\mathcal{Y}}\|y\|}{2\sigma^2\exp^2\left(-\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})^2\right)} \leq L_\Pi\\
&\quad\iff \sigma^2\exp^2\left(-\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})^2\right) \geq \frac{1}{L_\Pi}\operatorname{diam}(\mathcal{Y})\operatorname{diam}(\mathcal{U})\max_{y\in\mathcal{Y}}\|y\|.
\end{aligned}$$

Note that such $\sigma^2 > 0$ exists, as $\sigma^2\exp^2\left(-\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})^2\right) \to +\infty$ as $\sigma^2 \to +\infty$. □

F.12 PROOF OF THEOREM 4.3.3

Proof of Theorem 4.3.3. Keeping in mind that we have the **centralized training** system for stationary policy $\hat{\pi}^\theta$ parametrized by θ,

$$\tilde{\xi}_t \sim \hat{\pi}^\theta(\tilde{\mu}_t^N), \quad \check{\pi}_t = \Lambda(\tilde{\xi}_t)$$
$$\tilde{y}_t^i \sim p_{\mathcal{Y}}(\tilde{y}_t^i \mid \tilde{x}_t^i, \tilde{\mu}_t^N), \quad \tilde{u}_t^i \sim \check{\pi}_t(\tilde{u}_t^i \mid \tilde{y}_t^i), \quad \tilde{x}_{t+1}^i \sim p(\tilde{x}_{t+1}^i \mid \tilde{x}_t^i, \tilde{u}_t^i, \tilde{\mu}_t^N), \quad \forall i \in [N],$$

which we obtained by parametrizing the MDP actions via parametrizations $\xi \in \Xi$, the equivalent Dec-MFC MDP system concomitant with Eq. (4.3.36) under parametrization $\Lambda(\xi)$ for lower-level policies is

$$\xi_t \sim \hat{\pi}^\theta(\hat{\mu}_t), \quad \hat{\mu}_{t+1} = \hat{T}(\hat{\mu}_t, \xi_t) \coloneqq \iiint p(x,u,\hat{\mu}_t)\Lambda(\xi_t)(\mathrm{d}u \mid y)p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}_t)\hat{\mu}_t(\mathrm{d}x) \tag{F.12.20}$$

where we now sample ξ_t instead of h_t. Note that for kernel representations, this new $\hat{T}$ is indeed Lipschitz, which follows from Lipschitzness of $\hat{\mu}_t \otimes p_{\mathcal{Y}}(\hat{\mu}_t) \otimes \Lambda(\xi_t)$ in $(\hat{\mu}_t, \xi_t)$.

Lemma F.12.1. *Under Assumptions 4.3.1a and 4.3.3, the transitions $\hat{T}$ of the system with parametrized actions are $L_{\hat{T}}$-Lipschitz with $L_{\hat{T}} \coloneqq 2L_p + L_p L_\lambda + 2L' L_{p\mathcal{Y}}$.*

Proofs for lemmas are found in their respective following sections.

First, we prove $d^N_{\hat{\pi}^\theta} \to d_{\hat{\pi}^\theta}$ in $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ by showing at any time t that under $\hat{\pi}^\theta$, the centralized training system MF $\tilde{\mu}^N_t$ converges to the limiting Dec-MFC MF $\hat{\mu}_t$ in Eq. (F.12.20). The convergence is in the same sense as in Theorem 4.3.1.

Lemma F.12.2. *For any equicontinuous family of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{P}(\mathcal{X})}$, under Assumptions 4.3.1a, 4.3.1b and 4.3.3, at all times t we have*

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(\tilde{\mu}^N_t) - f(\hat{\mu}_t) \right] \right| \to 0. \tag{F.12.21}$$

We also show that $\tilde{Q}^\theta(\mu, \xi) \to Q^\theta(\mu, \xi)$, since we can show the same convergence as in Eq. (F.12.21) for new conditional systems, where for any μ, ξ we let $\tilde{\mu}_0 = \mu = \mu_0$ and $\tilde{\xi}_0 = \xi = \xi_0$ at time zero, where $\tilde{\mu}_0$ is the initial state distribution of the centralized training system.

Lemma F.12.3. *Under Assumptions 4.3.1a and 4.3.3, as $N \to \infty$, we have for any $\mu \in \mathcal{P}(\mathcal{X})$, $\xi \in \Xi$ that*

$$\left| \tilde{Q}^\theta(\mu, \xi) - Q^\theta(\mu, \xi) \right| \to 0.$$

Furthermore, $Q^\theta(\mu, \xi)$ is also continuous by a similar argument.

Lemma F.12.4. *For any equicontinuous family of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{P}(\mathcal{X})}$, under Assumptions 4.3.1a and 4.3.3, at all times $t \in \mathcal{T}$, the conditional expectations version of the MF is continuous in the starting conditions, in the sense that for any $(\mu_n, \xi_n) \to (\mu, \xi)$,*

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(\hat{\mu}_t) \mid \hat{\mu}_0 = \mu_n, \xi_0 = \xi_n \right] - \mathbb{E}\left[f(\hat{\mu}_t) \mid \hat{\mu}_0 = \mu, \xi_0 = \xi \right] \right| \to 0.$$

Lastly, keeping in mind $d^N_{\hat{\pi}^\theta} = (1-\gamma) \sum_{t \in \mathcal{T}} \gamma^t \mathcal{L}_{\hat{\pi}^\theta}(\tilde{\mu}^N_t)$, we have the desired statement

$$\begin{aligned}
&\left\| (1-\gamma)^{-1} \mathbb{E}_{\mu \sim d^N_{\hat{\pi}^\theta}, \xi \sim \hat{\pi}^\theta(\mu)} \left[\tilde{Q}^\theta(\mu, \xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \mu) \right] - \nabla_\theta J(\hat{\pi}^\theta) \right\| \\
&\quad \le (1-\gamma)^{-1} \left\| \mathbb{E}_{\mu \sim d^N_{\hat{\pi}^\theta}, \xi \sim \hat{\pi}^\theta(\mu)} \left[\left(\tilde{Q}^\theta(\mu, \xi) - Q^\theta(\mu, \xi) \right) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \mu) \right] \right\| \\
&\qquad + \left\| (1-\gamma)^{-1} \mathbb{E}_{\mu \sim d^N_{\hat{\pi}^\theta}, \xi \sim \hat{\pi}^\theta(\mu)} \left[Q^\theta(\mu, \xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \mu) \right] - \nabla_\theta J(\hat{\pi}^\theta) \right\| \\
&\quad \le (1-\gamma)^{-1} \left\| \mathbb{E}_{\mu \sim d^N_{\hat{\pi}^\theta}, \xi \sim \hat{\pi}^\theta(\mu)} \left[\left(\tilde{Q}^\theta(\mu, \xi) - Q^\theta(\mu, \xi) \right) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \mu) \right] \right\| \\
&\qquad + \left\| \sum_{t=T}^{\infty} \gamma^t \mathbb{E}_{\xi \sim \hat{\pi}^\theta(\tilde{\mu}^N_t)} \left[Q^\theta(\tilde{\mu}^N_t, \xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \tilde{\mu}^N_t) - Q^\theta(\hat{\mu}_t, \xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \hat{\mu}_t) \right] \right\| \\
&\qquad + \left\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\xi \sim \hat{\pi}^\theta(\tilde{\mu}^N_t)} \left[Q^\theta(\tilde{\mu}^N_t, \xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \tilde{\mu}^N_t) - Q^\theta(\hat{\mu}_t, \xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \hat{\mu}_t) \right] \right\| \\
&\quad \to 0
\end{aligned}$$

for the **first** term from $\tilde{Q}^\theta(\mu,\xi) \to Q^\theta(\mu,\xi)$ uniformly by Lemma F.12.3 and compactness of the domain, for the **second** by Assumptions 4.3.1a and 4.3.3 uniformly bounding $\nabla_\theta \log \pi^\theta$, Q^θ and choosing sufficiently large T, and for the **third** by repeating the argument for Q: Notice that

$$\begin{aligned}
&\left\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\xi\sim\hat{\pi}^\theta(\tilde{\mu}_t^N)} \left[Q^\theta(\tilde{\mu}_t^N,\xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \tilde{\mu}_t^N) - Q^\theta(\hat{\mu}_t,\xi) \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \hat{\mu}_t) \right] \right\| \\
&\leq \left\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\xi\sim\hat{\pi}^\theta(\tilde{\mu}_t^N)} \left[\sum_{t'=T'}^{\infty} \gamma^{t'} \Big(\mathbb{E}\left[r(\hat{\mu}_{t'}) \mid \hat{\mu}_0 = \tilde{\mu}_t^N, \xi_0 = \xi \right] \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \tilde{\mu}_t^N) \right. \right. \right. \\
&\qquad \left. \left. \left. - \mathbb{E}\left[r(\hat{\mu}_{t'}) \mid \hat{\mu}_0 = \hat{\mu}_t, \xi_0 = \xi \right] \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \hat{\mu}_t) \Big) \right] \right\| \\
&\quad + \left\| \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{\xi\sim\hat{\pi}^\theta(\tilde{\mu}_t^N)} \left[\sum_{t'=0}^{T'-1} \gamma^{t'} \Big(\mathbb{E}\left[r(\hat{\mu}_{t'}) \mid \hat{\mu}_0 = \tilde{\mu}_t^N, \xi_0 = \xi \right] \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \tilde{\mu}_t^N) \right. \right. \right. \\
&\qquad \left. \left. \left. - \mathbb{E}\left[r(\hat{\mu}_{t'}) \mid \hat{\mu}_0 = \hat{\mu}_t, \xi_0 = \xi \right] \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \hat{\mu}_t) \Big) \right] \right\|,
\end{aligned}$$

where the inner expectations are on the conditional system. Letting T' sufficiently large bounds the **former** term by uniform bounds on the summands from Assumption 4.3.3. Then, for the **latter** term, apply Lemma F.12.2 at times $t' < T'$ to the functions $f(\mu) = \int \mathbb{E}\left[r(\hat{\mu}_{t'}) \mid \hat{\mu}_0 = \mu, \xi_0 = \xi \right] \nabla_\theta \log \hat{\pi}^\theta(\xi \mid \mu) \hat{\pi}^\theta(\mathrm{d}\xi \mid \mu) \mathrm{d}\xi$, which are continuous up to any finite time t' by Lemma F.12.4 and Assumption 4.3.3. □

F.13 PROOF OF LEMMA F.12.1

Proof of Lemma F.12.1. We have by definition

$$\begin{aligned}
&\iiint p(x,u,\hat{\mu}) \Lambda(\xi)(\mathrm{d}u \mid y) p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}) \hat{\mu}(\mathrm{d}x) \\
&\quad = \iiint p(x,u,\hat{\mu}) \frac{\sum_{b\in[M_{\mathcal{Y}}]} \kappa(y_b,y) \lambda_b(\xi)(\mathrm{d}u)}{\sum_{b\in[M_{\mathcal{Y}}]} \kappa(y_b,y)} p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}) \hat{\mu}(\mathrm{d}x).
\end{aligned}$$

Consider any $\xi, \xi' \in \Xi$, $\hat{\mu}, \hat{\mu}' \in \mathcal{P}(\mathcal{X})$. Then, for readability, write

$$\begin{gathered}
\hat{\mu}_{xy} \coloneqq \hat{\mu} \otimes p_{\mathcal{Y}}(\hat{\mu}), \quad \hat{\mu}_{xyu} \coloneqq \hat{\mu}_{xy} \otimes \Lambda(\xi), \quad \hat{\mu}_{xyux'} \coloneqq \hat{\mu}_{xyu} \otimes p(\hat{\mu}), \\
\hat{\mu}'_{xy} \coloneqq \hat{\mu}' \otimes p_{\mathcal{Y}}(\hat{\mu}'), \quad \hat{\mu}'_{xyu} \coloneqq \hat{\mu}'_{xy} \otimes \Lambda(\xi'), \quad \hat{\mu}'_{xyux'} \coloneqq \hat{\mu}'_{xyu} \otimes p(\hat{\mu}'), \\
\Delta p(\cdot \mid x,u) \coloneqq p(\cdot \mid x,u,\hat{\mu}) - p(\cdot \mid x,u,\hat{\mu}'), \\
\Delta\Lambda(\cdot \mid y) \coloneqq \frac{\sum_b \kappa(y_b,y) \left(\lambda_b(\xi)(\cdot) - \lambda_b(\xi')(\cdot) \right)}{\sum_b \kappa(y_b,y)}, \\
\Delta p_{\mathcal{Y}}(\cdot \mid x) \coloneqq p_{\mathcal{Y}}(\cdot \mid x,\hat{\mu}) - p_{\mathcal{Y}}(\cdot \mid x,\hat{\mu}'), \quad \Delta\mu \coloneqq \hat{\mu} - \hat{\mu}'
\end{gathered}$$

to obtain

$$\begin{aligned}
&W_1 \left(\iiint p(x,u,\hat{\mu}) \frac{\sum_b \kappa(y_b,y) \lambda_b(\xi)(\mathrm{d}u)}{\sum_b \kappa(y_b,y)} p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}) \hat{\mu}(\mathrm{d}x), \right. \\
&\qquad \left. \iiint p(x,u,\hat{\mu}') \frac{\sum_b \kappa(y_b,y) \lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b,y)} p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}') \hat{\mu}'(\mathrm{d}x) \right) \\
&= \sup_{f\in\mathrm{Lip}(1)} \iiiint f(x') \left(\hat{\mu}_{xyux'}(\mathrm{d}x,\mathrm{d}y,\mathrm{d}u,\mathrm{d}x') - \hat{\mu}'_{xyux'}(\mathrm{d}x,\mathrm{d}y,\mathrm{d}u,\mathrm{d}x') \right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{f\in\mathrm{Lip}(1)} \iiiint f(x')\Delta p(\mathrm{d}x' \mid x,u)\hat{\mu}_{xyu}(\mathrm{d}x,\mathrm{d}y,\mathrm{d}u) \\
&\quad + \sup_{f\in\mathrm{Lip}(1)} \iiiint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')\Delta\Lambda(\mathrm{d}u \mid y)\hat{\mu}_{xy}(\mathrm{d}x,\mathrm{d}y)) \\
&\quad + \sup_{f\in\mathrm{Lip}(1)} \iiiint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')\frac{\sum_b \kappa(y_b,y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b,y)}\Delta p_{\mathcal{Y}}(\mathrm{d}y \mid x)\hat{\mu}(\mathrm{d}x) \\
&\quad + \sup_{f\in\mathrm{Lip}(1)} \iiiint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')\frac{\sum_b \kappa(y_b,y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b,y)}p_{\mathcal{Y}}(\mathrm{d}y \mid x,\hat{\mu}')\Delta\mu(\mathrm{d}x) \\
&\leq \sup_{f\in\mathrm{Lip}(1)} \sup_{(x,y,u)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{U}} \left| \int f(x')\Delta p(\mathrm{d}x' \mid x,u) \right| \\
&\quad + \sup_{f\in\mathrm{Lip}(1)} \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left| \iint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')\Delta\Lambda(\mathrm{d}u \mid y) \right| \\
&\quad + \sup_{f\in\mathrm{Lip}(1)} \sup_{x\in\mathcal{X}} \left| \iiint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')\frac{\sum_b \kappa(y_b,y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b,y)}\Delta p_{\mathcal{Y}}(\mathrm{d}y \mid x) \right| \\
&\quad + \sup_{f\in\mathrm{Lip}(1)} \left| \iiiint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')\frac{\sum_b \kappa(y_b,y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b,y)}p_{\mathcal{Y}}(\mathrm{d}y \mid x,\hat{\mu}')\Delta\mu(\mathrm{d}x) \right|
\end{aligned}
$$

bounded by the same arguments as in Theorem 4.3.1:

For the **first** term, we have that the function $x' \mapsto f(x')$ is 1-Lipschitz, and therefore

$$
\sup_{f\in\mathrm{Lip}(1)} \sup_{(x,y,u)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{U}} \left| \int f(x') \left(p(\mathrm{d}x' \mid x,u,\hat{\mu}) - p(\mathrm{d}x' \mid x,u,\hat{\mu}') \right) \right| \leq L_p W_1(\hat{\mu},\hat{\mu}')
$$

by Assumption 4.3.1a.

For the **second** term, we have L_p-Lipschitz $u \mapsto \int f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')$, since for any $f \in \mathrm{Lip}(1)$ and $(x,y) \in \mathcal{X}\times\mathcal{Y}$, we obtain

$$
\begin{aligned}
&\left| \int f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}') - \int f(x')p(\mathrm{d}x' \mid x,u',\hat{\mu}') \right| \\
&\qquad \leq W_1(p(x,u,\hat{\mu}'),p(x,u',\hat{\mu}')) \leq L_p d(u,u')
\end{aligned}
$$

for any $u,u' \in \mathcal{U}$ by Assumption 4.3.1a, and therefore

$$
\begin{aligned}
&\sup_{f\in\mathrm{Lip}(1)} \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left| \iint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')\frac{\sum_b \kappa(y_b,y)\left(\lambda_b(\xi)(\mathrm{d}u) - \lambda_b(\xi')(\mathrm{d}u)\right)}{\sum_b \kappa(y_b,y)} \right| \\
&\quad \leq \frac{\sum_b \kappa(y_b,y)L_p W_1\left(\lambda_b(\xi),\lambda_b(\xi')\right)}{\sum_b \kappa(y_b,y)} \leq L_p L_\lambda d(\xi,\xi')
\end{aligned}
$$

by Assumption 4.3.3.

For the **third** term, we have L'-Lipschitz $y \mapsto \iint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}')\frac{\sum_b \kappa(y_b,y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b,y)}$ where we define $L' \coloneqq L_p \frac{\operatorname{diam}(\mathcal{Y})\operatorname{diam}(\mathcal{U})\max_{y\in\mathcal{Y}}\|y\|}{2\sigma^2\exp^2\left(-\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})^2\right)}$, since for any $f \in \mathrm{Lip}(1)$ and $x \in \mathcal{X}$, we obtain

$$
\begin{aligned}
&\left| \iint f(x')p(\mathrm{d}x' \mid x,u,\hat{\mu}') \left(\frac{\sum_b \kappa(y_b,y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b,y)} - \frac{\sum_b \kappa(y_b,y')\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b,y')} \right) \right| \\
&\quad \leq L_p \cdot \frac{\operatorname{diam}(\mathcal{Y})\operatorname{diam}(\mathcal{U})\max_{y\in\mathcal{Y}}\|y\|}{2\sigma^2\exp^2\left(-\frac{1}{2\sigma^2}\operatorname{diam}(\mathcal{Y})^2\right)} d(y,y')
\end{aligned}
$$

for any $y, y' \in \mathcal{Y}$ by Proposition 4.3.3 and the prequel, and therefore, for any $f \in \mathrm{Lip}(1)$, $x \in \mathcal{X}$

$$\left|\iiint f(x')p(\mathrm{d}x' \mid x, u, \hat{\mu}')\frac{\sum_b \kappa(y_b, y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b, y)}\left(p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}) - p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}')\right)\right| \leq L' L_{p_{\mathcal{Y}}} W_1(\hat{\mu}, \hat{\mu}')$$

by Assumption 4.3.1a.

Lastly, for the **fourth** term, $x \mapsto \iiint f(x')p(\mathrm{d}x' \mid x, u, \hat{\mu}')\frac{\sum_b \kappa(y_b, y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b, y)}p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}')$ is similarly $(L_p + L' L_{p_{\mathcal{Y}}})$-Lipschitz, since for any $f \in \mathrm{Lip}(1)$, we obtain

$$\begin{aligned}
&\left|\iiint f(x')p(\mathrm{d}x' \mid x, u, \hat{\mu}')\frac{\sum_b \kappa(y_b, y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b, y)}p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}')\right. \\
&\qquad \left. - \iiint f(x')p(\mathrm{d}x' \mid x'', u, \hat{\mu}')\frac{\sum_b \kappa(y_b, y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b, y)}p_{\mathcal{Y}}(\mathrm{d}y \mid x'', \hat{\mu}')\right| \\
&\leq \left|\iiint f(x')\left(p(\mathrm{d}x' \mid x, u, \hat{\mu}') - p(\mathrm{d}x' \mid x'', u, \hat{\mu}')\right)\frac{\sum_b \kappa(y_b, y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b, y)}p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}')\right| \\
&+ \left|\iiint f(x')p(\mathrm{d}x' \mid x'', u, \hat{\mu}')\frac{\sum_b \kappa(y_b, y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b, y)}\left(p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}') - p_{\mathcal{Y}}(\mathrm{d}y \mid x'', \hat{\mu}')\right)\right| \\
&\leq L_p d(x, x'') + L' L_{p_{\mathcal{Y}}} d(x, x'') = (L_p + L' L_{p_{\mathcal{Y}}})d(x, x'')
\end{aligned}$$

for any $x, x'' \in \mathcal{X}$ by the prequel, which implies

$$\begin{aligned}
&\sup_{f \in \mathrm{Lip}(1)}\left|\iiiint f(x')p(\mathrm{d}x' \mid x, u, \hat{\mu}')\frac{\sum_b \kappa(y_b, y)\lambda_b(\xi')(\mathrm{d}u)}{\sum_b \kappa(y_b, y)}p_{\mathcal{Y}}(\mathrm{d}y \mid x, \hat{\mu}')\left(\hat{\mu}(\mathrm{d}x) - \hat{\mu}'(\mathrm{d}x)\right)\right| \\
&\quad \leq (L_p + L' L_{p_{\mathcal{Y}}}) W_1(\hat{\mu}, \hat{\mu}').
\end{aligned}$$

Overall, the map $\hat{T}$ is therefore Lipschitz with constant $L_{\hat{T}} = 2L_p + L_p L_\lambda + 2L' L_{p_{\mathcal{Y}}}$. □

F.14 PROOF OF LEMMA F.12.2

Proof of Lemma F.12.2. The proof is the same as the proof of Theorem 4.3.1. The only difference is that for the weak LLN argument, we condition not only on $\tilde{x}_t^N$, but also on $\tilde{\xi}_t$, while for the induction assumption, we still apply to equicontinuous functions by Assumption 4.3.3 and Lemma F.12.1.

In other words, for the weak LLN we use

$$\begin{aligned}
\mathbb{E}\left[d_\Sigma\left(\tilde{\mu}_{t+1}^N, \tilde{T}\left(\tilde{\mu}_t^N, \tilde{\xi}_t\right)\right)\right] &= \sum_{m=1}^{\infty} 2^{-m}\mathbb{E}\left[\left|\int f_m \,\mathrm{d}\left(\tilde{\mu}_{t+1}^N - \tilde{T}\left(\tilde{\mu}_t^N, \tilde{\xi}_t\right)\right)\right|\right] \\
&\leq \sup_{m \geq 1}\mathbb{E}\left[\mathbb{E}\left[\left|\int f_m \,\mathrm{d}\left(\tilde{\mu}_{t+1}^N - \tilde{T}\left(\tilde{\mu}_t^N, \tilde{\xi}_t\right)\right)\right| \,\middle|\, \tilde{x}_t^N, \tilde{\xi}_t\right]\right].
\end{aligned}$$

and obtain

$$\begin{aligned}
&\mathbb{E}\left[\left|\int f_m \,\mathrm{d}\left(\tilde{\mu}_{t+1}^N - \tilde{T}\left(\tilde{\mu}_t^N, \tilde{\xi}_t\right)\right)\right| \,\middle|\, \tilde{x}_t^N, \tilde{\xi}_t\right]^2 \\
&\quad = \mathbb{E}\left[\left|\frac{1}{N}\sum_{i \in [N]}\left(f_m(x_{t+1}^i) - \mathbb{E}\left[f_m(x_{t+1}^i) \,\middle|\, \tilde{x}_t^N, \tilde{\xi}_t\right]\right)\right| \,\middle|\, \tilde{x}_t^N, \tilde{\xi}_t\right]^2 \leq \frac{4}{N} \to 0,
\end{aligned}$$

while for the induction assumption we use the equicontinuous functions $\mu \mapsto \int f(\hat{T}(\mu, \xi))\hat{\pi}^\theta(\xi \mid \mu)\mathrm{d}\xi$ by Assumption 4.3.3 and Lemma F.12.1. □

F.15 PROOF OF LEMMA F.12.3

Proof of Lemma F.12.3. We show the required statement by first showing at all times t that

$$\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\tilde{\mu}_t^N) - f(\hat{\mu}_t) \;\middle|\; \tilde{\mu}_0 = \mu = \hat{\mu}_0, \tilde{\xi}_0 = \xi = \xi_0\right]\right| \to 0. \tag{F.15.22}$$

This is clear at time $t = 0$ by $\tilde{\mu}_0 = \mu = \hat{\mu}_0$, $\tilde{\xi}_0 = \xi = \xi_0$ and the weak LLN argument as in the proof of Lemma F.12.2. At time $t = 1$, we analogously have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\tilde{\mu}_1^N) - f(\hat{\mu}_1) \;\middle|\; \tilde{\mu}_0 = \mu = \hat{\mu}_0, \tilde{\xi}_0 = \xi = \xi_0\right]\right| \\
&\quad\le \sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\tilde{\mu}_1^N) - f(\hat{T}(\tilde{\mu}_0^N, \tilde{\xi}_0)) \;\middle|\; \tilde{\mu}_0 = \mu, \tilde{\xi}_0 = \xi\right]\right| \\
&\qquad + \sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\hat{T}(\tilde{\mu}_0^N, \tilde{\xi}_0)) - f(\hat{\mu}_1) \;\middle|\; \tilde{\mu}_0 = \mu = \hat{\mu}_0, \tilde{\xi}_0 = \xi = \xi_0\right]\right|,
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\tilde{\mu}_{t+1}^N) - f(\hat{\mu}_{t+1}) \;\middle|\; \tilde{\mu}_0 = \mu = \hat{\mu}_0, \tilde{\xi}_0 = \xi = \xi_0\right]\right| \\
&\quad\le \sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\tilde{\mu}_{t+1}^N) - \int f(\hat{T}(\tilde{\mu}_t^N, \xi'))\hat{\pi}^\theta(\xi' \mid \tilde{\mu}_t^N)\mathrm{d}\xi' \;\middle|\; \tilde{\mu}_0 = \mu, \tilde{\xi}_0 = \xi\right]\right| \\
&\qquad + \sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\int f(\hat{T}(\tilde{\mu}_t^N, \xi'))\hat{\pi}^\theta(\xi' \mid \tilde{\mu}_t^N)\mathrm{d}\xi' - f(\hat{\mu}_{t+1}) \;\middle|\; \tilde{\mu}_0 = \mu = \hat{\mu}_0, \tilde{\xi}_0 = \xi = \xi_0\right]\right|,
\end{aligned}$$

for times $t + 1 \ge 1$, each with the weak LLN arguments applied to the **former** terms (conditioning not only on $\tilde{x}_t^N$, but also $\tilde{\xi}_t$), and the induction assumption applied to the **latter** terms, using the equicontinuous functions $\mu \mapsto \int f(\hat{T}(\mu, \xi))\hat{\pi}^\theta(\xi \mid \mu)\mathrm{d}\xi$ by Assumption 4.3.3 and Lemma F.12.1. □

F.16 PROOF OF LEMMA F.12.4

Proof of Lemma F.12.4. For any $(\mu_n, \xi_n) \to (\mu, \xi)$, we show again by induction over all times t that for any equicontinuous family $\mathcal{F}$,

$$\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\hat{\mu}_t) \mid \hat{\mu}_0 = \mu_n, \xi_0 = \xi_n\right] - \mathbb{E}\left[f(\hat{\mu}_t) \mid \hat{\mu}_0 = \mu, \xi_0 = \xi\right]\right| \to 0 \tag{F.16.23}$$

as $N \to \infty$, from which the result follows. At time $t = 0$, we have by definition

$$\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\hat{\mu}_0) \mid \hat{\mu}_0 = \mu_n, \xi_0 = \xi_n\right] - \mathbb{E}\left[f(\hat{\mu}_0) \mid \hat{\mu}_0 = \mu, \xi_0 = \xi\right]\right| = 0.$$

Analogously, at time $t = 1$ we have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\hat{\mu}_1) \mid \hat{\mu}_0 = \mu_n, \xi_0 = \xi_n\right] - \mathbb{E}\left[f(\hat{\mu}_1) \mid \hat{\mu}_0 = \mu, \xi_0 = \xi\right]\right| \\
&\quad= \sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(\hat{T}(\hat{\mu}_0, \xi_0)) \;\middle|\; \hat{\mu}_0 = \mu_n, \xi_0 = \xi_n\right] - \mathbb{E}\left[f(\hat{T}(\hat{\mu}_0, \xi_0)) \;\middle|\; \hat{\mu}_0 = \mu, \xi_0 = \xi\right]\right| \\
&\quad= \sup_{f\in\mathcal{F}}\left|f(\hat{T}(\hat{\mu}_n, \xi_n)) - f(\hat{T}(\hat{\mu}, \xi))\right| \to 0
\end{aligned}$$

by equicontinuous f and continuous $\hat{T}$ from Lemma F.12.1.

Now assuming that Eq. (F.16.23) holds at time $t \geq 1$, then at time $t+1$ we have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(\hat{\mu}_{t+1}) \mid \hat{\mu}_0 = \mu_n, \xi_0 = \xi_n\right] - \mathbb{E}\left[f(\hat{\mu}_{t+1}) \mid \hat{\mu}_0 = \mu, \xi_0 = \xi\right]\right| \\
&\quad = \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\int f(\hat{T}(\hat{\mu}_t, \xi'))\hat{\pi}^\theta(\xi' \mid \hat{\mu}_t)\mathrm{d}\xi' \;\middle|\; \hat{\mu}_0 = \mu_n, \xi_0 = \xi_n\right]\right. \\
&\qquad\qquad \left. -\mathbb{E}\left[\int f(\hat{T}(\hat{\mu}_t, \xi'))\hat{\pi}^\theta(\xi' \mid \hat{\mu}_t)\mathrm{d}\xi' \;\middle|\; \hat{\mu}_0 = \mu, \xi_0 = \xi\right]\right| \\
&\quad = \sup_{g\in\mathcal{G}} \left|\mathbb{E}\left[g(\hat{\mu}_t) \mid \hat{\mu}_0 = \mu_n, \xi_0 = \xi_n\right] - \mathbb{E}\left[g(\hat{\mu}_t) \mid \hat{\mu}_0 = \mu, \xi_0 = \xi\right]\right| \to 0
\end{aligned}$$

by induction assumption on equicontinuous functions $g \in \mathcal{G}$ by Assumptions 4.3.1a and 4.3.3, Lemma F.12.1, and equicontinuous $f \in \mathcal{F}$, as in Theorem 4.3.1.

The convergence of $\left|Q^\theta(\mu_n, \xi_n) - Q^\theta(\mu, \xi)\right| \to 0$ thus follows by Assumption 4.3.1a. □

F.17 ADDITIONAL EXPERIMENTS

In this section, we give additional details on experiments. The mathematical description of problems can be found in Appendix F.19.

We use the manifolds as depicted in Figure F.1 and as described in the following. Here, we visualize the qualitative results as in the main text for the remaining topologies. Due to technical limitations, all agents are drawn, including the ones behind a surface. To indicate where an agent belongs, we colorize the inside of the agent with the color of its corresponding surface.

TORUS MANIFOLD. The (flat) torus manifold is obtained from the square $[-1, 1]^2$ by identifying $(x, -1) \sim (x, 1)$ and $(-1, y) \sim (1, y)$ for all $x, y \in [-1, 1]$. For the metric, we use the toroidal distance inherited from the Euclidean distance d_2, which can be computed as

$$d(x, y) = \min_{t_1,t_2\in\{-1,0,1\}} d_2(x, y + 2t_1e_1 + 2t_2e_2)$$

where e_1, e_2 denote unit vectors. In Figure F.1, we visualize the torus by mapping each point $(x, y) \in [-1, 1]^2$ to a point (X, Y, Z) in 3D space, given by

$$\begin{aligned}
X &= (2 + 0.75\cos(\pi(x + 1)))\cos(\pi(y + 1)), \\
Y &= (2 + 0.75\cos(\pi(x + 1)))\sin(\pi(y + 1)), \\
Z &= 0.75\sin(\pi(x + 1)).
\end{aligned}$$

The results have been described in the main text in Figure 4.20. Here, also note that the torus – by periodicity and periodic boundary conditions – can essentially be understood as the case of an *infinite plane*, consisting of infinitely many copies of the square $[-1, 1]^2$ laid next to each other.

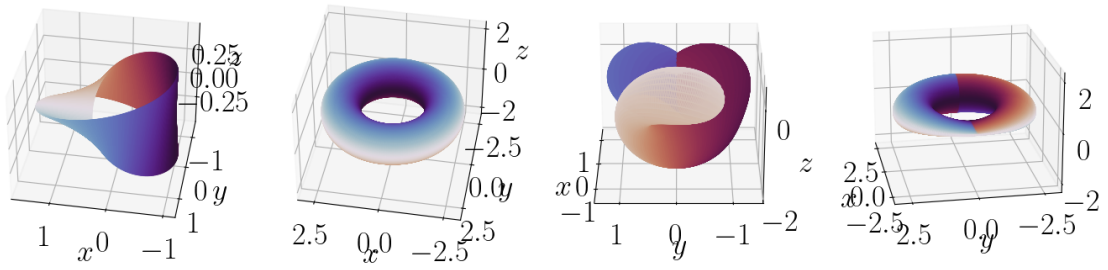


FIGURE F.1: Two-dimensional manifolds visualized in three-dimensional space. In order from left to right: Möbius strip, torus, projective plane (Boy's surface), and Klein bottle (pinched torus).

MÖBIUS STRIP. The Möbius strip is obtained from the square $[-1,1]^2$ by instead only identifying $(-1,-y) \sim (1,y)$ for all $y \in [-1,1]$, i.e. only the top and bottom side of the square, where directions are flipped. We then use the inherited distance

$$d(x,y) = \min_{t_2 \in \{-1,0,1\}} d_2(x, (1 - 2 \cdot \mathbf{1}_{t_2 \neq 0}, 1)^T \odot y + 2t_2 e_2)$$

where $\odot$ denotes the elementwise (Hadamard) product.

We visualize the Möbius strip in Figure F.1 by mapping each point $(x,y) \in [-1,1]^2$ to

$$\begin{aligned} X &= \left(1 + \frac{x}{2} \cos\left(\frac{\pi}{2}(y+1)\right)\right) \cos\left(\pi(y+1)\right), \\ Y &= \left(1 + \frac{x}{2} \cos\left(\frac{\pi}{2}(y+1)\right)\right) \sin\left(\pi(y+1)\right), \\ Z &= \frac{x}{2} \sin\left(\frac{\pi}{2}(y+1)\right). \end{aligned}$$

As we can see in Figure F.2, the behavior of agents is learned as expected: Agents learn to align along one direction on the Möbius strip.

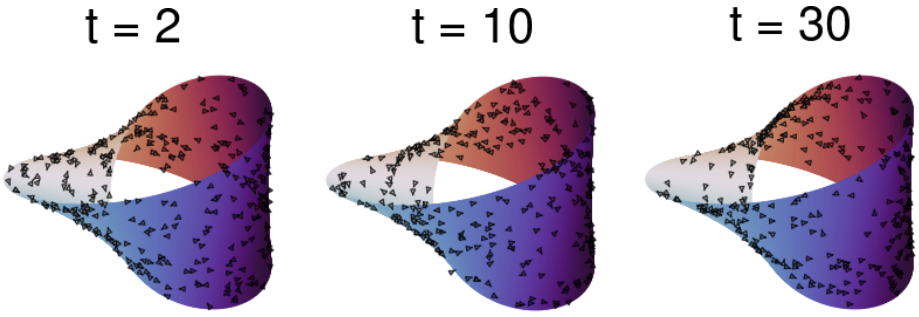


FIGURE F.2: Qualitative visualization of Vicsek behavior on the Möbius strip manifold for uniform initialization. The 300 agents (triangles) align into one direction on the Möbius strip.

PROJECTIVE PLANE. Analogously, the projective plane is obtained by identifying and flipping both sides of the square $[-1,1]^2$, i.e. $(-x,-1) \sim (x,1)$ and $(-1,-y) \sim (1,y)$ for all $x,y \in [-1,1]$. We use the inherited distance

$$d(x,y) = \min_{t_1,t_2 \in \{-1,0,1\}} d_2(x, (1 - 2 \cdot \mathbf{1}_{t_2 \neq 0}, 1 - 2 \cdot \mathbf{1}_{t_1 \neq 0})^T \odot y + 2t_1 e_1 + 2t_2 e_2)$$

and though an accurate visualization in less than four dimensions is difficult, we visualize in Figure F.3 by mapping each point $(x, y) \in [-1, 1]^2$ to a point (X, Y, Z) on the so-called Boy's surface, with

$$\begin{aligned}
z &= \frac{x+1}{2} \exp\left(i\pi(y+1)\right), \\
g_1 &= -\frac{3}{2} \operatorname{Im}\left(\frac{z(1-z^4)}{z^6 + \sqrt{5}z^3 - 1}\right), \\
g_2 &= -\frac{3}{2} \operatorname{Re}\left(\frac{z(1-z^4)}{z^6 + \sqrt{5}z^3 - 1}\right), \\
g_3 &= \operatorname{Im}\left(\frac{1+z^6}{z^6 + \sqrt{5}z^3 - 1}\right) - \frac{1}{2}, \\
X &= \frac{g_1}{g_1^2 + g_2^2 + g_3^2}, \\
Y &= \frac{g_2}{g_1^2 + g_2^2 + g_3^2}, \\
Z &= \frac{g_2}{g_1^2 + g_2^2 + g_3^2}.
\end{aligned}$$

As we can see in Figure F.3, under the inherited metric and radial parametrization, agents tend to gather at the bottom of the surface.

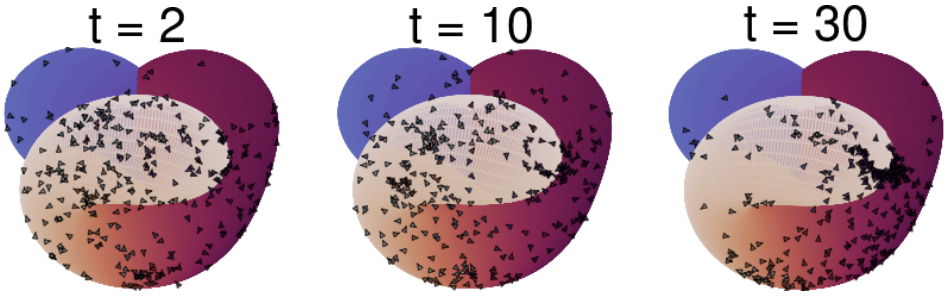


FIGURE F.3: Qualitative visualization of Vicsek behavior on the projective plane manifold for uniform initialization. The 300 agents (triangles) align over time by gathering at the bottom right.

KLEIN BOTTLE. Similarly, the Klein bottle is obtained by identifying both sides of the square $[-1, 1]^2$ and flipping one side, i.e. $(x, -1) \sim (x, 1)$ and $(-1, -y) \sim (1, y)$ for all $x, y \in [-1, 1]$. We use the inherited distance

$$d(x, y) = \min_{t_1, t_2 \in \{-1, 0, 1\}} d_2(x, (1 - 2 \cdot \mathbf{1}_{t_2 \neq 0}, 1)^T \odot y + 2t_1 e_1 + 2t_2 e_2)$$

and visualize in Figure F.4 by the pinched torus, i.e. mapping each $(x, y) \in [-1, 1]^2$ to a point (X, Y, Z) with

$$\begin{aligned}
X &= (2 + 0.75 \cos\left(\pi(x+1)\right)) \cos\left(\pi(y+1)\right), \\
Y &= (2 + 0.75 \cos\left(\pi(x+1)\right)) \sin\left(\pi(y+1)\right), \\
Z &= 0.75 \sin\left(\pi(x+1)\right)) \cos\left(\frac{\pi}{2}(y+1)\right).
\end{aligned}$$

As we can see in Figure F.4, agents may align by aggregating on the inner and outer ring, such that they may avoid switching sides at the pinch.

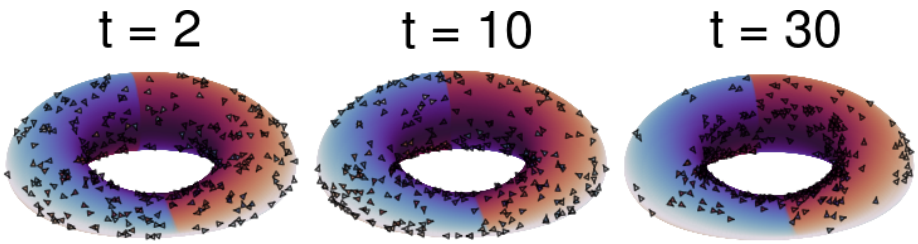


FIGURE F.4: Qualitative visualization of behavior on Klein bottle topology for uniform initialization: The visualization in 3D is limited. Here, we use a pinched torus visualization that inverts itself at the flat pinch (i.e. there is no "connection" between blue and red surfaces at the bottom). Over time, 300 agents (triangles) sometimes align by aggregating on the inner, avoiding side switches at the pinch.

BOX. Lastly, the box manifold is the square $[-1, 1]^2$ equipped with the standard Euclidean topology, i.e. distances between two points $x, y \in [-1, 1]^2$ are given by

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

while the sides of the square are not connected to anything else. We use the box manifold for the following experiments in Aggregation, and mention it here for sake of completeness.

ABLATION ON NUMBER OF AGENTS. As seen in Figures F.5 and F.6, we can successfully train on various numbers of agents, despite the inaccuracy of the MF approximation for fewer agents as inferred from Figure 4.18. This indicates that our algorithm is general and – at least in the considered problems – scales to arbitrary numbers of agents.

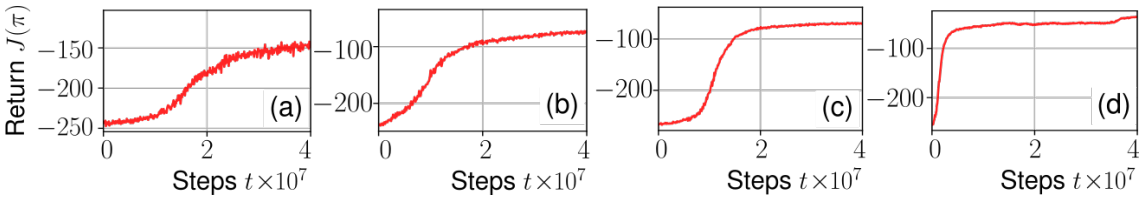


FIGURE F.5: Training curves for Vicsek (torus), using RBF / discretization solutions. (a): RBF, $N = 25$; (b): Discretization, $N = 25$; (c): RBF, $N = 50$; (d) : Discretization, $N = 50$.

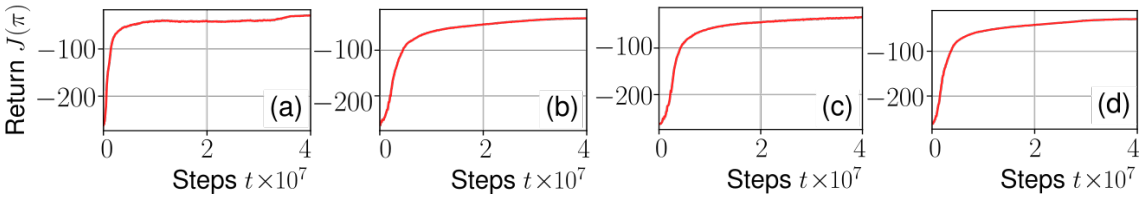


FIGURE F.6: Training curves for Vicsek (torus), using RBF / discretization solutions. (a): RBF, $N = 100$; (b): Discretization, $N = 100$; (c): RBF, $N = 150$; (d) : Discretization, $N = 150$.

QUALITATIVE RESULTS FOR KURAMOTO. The Kuramoto model, see Figure F.7, demonstrates instability during training and subsequent lower-grade qualitative behavior compared to the Vicsek model. This disparity persists even when considering more intricate topologies, despite being a specialization of the Vicsek model. One explanation is that the added movement makes it easier

to align agents over time. Another general explanation could be that, despite initially distributing agents uniformly across the region of interest, the learned policy causes the agents to aggregate into a few or even a single cluster (though we do not observe such behavior in Figure F.7). The closer particles are to each other, the greater the likelihood that they perceive a similar or identical MF, prompting alignment only in local clusters. A similar behavior is observed in the classical Vicsek model, where agents tend to move in the same direction after interaction. Consequently, they remain within each other's interaction region and have the potential to form compounds provided there are no major disturbances. These can come from either other particles or excessively high levels of noise [394]. Although agents are able to align, the desired alignment remains to be improved, either via more parameter tuning or improved algorithms.

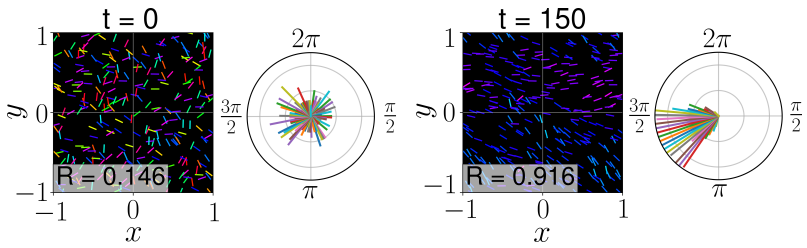


FIGURE F.7: Qualitative behavior of the learned behavior in the Kuramoto model with histogram over angles, where in contrast to Vicsek, agents remain fixed in their initial position.

EFFECT OF KERNEL METHOD. While for low dimensions, the effect of kernel methods is not as pronounced and mostly ensure theoretical guarantees, in Figures F.8 and F.9 we can see that training via our RBF-based methods outperforms discretization-based methods for dimensions higher than 3 as compared to a simple gridding of the probability simplex with associated histogram representation of the MF. Here, for the RBF method in Aggregation, we place 5 equidistant points y_b on the axis of each dimension. This is also the reason for why the discretization-based approach is better for low dimensions $d = 2$ or $d = 3$, as more points will have more control over actions of agents, and can therefore achieve better results, in exchange for tractability in high dimensions. This shows the advantage of RBF-based methods in more complex, high-dimensional problems. While the RBF-based method continues to learn even for higher dimensions up to $d = 5$, the discretization-based solution eventually stops learning due to very large action spaces leading to increased noise on the gradient. The advantage is not just in terms of sample complexity, but also in terms of wall clock time, as the computation over exponentially many bins also takes more CPU time as shown in Table F.1.

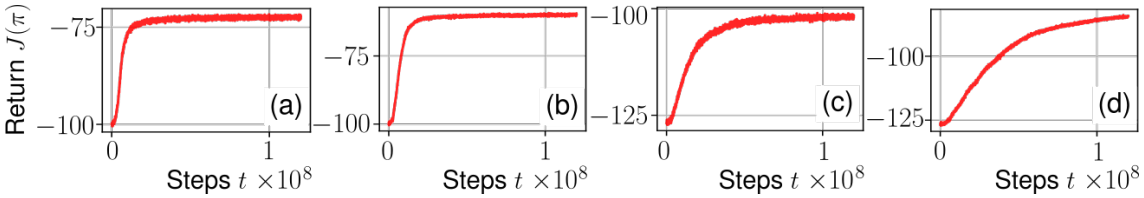


FIGURE F.8: Training curves for d-dimensional Aggregation, RBF vs. discretization. (a): RBF, $d = 2$; (b): Discretization, $d = 2$; (c): RBF, $d = 3$; (d) : Discretization, $d = 3$.

ABLATIONS ON TIME DEPENDENCY AND STARTING CONDITIONS. As discussed in the main text, we also verify the effect of using a non-time-dependent open-loop sequence of lower-level policies, and also an ablation over different starting conditions. In particular, for starting conditions, to begin we will consider the **uniform** initialization as well as the **beta-1**, **beta-2** and **beta-3**

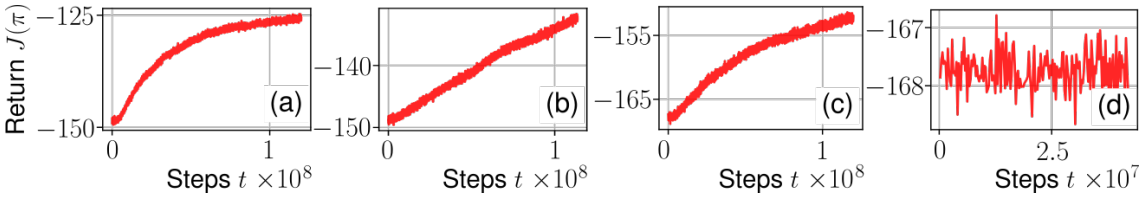


FIGURE F.9: Training curves for d-dimensional Aggregation, RBF vs. discretization. (a): RBF, $d = 4$; (b): Discretization, $d = 4$; (c): RBF, $d = 5$; (d) : Discretization, $d = 5$.

TABLE F.1: Wall clock time for one training step averaged over 500 iterations in d-dimensional Aggregation, for 50 agents.

Dimensionality d	RBF MFC [s]	Discretization MFC [s]	MARL (IPPO) [s]
2	5.64	5.69	146.58
3	6.16	7.96	147.03
4	6.97	17.26	147.29
5	8.31	76.33	146.91

initializations with a beta distribution over each dimension of the states, using $\alpha = \beta = 0.5$, $\alpha = \beta = 0.25$ and $\alpha = \beta = 0.75$ respectively.

As we can see in Figure F.10, the behavior learned for the Vicsek problem on the torus with $N = 200$ agents allows for using the first lower-level policy $\check{\pi}_0$ at all times t under the Gaussian initialization used in training to nonetheless achieve alignment. This validates the fact that a time-variant open-loop sequence of lower-level policies is not always needed, and the results even hold for slightly different initial conditions from the ones used in training.

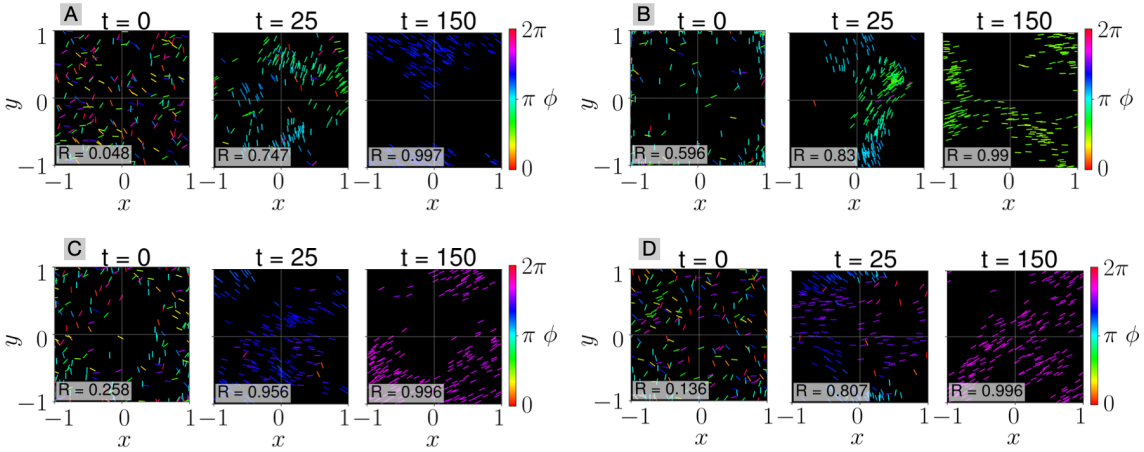


FIGURE F.10: Open-loop behavior by using the lower-level decision policy at time $t = 0$ for all times, on Vicsek (torus) with $N = 200$ agents and various initializations. A: uniform initialization; B: beta-2 initialization; C: beta-1 initialization; D: beta-3 initialization.

Analogously, we consider some more strongly concentrated and heterogeneous initializations: The **peak-normal** initialization is given by a more concentrated zero-centered diagonal Gaussian with covariance $\sigma^2 = 0.1$. The **squeezed-normal** is the same initialization as in training, except for dividing the variance in the y-axis by 10. The **multiheaded-normal** initialization is a mixture of two equisized Gaussian distributions in the upper-right and lower-left quadrant, where in comparison to

the training initialization, position variances are halved. Finally, the **bernoulli-multiheaded-normal** additionally changes the weights of two Gaussians to 0.75 and 0.25 respectively.

As seen in Figure F.11, the lower-level policy $\check{\pi}_0$ for Gaussian initialization from training easily transfers and generalizes to more complex initializations. However, the behavior may naturally be more suboptimal due to the training process likely never seeing more strongly concentrated and heterogeneous distributions of agents. For example, in the peak-normal initialization in Figure F.11, we see that the agents begin relatively aligned, but will first misalign in order to align again, as the learned policy was trained to handle only the wider Gaussian initialization.

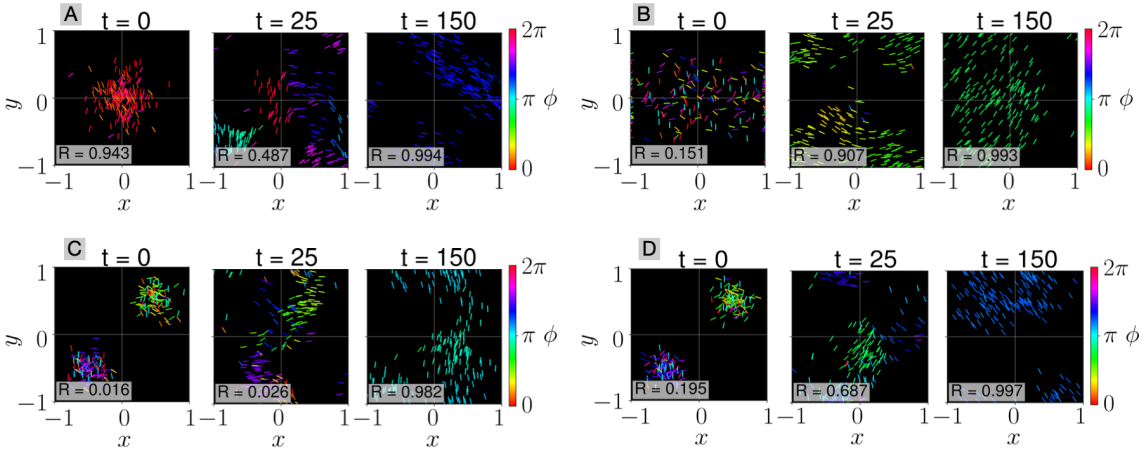


FIGURE F.11: Open-loop behavior by using the lower-level decision policy at time $t = 0$ for all times, on Vicsek (torus) with $N = 200$ agents and various initializations. A: peak-normal initialization; B: squeezed-normal initialization; C: multiheaded-normal initialization; D: bernoulli-multiheaded-normal initialization.

NO OBSERVATIONS. As an additional verification of the positive effect of MF guidance on PG training, we also perform experiments for training PPO without any RL observations, as in the previous paragraph we verified the applicability of learned behavior even without observing the MF that is observed by RL during training. In Figure F.12 we see that PPO is unable to learn useful behavior, despite the existence of such a time-invariant lower-level policy from the preceding paragraph, underlining the empirical importance of MF guidance that we derived.

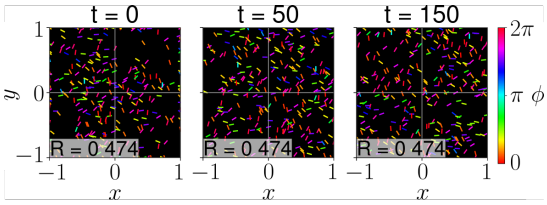


FIGURE F.12: Qualitative behavior of training *without observations* for Vicsek (torus).

TRANSFER TO DIFFERING AGENT COUNTS. In Figure F.13, we see qualitatively that the behavior learned for $N = 200$ agents transfers to different, lower numbers of agents as well. The result is congruent with the results shown in the main text, such as in Figure 4.18, and further supports the fact that our method scales to nearly arbitrary numbers of agents.

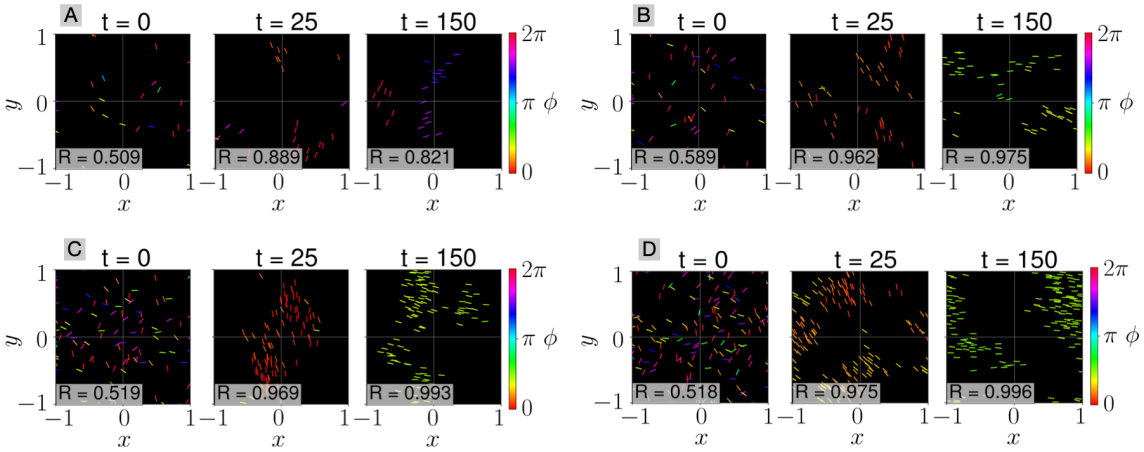


FIGURE F.13: Qualitative behavior of the policy learned for $N = 200$ on Vicsek (torus), transferred to different numbers of agents N. A: $N = 25$; B: $N = 50$; C: $N = 100$; D: $N = 150$.

FORWARD VELOCITY CONTROL. We also allow agents to alternatively control their maximum velocity in the range $[0, v_0]$. Forward velocity can similarly be controlled, and allows for more uniform spreading of agents in contrast to the case where velocity cannot be controlled. This shows some additional generalization of our algorithm to variants of collective behavior problems. The corresponding final qualitative behavior is depicted in Figure F.14.

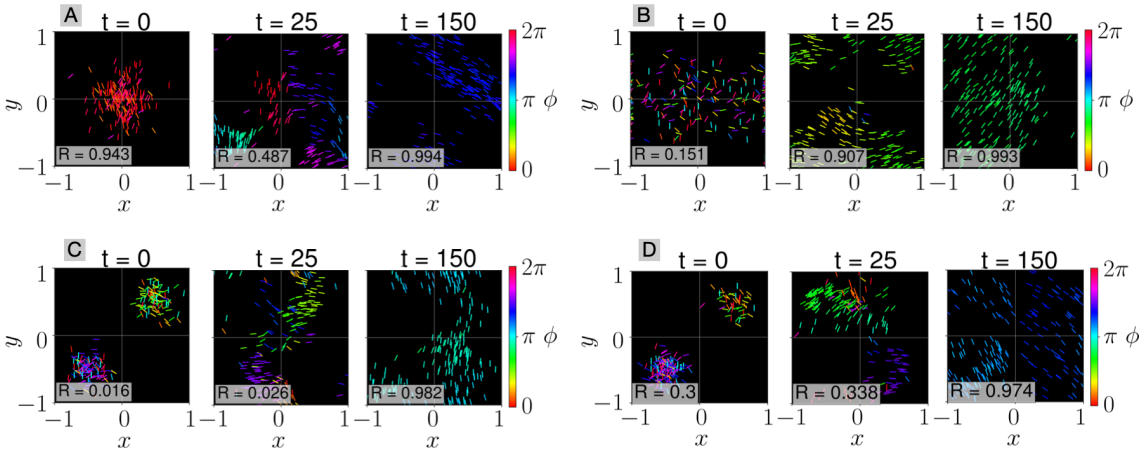


FIGURE F.14: Qualitative behavior on Vicsek (torus) with $N = 200$ agents, additional forward velocity control, and various initializations. A: peak-normal initialization; B: squeezed-normal initialization; C: multiheaded-normal initialization; D: bernoulli-multiheaded-normal initialization.

COMPARISON OF IPPO AND MAPPO FOR LOW NUMBERS OF AGENTS. Lastly, for completeness we show the comparison of IPPO and MAPPO training results for various numbers of agents in Figures F.15 and F.16. The overall achieved performances are overall comparable to the results of the Dec-POMFPPO method in Figure 4.18.

F.18 EXPERIMENTAL DETAILS

We use the RLlib 2.0.1 (Apache-2.0 license) [76] implementation of PPO [73] for both MARL via IPPO, and our Dec-POMFPPO. For MAPPO, we used the MARLlib 1.0.0 framework [91], which builds

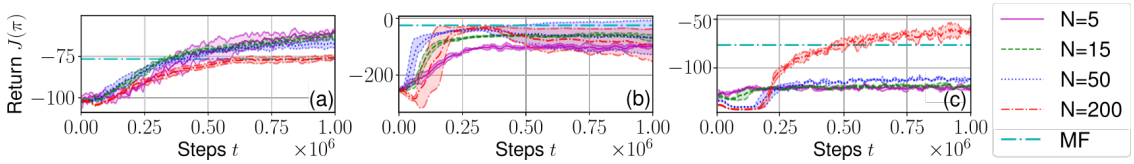


FIGURE F.15: IPPO training curves (episode return) with shaded standard deviation over 3 seeds and various N, in (a) Aggregation (box), (b) Vicsek (torus), (c) Kuramoto (torus). For comparison, we also plot the best return averaged over 3 seeds for Dec-POMFPPO in Figure 4.16 (MF).

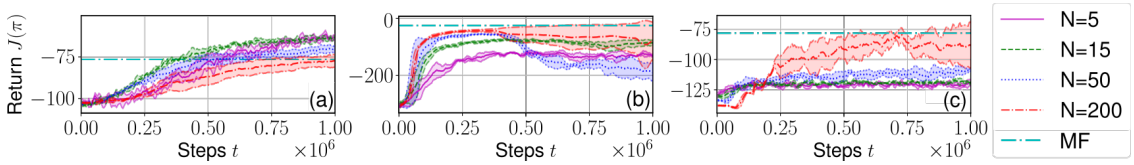


FIGURE F.16: MAPPO training curves (episode return) with shaded standard deviation over 3 seeds and various N, in (a) Aggregation (box), (b) Vicsek (torus), (c) Kuramoto (torus). For comparison, we also plot the best return averaged over 3 seeds for Dec-POMFPPO in Figure 4.16 (MF).

upon RLlib. For our experiments, we used no GPUs and around 60 000 Intel Xeon Platinum 9242 CPU core hours, and each training run usually took at most three days by training on up to 96 CPU cores. Implementation-wise, for the upper-level policy NNs learned by PPO, we use two hidden layers with 256 nodes and $\tanh$ activations, parametrizing diagonal Gaussians over the MDP actions $\xi \in \Xi$ (parametrizations of lower-level policies).

In Aggregation, we define the parameters $\xi \in \Xi$ for continuous spaces $\mathcal{X}, \mathcal{Y}, \mathcal{U} \subseteq \mathbb{R}^d$ by values in $\Xi := [-1, 1]^{2d}$. Each component of ξ is then mapped affinely to mean in $\mathcal{U}$ or diagonal covariance in $[\epsilon, 0.5 + \epsilon]$ with $\epsilon = 10^{-10}/4$, of each dimension. Meanwhile, in Vicsek and Kuramoto, we pursue a "discrete action space" approach, letting $\Xi := [-1, 1]^3$. We then affinely map components of $\xi \in \Xi$ to $[\epsilon, 0.5 + \epsilon]$, which are normalized to constitute probabilities of actions in $\{-1, 0, 1\} \subseteq \mathcal{U}$.

For the kernel-based representation of MFs in d-dimensional state spaces $\mathcal{X}$, we define points x_b by the center points of a d-dimensional gridding of spaces via equisized ($M_\mathcal{X} = 5^d$ hypercubes) partitions. For the histogram, we similarly use the equisized hypercube partitions. For observation spaces $\mathcal{Y}$ and the kernel-based representation of lower-level policies, unless noted otherwise (e.g. in the high-dimensional experiments below, where we use less than exponentially many points), we do the same but additionally rescale the center points $\tilde{y}_b$ around zero, giving $y_b = c\tilde{y}_b$ for some $c > 0$ and $M_\mathcal{Y} = 5^d$. We use $c = 0.75$ for Aggregation and $c = 0.1$ for Vicsek and Kuramoto. For the (diagonal) bandwidths of RBF kernels, in Aggregation we use $\sigma = 0.12/\sqrt{M_\mathcal{X}}$ for states and $\sigma = 0.12c$ for observations. In Vicsek and Kuramoto, we use $\sigma = 0.12/\sqrt{2}$ for state positions, $\sigma = 0.12\pi$ for state angles, and $\sigma = 0.06c$ or $\sigma = 0.12\pi c$ for the first or second component of observations respectively. For IPPO and MAPPO, we observe the observations y_t directly. For hyperparameters of PPO, see Table F.2.

F.19 PROBLEM DETAILS

In this section, we will discuss in more detail the problems considered in our experiments.

AGGREGATION. The Aggregation problem is a problem where agents must aggregate into a single point. Here, $\mathcal{X} = \mathcal{Y} = [-1, 1]^d \subseteq \mathbb{R}^d$ for some dimensionality parameter $d \in \mathbb{N}$, and

TABLE F.2: Hyperparameter configurations for PPO.

Hyperparameter	Value
Discount factor γ	0.99
GAE lambda	1
KL coefficient	0.03
Clip parameter	0.2
Learning rate	0.00005
Training batch size B_{len}	4000
Mini-batch size b_{len}	1000
Steps per batch N_{PPO}	5

analogously $\mathcal{U} = [-1, 1]^d$ for per-dimension movement actions. Observations are the own, noisily observed position, and movements are similarly noisy, using Gaussian noise. Overall, the dynamics are given by

$$\begin{aligned} y_t &\sim \mathcal{N}\left(x_t, \operatorname{diag}(\sigma_y^2, \dots \sigma_y^2)\right), \\ x_{t+1} &\sim \mathcal{N}\left(x_t + v_0 \frac{u_t}{\max(1, \|u_t\|_2)}, \operatorname{diag}(\sigma_x^2, \dots \sigma_x^2)\right) \end{aligned}$$

for some velocity v_0, where additionally, observations and states that are outside of the box $[-1, 1]^d$ are projected back to the box.

The reward function for aggregation of agents is defined as

$$r(\mu_t) = -c_d \iint \|x - y\|_1 \mu_t(\mathrm{d}x)\mu_t(\mathrm{d}y) - c_u \iint \left\| \frac{u}{\max(1, \|u\|_2)} \right\|_1 \pi_t(\mathrm{d}u \mid x)\mu_t(\mathrm{d}x),$$

for some disaggregation cost $c_d > 0$ and action cost $c_u > 0$, where we allow the dependence of rewards on actions as well: Note that our framework still applies to the above dependence on actions, as discussed in Section 4.3.1, by rewriting the system in the following way. At any even time $2t$, the agents transition from state $x \in \mathcal{X}$ to state-actions $(x, u) \in \mathcal{X} \times \mathcal{U}$, which will constitute the states of the new system. At the following odd times $2t + 1$, the transition is sampled for the given state-actions. In this way, the MF is over $\mathcal{X} \cup (\mathcal{X} \times \mathcal{U})$ and allows description of dependencies on the state-action distributions instead of only the state distribution.

For the experiments, we use $\sigma_x^2 = 0.04$, $\sigma_y^2 = 0.04$ and $v_0 = 0.1$. The initial distribution of agent positions is a Gaussian centered around zero, with variance 0.4. The cost coefficients are $c_d = 1$ and $c_u = 0.1$. For simulation purposes, we consider episodes with length $T = 100$.

VICSEK. In classical Vicsek models, each agent is coupled to every other agent within a predefined interaction region. The agents have a fixed maximum velocity $v_0 > 0$, and attempt to align themselves with the neighboring particles within their interaction range $D > 0$. The equations governing the dynamics of the i-th agent in the classical Vicsek model are given in continuous time by

$$\begin{aligned} \mathrm{d}p^i &= (v_0 \sin(\phi^i), v_0 \cos(\phi^i))^T \mathrm{d}t \\ \mathrm{d}\phi^i &= \frac{1}{|N_i|} \sum_{j \in N_i} \sin\left(\phi^j - \phi^i\right) \mathrm{d}t + \sigma \mathrm{d}W \end{aligned}$$

for all agents i, where N_i denotes the set of agents within the interaction region, $N_i \coloneqq \{j \in [N] \mid d(x^i, x^j) \leq D\}$, and W denotes Brownian motion.

We consider a discrete-time variant where agents may *control* independently how to adjust their angles in order to achieve a global objective (e.g. alignment, misalignment, aggregation). For states $x_t \equiv (p_t, \phi_t)$, actions u_t and observations y_t, we have

$$\begin{aligned}
(\bar{x}, \bar{y})^T &= \left(\iint \sin(\phi - \phi_t) \mathbf{1}_{d(p_t,p) \leq D} \mu_t(\mathrm{d}p, \mathrm{d}\phi), \iint \cos(\phi - \phi_t) \mathbf{1}_{d(p_t,p) \leq D} \mu_t(\mathrm{d}p, \mathrm{d}\phi) \right)^T, \\
y_t &= \left(\|(\bar{x}, \bar{y})^T\|_2, \operatorname{atan2}(\bar{x}, \bar{y}) \right)^T, \\
p_{t+1} &= (p_{t,1} + v_0 \sin(\phi_t), p_{t,2} + v_0 \cos(\phi_t))^T, \\
\phi_{t+1} &\sim \mathcal{N}(\phi_t + \omega_0 u_t, \sigma_\phi^2)
\end{aligned}$$

for some maximum angular velocity $\omega_0 > 0$ and noise covariance $\sigma_\phi^2 > 0$, where $\operatorname{atan2}(x, y)$ is the angle from the positive x-axis to the vector $(x, y)^T$. Therefore, we have $\mathcal{X} = [-1, 1]^2 \times [0, 2\pi)$, where positions are equipped with the corresponding topologies discussed in Appendix F.17, and standard Euclidean spaces $\mathcal{Y} = [-1, 1]^2$ and $\mathcal{U} = [-1, 1]$. Importantly, agents only observe the relative headings of other agents within the interaction region. As a result, it is impossible to model such a system using standard MFC techniques.

As cost functions, we consider rewards via the polarization, plus action cost as in Aggregation. Defining polarization similarly to e.g. [395],

$$\begin{aligned}
\mathrm{pol}_t &\coloneqq \iint \angle(x, \bar{x}_t) \mu_t(\mathrm{d}p, \mathrm{d}\phi), \\
\angle(x, y) &\coloneqq \arccos\left((\cos(\phi), \sin(\phi))^T \cdot \frac{y}{\|y\|_2} \right), \\
\bar{x}_t &\coloneqq \iint (\cos(\phi), \sin(\phi))^T \mu_t(\mathrm{d}p, \mathrm{d}\phi)
\end{aligned}$$

where high values of pol_t indicate misalignment, we define the rewards for alignment

$$r(\mu_t) = -c_a \mathrm{pol}_t - c_u \iint |u| \pi_t(\mathrm{d}u \mid x) \mu_t(\mathrm{d}x),$$

and analogously for misalignment

$$r(\mu_t) = +c_a \mathrm{pol}_t - c_u \iint |u| \pi_t(\mathrm{d}u \mid x) \mu_t(\mathrm{d}x).$$

For our training, unless noted otherwise, we let $D = 0.25$, $v_0 = 0.075$, $\omega_0 = 0.2$, $\sigma_\phi = 0.02$ and μ_0 as a zero-centered (clipped) diagonal Gaussian with variance 0.4. The cost coefficients are $c_a = 1$ and $c_u = 0.1$. For simulation purposes, we consider episodes with length $T = 200$.

KURAMOTO. The Kuramoto model can be obtained from the Vicsek model by setting the maximal velocity v_0 of the above equations to zero. Hence, we obtain a random geometric graph, where agents see only their neighbor's state distribution within the interaction region, and the neighbors are static per episode. For parameters, we let $D = 0.25$, $v_0 = 0$, $\omega_0 = 0.2$, $\sigma_\phi = 0$ and μ_0 as a zero-centered (clipped) Gaussian with variance 0.4. The cost coefficients are $c_a = 1$ and $c_u = 0.1$. For simulation purposes, we consider episodes with length $T = 200$.

SYMBOL	DESCRIPTION
$\mathbb{N}$	The set of natural numbers.
$\mathbb{N}_0$, $\mathbb{N}_{\geq 0}$	The set of natural numbers with element zero.
$\mathbb{R}$	The set of real numbers.
$\mathbb{R}_{\geq 0}$	The set of real numbers with elements greater or equal zero.
$\mathcal{P}(A)$	The space of probability measures over a metric space A, equipped with the 1-Wasserstein distance.
$\mathcal{B}_1(A)$	The space of Borel measures over a metric space A bounded by 1.
$\mathbf{1}_A$	The indicator function equal one if the argument is in A (if A is a set) or A is true (if A is a predicate), and zero otherwise.
δ_x	The Dirac measure equal one if the argument contains x, and zero otherwise.
W_1	The 1-Wasserstein distance.
$\mathcal{T}$	The space of discrete decision epochs.
$\mathcal{X}$	The space of agent states.
$\mathcal{U}$	The space of agent actions.
Π	The space of admissible agent policies.
$\mathcal{M}$	The space of mean fields $\mathcal{P}(\mathcal{X})^{\mathcal{T}}$, i.e. state measures at all times $t \in \mathcal{T}$.
$[k]$	The set of all integers $\{1, 2, \ldots, k\}$ up to $k \in \mathbb{N}$.
∇_x	The gradient w.r.t. x.
$\mathbb{E}[X]$	The expectation of a random variable X.
$\mathbb{E}_Y[X]$, $\mathbb{E}[X \mid Y]$	The conditional expectation of a random variable X given Y.
$\mathbb{V}[X]$	The variance of a random variable X.
$\mathcal{L}(X)$	The probability law of a random variable X.
$\mathrm{KL}(p \parallel q)$	The Kullback-Leibler divergence between probability measures p and q.
$\mathrm{Unif}(\cdot)$	The uniform distribution on some implicitly defined space.
$\mathcal{N}(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	The multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
$\mathcal{W}$	The space of graphons.
$\lVert\cdot\rVert_\square$	The cut norm.
$\mathcal{W}_k$	The space of k-uniform hypergraphons.
$\lVert\cdot\rVert_{\square^{k-1}}$	The generalized cut (semi-)norm for k-uniform hypergraphons.
$\bigsqcup$	The disjoint union.
$r(A, m)$	The set of all distinct non-empty subsets of A with at most m elements.
$r_<(A)$	The set of all distinct non-empty, proper subsets of A.
$r(A)$	The set of all distinct non-empty subsets of A.
$\mathrm{Sym}(A)$	The set of all permutations of a set A.
$\mathrm{Sym}^{\mathrm{ind}}_<[k]$	The space of bounded, symmetric functions $f\colon r_<[k] \to \mathbb{R}$.
$\mathrm{Sym}^{\mathrm{ind}}_\leq[k]$	The space of bounded, symmetric functions $f\colon r[k] \to \mathbb{R}$.
$\mathrm{Sym}^{\overline{\mathrm{ind}}}[k]$	The space of bounded, symmetric functions $f\colon [k] \to \mathbb{R}$.

ACRONYMS

A2C Advantage Actor Critic

APF Artificial Potential Field

CTDE Centralized Training Decentralized Execution

DEC-MFC Decentralized Mean-Field-Observable Mean Field Control

DEC-POMDP Decentralized Partially-Observable Markov Decision Process

DEC-POMFC Decentralized Partially-Observable Mean Field Control

DEC-POMFPPO Decentralized Partially-Observable Mean Field PPO

DPP Dynamic Programming Principle

DQN Deep Q-Network

ER Erdős–Rényi

FP Fictitious Play

FPI Fixed Point Iteration

GMFE Graphon Mean Field Equilibrium

GMFG Graphon Mean Field Game

HMFE Hypergraphon Mean Field Equilibrium

IPPO Independent PPO

LLM Large Language Model

LLN Law of Large Numbers

M2M Machine-To-Machine

M3FA2C Major-Minor Mean Field Advantage Actor Critic

M3FC Major-Minor Mean Field Control

M3FG Major-Minor Mean Field Game

M3FMARL Major-Minor Mean Field Multi-Agent Reinforcement Learning

M3FNE Major-Minor Mean Field Nash Equilibrium

M3FPPO Major-Minor Mean Field PPO

MAPPO Multi-Agent PPO

MARL Multi-Agent Reinforcement Learning

MDP Markov Decision Process

MEC Multi-access Edge Computing

MF Mean Field

MFC Mean Field Control

MFE Mean Field Equilibrium

MFG Mean Field Game

MMDP Multi-agent Markov Decision Process

PG Policy Gradient

POMDP Partially-Observable Markov Decision Process

PPO Proximal Policy Optimization

RBF Radial Basis Function

RL Reinforcement Learning

RQ Research Question

SG Stochastic Game

UAV Unmanned Aerial Vehicle

UE User Edge device

BIBLIOGRAPHY

[1] K. Cui, C. Fabian, and H. Koeppl, "Major-minor mean field multi-agent reinforcement learning", *Proc. ICML*, 2024.

[2] K. Cui, S. Hauck, C. Fabian, and H. Koeppl, "Learning decentralized partially observable mean field control for artificial collective behavior", in *Proc. ICLR*, 2024, pp. 1–40.

[3] K. Cui, G. Dayanıklı, M. Laurière, M. Geist, O. Pietquin, and H. Koeppl, "Learning discrete-time major-minor mean field games", in *Proc. AAAI*, vol. 38, 2024, pp. 9616–9625.

[4] K. Cui, L. Baumgärtner, M. B. Yilmaz, M. Li, C. Fabian, B. Becker, L. Xiang, M. Bauer, and H. Koeppl, "UAV swarms for joint data ferrying and dynamic cell coverage via optimal transport descent and quadratic assignment", in *Proc. LCN*, 2023, pp. 1–8.

[5] K. Cui, M. Li, C. Fabian, and H. Koeppl, "Scalable task-driven robotic swarm control via collision avoidance and learning mean-field control", in *Proc. ICRA*, 2023, pp. 1192–1199.

[6] K. Cui, M. B. Yilmaz, A. Tahir, A. Klein, and H. Koeppl, "Optimal offloading strategies for edge-computing via mean-field games and control", in *Proc. GLOBECOM*, 2022, pp. 976–981.

[7] K. Cui and H. Koeppl, "Learning graphon mean field games and approximate Nash equilibria", in *Proc. ICLR*, 2022, pp. 1–31.

[8] K. Cui, A. Tahir, M. Sinzger, and H. Koeppl, "Discrete-time mean field control with environment states", in *Proc. CDC*, 2021, pp. 5239–5246.

[9] K. Cui and H. Koeppl, "Approximately solving mean field games via entropy-regularized deep reinforcement learning", in *Proc. AISTATS*, 2021, pp. 1909–1917.

[10] A. Tahir, K. Cui, A. Rizk, and H. Koeppl, "Collaborative optimization of the age of information under partial observability", *to appear in IFIP Networking 2024, arXiv:2312.12977*, 2023.

[11] A. K. Sreedhara, D. Padala, S. Mahesh, K. Cui, M. Li, and H. Koeppl, "Optimal collaborative transportation for under-capacitated vehicle routing problems using aerial drone swarms", in *Proc. ICRA*, IEEE, 2024, pp. 8401–8407.

[12] M. Li, K. Cui, and H. Koeppl, "A modular aerial system based on homogeneous quadrotors with fault-tolerant control", pp. 8408–8414, 2024.

[13] C. Fabian, K. Cui, and H. Koeppl, "Learning mean field games on sparse graphs: A hybrid graphex approach", in *Proc. ICLR*, 2024, pp. 1–39.

[14] C. Fabian, K. Cui, and H. Koeppl, "Learning sparse graphon mean field games", in *Proc. AISTATS*, 2023, pp. 4486–4514.

[15] A. Tahir, K. Cui, and H. Koeppl, "Learning mean-field control for delayed information load balancing in large queuing systems", in *Proc. ICPP*, 2022, pp. 1–11.

[16] R. Ourari, K. Cui, A. Elshamanhory, and H. Koeppl, "Nearest-neighbor-based collision avoidance for quadrotors via reinforcement learning", in *Proc. ICRA*, 2022, pp. 293–300.

[17] K. Cui, W. R. KhudaBukhsh, and H. Koeppl, "Hypergraphon mean field games", *Chaos*, vol. 32, no. 11, 2022.

[18] K. Cui, W. R. KhudaBukhsh, and H. Koeppl, "Motif-based mean-field approximation of interacting particles on clustered networks", *Phys. Rev. E*, vol. 105, no. 4, p. L042301, 2022.

[19] C. Fabian, K. Cui, and H. Koeppl, "Mean field games on weighted and directed graphs via colored digraphons", *IEEE Control Syst. Lett.*, vol. 7, pp. 877–882, 2022.

[20] K. Cui, A. Tahir, G. Ekinci, A. Elshamanhory, Y. Eich, M. Li, and H. Koeppl, "A survey on large-population systems and scalable multi-agent reinforcement learning", *in preparation for AI Review, arXiv:2209.03859*, 2022.

[21] A. Tahir, K. Cui, and H. Koeppl, "Sparse mean field load balancing in large localized queueing systems", *submitted to MobiHoc 2024, arXiv:2312.12973*, 2023.

[22] J.-M. Lasry and P.-L. Lions, "Mean field games", *Japanese J. Math*, vol. 2, pp. 229–260, 2007.

[23] M. Huang, R. P. Malhamé, and P. E. Caines, "Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle", *Commun. Inf. Syst.*, vol. 6, no. 3, pp. 221–252, 2006.

[24] N. Saldi, T. Basar, and M. Raginsky, "Markov–Nash equilibria in mean-field games with discounted cost", *SIAM J. Control Optim.*, vol. 56, no. 6, pp. 4256–4287, 2018.

[25] M. Laurière, S. Perrin, M. Geist, and O. Pietquin, "Learning mean field games: A survey", *arXiv:2205.12944*, 2022.

[26] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT press, 2018.

[27] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey", *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.

[28] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey", *IEEE Trans. Transp. Syst.*, pp. 4909–4926, 2022.

[29] M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang, "Autonomous navigation of stratospheric balloons using reinforcement learning", *Nature*, vol. 588, no. 7836, pp. 77–82, 2020.

[30] J. Tožička, B. Szulyovszky, G. de Chambrier, V. Sarwal, U. Wani, and M. Gribulis, "Application of deep reinforcement learning to UAV fleet control", in *Proc. SAI Intell. Syst. Conf.*, 2018, pp. 1169–1177.

[31] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning", in *Proc. IROS*, 2017, pp. 1343–1350.

[32] M. Everett, Y. F. Chen, and J. P. How, "Collision avoidance in pedestrian-rich environments with deep reinforcement learning", *IEEE Access*, vol. 9, pp. 10 357–10 377, 2021.

[33] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning", *arXiv:1312.5602*, 2013.

[34] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search", *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[35] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, *et al.*, "Starcraft II: A new challenge for reinforcement learning", *arXiv:1708.04782*, 2017.

[36] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, *et al.*, "Mastering Atari, Go, Chess and Shogi by planning with a learned model", *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[37] Y.-J. Hu and S.-J. Lin, "Deep reinforcement learning for optimizing finance portfolio management", in *Proc. AICAI*, 2019, pp. 14–20.

[38] A. Charpentier, R. Elie, and C. Remlinger, "Reinforcement learning in economics and finance", *Comput. Econ.*, pp. 1–38, 2021.

[39] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback", in *Proc. NeurIPS*, vol. 35, 2022, pp. 27 730–27 744.

[40] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms", in *Handbook of Reinforcement Learning and Control*, K. G. Vamvoudakis, Y. Wan, F. L. Lewis, and D. Cansever, Eds., Cham: Springer International Publishing, 2021, pp. 321–384.

[41] P. Muller, S. Omidshafiei, M. Rowland, K. Tuyls, J. Perolat, S. Liu, D. Hennes, L. Marris, M. Lanctot, E. Hughes, *et al.*, "A generalized training approach for multiagent learning", in *Proc. ICLR*, 2020, pp. 1–35.

[42] G. Papoudakis, F. Christianos, L. Schäfer, and S. V. Albrecht, "Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks", in *Proc. NeurIPS Track Datasets Benchmarks*, 2021.

[43] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

[44] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. De Cote, "A survey of learning in multiagent environments: Dealing with non-stationarity", *arXiv:1707.09183*, 2017.

[45] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, "The complexity of computing a Nash equilibrium", *SIAM J. Comput.*, vol. 39, no. 1, pp. 195–259, 2009.

[46] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes", *Math. Oper. Res.*, vol. 27, no. 4, pp. 819–840, 2002.

[47] R. J. Glauber, "Time-dependent statistics of the Ising model", *J. Math. Phys.*, vol. 4, no. 2, pp. 294–307, 1963.

[48] D. F. Anderson and T. D. Nguyen, "Deficiency zero for random reaction networks under a stochastic block model framework", *J. Math. Chem.*, vol. 59, no. 9, pp. 2063–2097, 2021.

[49] I. Z. Kiss, J. C. Miller, and P. L. Simon, *Mathematics of Epidemics on Networks: From Exact to Approximate Models*. Springer, 2017, vol. 46.

[50] D. Bruneo, M. Scarpa, A. Bobbio, D. Cerotti, and M. Gribaudo, "Markovian agent modeling swarm intelligence algorithms in wireless sensor networks", *Perform. Eval.*, vol. 69, no. 3–4, pp. 135–149, Mar. 2012.

[51] K. Sugishita, M. A. Porter, M. Beguerisse-Díaz, and N. Masuda, "Opinion dynamics on tie-decay networks", *Phys. Rev. Research*, vol. 3, p. 023 249, 2 Jun. 2021.

[52] A. Barrat, M. Barthlemy, and A. Vespignani, *Dynamical Processes on Complex Networks*. USA: Cambridge University Press, 2008.

[53] A. C. Kizilkale and R. P. Malhame, "Collective target tracking mean field control for electric space heaters", in *Proc. Mediterranean Conf. Control Automat.*, 2014, pp. 829–834.

[54] T. Cabannes, M. Laurière, J. Perolat, R. Marinier, S. Girgin, S. Perrin, O. Pietquin, A. M. Bayen, E. Goubault, and R. Elie, "Solving n-player dynamic routing games with congestion: A mean-field approach", in *Proc. AAMAS*, vol. 21, 2022, pp. 1557–1559.

[55] K. Huang, X. Chen, X. Di, and Q. Du, "Dynamic driving and routing games for autonomous vehicles on networks: A mean field game approach", *Transp. Res. C: Emerg. Technol.*, vol. 128, p. 103 189, 2021.

[56] S. Kar, R. Rehrmann, A. Mukhopadhyay, B. Alt, F. Ciucu, H. Koeppl, C. Binnig, and A. Rizk, "On the throughput optimization in large-scale batch-processing systems", *Perform. Eval.*, vol. 144, p. 102 142, 2020.

[57] W. R. KhudaBukhsh, S. Kar, B. Alt, A. Rizk, and H. Koeppl, "Generalized cost-based job scheduling in very large heterogeneous cluster systems", *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 11, pp. 2594–2604, 2020.

[58] W. R. KhudaBukhsh, J. Rückert, J. Wulfheide, D. Hausheerv, and H. Koeppl, "Analysing and leveraging client heterogeneity in swarming-based live streaming", in *Proc. IFIP Networking*, 2016, pp. 386–394.

[59] S. Eshghi, M. H. R. Khouzani, S. Sarkar, and S. S. Venkatesh, "Optimal patching in clustered malware epidemics", *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 283–298, 2016.

[60] A. Aurell and B. Djehiche, "Mean-field type modeling of nonlocal crowd aversion in pedestrian crowd dynamics", *SIAM J. Control Optim.*, vol. 56, no. 1, pp. 434–455, 2018.

[61] M. Wu, X. Wang, Y. Yin, and H. Liu, "Leveraging connected and automated vehicles for participatory traffic control", University of Michigan, Center for Connected and Automated Transportation, Tech. Rep., 2023.

[62] A. Aurell, R. Carmona, G. Dayanıklı, and M. Laurière, "Optimal incentives to mitigate epidemics: A Stackelberg mean field game approach", *SIAM J. Control Optim.*, vol. 60, no. 2, S294–S322, 2022.

[63] G. Fu, P. Graewe, U. Horst, and A. Popier, "A mean field game of optimal portfolio liquidation", *Math. Oper. Res.*, vol. 46, no. 4, pp. 1250–1281, 2021.

[64] R. Carmona, "Applications of mean field games in financial engineering and economic theory", *arXiv:2012.05237*, 2020.

[65] B. Djehiche, A. Tcheukam, and H. Tembine, "Mean-field-type games in engineering", *AIMS Electron. Electr. Eng.*, vol. 1, no. 1, pp. 18–73, 2017.

[66] K. Doya, "Reinforcement learning in continuous time and space", *Neural Comput.*, vol. 12, no. 1, pp. 219–245, 2000.

[67] R. Carmona and F. Delarue, *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

[68] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[69] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer Science & Business Media, 2012, vol. 30.

[70] L. Buşoniu, T. De Bruin, D. Tolić, J. Kober, and I. Palunko, "Reinforcement learning for control: Performance, stability, and deep approximators", *Annu. Rev. Control*, vol. 46, pp. 8–28, 2018.

[71] C. J. Watkins and P. Dayan, "Q-Learning", *Mach. Learn.*, vol. 8, pp. 279–292, 1992.

[72] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning", *Mach. Learn.*, vol. 8, pp. 229–256, 1992.

[73] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms", *arXiv:1707.06347*, 2017.

[74] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization", in *Proc. ICML*, 2015, pp. 1889–1897.

[75] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, *et al.*, "Dota 2 with large scale deep reinforcement learning", *arXiv:1912.06680*, 2019.

[76] E. Liang, R. Liaw, R. Nishihara, P. Moritz, R. Fox, K. Goldberg, J. Gonzalez, M. Jordan, and I. Stoica, "RLlib: Abstractions for distributed reinforcement learning", in *Proc. ICML*, 2018, pp. 3053–3062.

[77] D. Fudenberg and J. Tirole, *Game Theory*. MIT press, 1991.

[78] G. W. Brown, "Iterative solution of games by fictitious play", *Act. Anal. Prod. Alloc.*, vol. 13, no. 1, pp. 374–376, 1951.

[79] N. Saldi, T. Başar, and M. Raginsky, "Partially-observed discrete-time risk-sensitive mean-field games", in *Proc. CDC*, 2019, pp. 317–322.

[80] C. Boutilier, "Planning, learning and coordination in multiagent decision processes", in *Proc. 6th Conf. Theor. Asp. Ration. Knowl.*, 1996, pp. 195–210.

[81] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of Markov decision processes", *Math. Oper. Res.*, vol. 12, no. 3, pp. 441–450, 1987.

[82] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains", *Artif. Intell.*, vol. 101, no. 1-2, pp. 99–134, 1998.

[83] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, *et al.*, "Value-decomposition networks for cooperative multi-agent learning based on team reward", in *Proc. AAMAS*, vol. 17, 2018, pp. 2085–2087.

[84] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning", in *Proc. ICML*, 2018, pp. 4295–4304.

[85] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, "Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning", in *Proc. NeurIPS*, vol. 33, 2020, pp. 10 199–10 210.

[86] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents", in *Proc. ICML*, 1993, pp. 330–337.

[87] C. S. de Witt, T. Gupta, D. Makoviichuk, V. Makoviychuk, P. H. Torr, M. Sun, and S. Whiteson, "Is independent learning all you need in the Starcraft multi-agent challenge?", *arXiv:2011.09533*, 2020.

[88] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games", in *Proc. NeurIPS Datasets and Benchmarks*, 2022.

[89] W. Fu, C. Yu, Z. Xu, J. Yang, and Y. Wu, "Revisiting some common practices in cooperative multi-agent reinforcement learning", in *Proc. ICML*, 2022, pp. 6863–6877.

[90] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning", in *Proc. AAMAS*, 2017, pp. 66–83.

[91] S. Hu, Y. Zhong, M. Gao, W. Wang, H. Dong, X. Liang, Z. Li, X. Chang, and Y. Yang, "MARLlib: A scalable and efficient library for multi-agent reinforcement learning", *J. Mach. Learn. Res.*, vol. 24, pp. 1–23, 2023.

[92] G. Qu, A. Wierman, and N. Li, "Scalable reinforcement learning of localized policies for multi-agent networked systems", in *Proc. Learn. Dyn. Control*, 2020, pp. 256–266.

[93] G. Qu, Y. Lin, A. Wierman, and N. Li, "Scalable multi-agent reinforcement learning for networked systems with average reward", in *Proc. NeurIPS*, vol. 33, 2020, pp. 2074–2086.

[94] P. Cardaliaguet and S. Hadikhanloo, "Learning in mean field games: The fictitious play", *ESAIM Control Optim. Calc. Var.*, vol. 23, no. 2, pp. 569–591, 2017.

[95] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games", in *Proc. NeurIPS*, 2019, pp. 4966–4976.

[96] R. Elie, J. Perolat, M. Laurière, M. Geist, and O. Pietquin, "On the convergence of model free learning in mean field games", in *Proc. AAAI*, vol. 34, 2020, pp. 7143–7150.

[97] X. Guo, A. Hu, R. Xu, and J. Zhang, "A general framework for learning mean-field games", *Math. Oper. Res.*, 2022.

[98] Y. Chen, J. Liu, and B. Khoussainov, "Agent-level maximum entropy inverse reinforcement learning for mean field games", *arXiv:2104.14654*, 2021.

[99] P. Lavigne and L. Pfeiffer, "Generalized conditional gradient and learning in potential mean field games", *Applied Mathematics & Optimization*, vol. 88, no. 3, p. 89, 2023.

[100] B. Anahtarci, C. D. Kariksiz, and N. Saldi, "Learning in discrete-time average-cost mean-field games", in *Proc. CDC*, 2021, pp. 3048–3053.

[101] X. Guo, R. Xu, and T. Zariphopoulou, "Entropy regularization for mean field games with learning", *Math. Oper. Res.*, 2022.

[102] J. Arabneydi and A. Mahajan, "Team optimal control of coupled subsystems with mean-field sharing", in *Proc. CDC*, 2014, pp. 1669–1674.

[103] H. Pham and X. Wei, "Bellman equation and viscosity solutions for mean-field stochastic control problem", *ESAIM Control Optim. Calc. Var.*, vol. 24, no. 1, pp. 437–461, 2018.

[104] H. Gu, X. Guo, X. Wei, and R. Xu, "Mean-field controls with Q-learning for cooperative MARL: Convergence and complexity analysis", *SIAM J. Math. Data Sci.*, vol. 3, no. 4, pp. 1168–1196, 2021.

[105] W. U. Mondal, M. Agarwal, V. Aggarwal, and S. V. Ukkusuri, "On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC)", *J. Mach. Learn. Res.*, vol. 23, no. 129, pp. 1–46, 2022.

[106] W. U. Mondal, V. Aggarwal, and S. V. Ukkusuri, "Can mean field control (MFC) approximate cooperative multi agent reinforcement learning (MARL) with non-uniform interaction?", in *Proc. UAI*, PMLR, 2022, pp. 1371–1380.

[107] R. Carmona, M. Laurière, and Z. Tan, "Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning", *Ann. Appl. Probab.*, vol. 33, no. 6B, pp. 5334–5381, 2023.

[108] T. Tanaka, E. Nekouei, A. R. Pedram, and K. H. Johansson, "Linearly solvable mean-field traffic routing games", *IEEE Trans. Automat. Control*, vol. 66, no. 2, pp. 880–887, 2020.

[109] A. F. Hanif, H. Tembine, M. Assaad, and D. Zeghlache, "Mean-field games for resource sharing in cloud-based networks", *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 624–637, 2015.

[110] A. Tcheukam, B. Djehiche, and H. Tembine, "Evacuation of multi-level building: Design, control and strategic flow", in *Proc. CCC*, 2016, pp. 9218–9223.

[111] B. Djehiche, A. Tcheukam, and H. Tembine, "A mean-field game of evacuation in multilevel building", *IEEE Trans. Automat. Control*, vol. 62, no. 10, pp. 5154–5169, 2017.

[112] N. Saldi, T. Başar, and M. Raginsky, "Approximate Nash equilibria in partially observed stochastic games with mean-field interactions", *Math. Oper. Res.*, vol. 44, no. 3, pp. 1006–1033, 2019.

[113] M. Nourian and P. E. Caines, "ϵ-Nash mean field game theory for nonlinear stochastic dynamical systems with major and minor agents", *SIAM J. Control Optim.*, vol. 51, no. 4, pp. 3302–3331, 2013.

[114] D. Mguni, J. Jennings, and E. M. de Cote, "Decentralised learning in systems with many, many strategic agents", in *Proc. AAAI*, vol. 32, 2018, pp. 4686–4693.

[115] J. Subramanian and A. Mahajan, "Reinforcement learning in stationary mean-field games", in *Proc. AAMAS*, vol. 18, 2019, pp. 251–259.

[116] B. Pásztor, A. Krause, and I. Bogunovic, "Efficient model-based multi-agent mean-field reinforcement learning", *Trans. Mach. Learn. Res.*, 2023.

[117] J. Pérolat, S. Perrin, R. Elie, M. Laurière, G. Piliouras, M. Geist, K. Tuyls, and O. Pietquin, "Scaling mean field games by online mirror descent", in *Proc. AAMAS*, 2022, pp. 1028–1037.

[118] S. Perrin, M. Laurière, J. Pérolat, R. Élie, M. Geist, and O. Pietquin, "Generalization in mean field games by learning master policies", in *Proc. AAAI*, vol. 36, 2022, pp. 9413–9421.

[119] J. Yang, X. Ye, R. Trivedi, H. Xu, and H. Zha, "Learning deep mean field games for modeling large population behavior", in *Proc. ICLR*, 2018, pp. 1–15.

[120] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning", in *Proc. ICML*, 2018, pp. 5571–5580.

[121] M. Motte and H. Pham, "Mean-field Markov decision processes with common noise and open-loop controls", *Ann. Appl. Probab.*, vol. 32, no. 2, pp. 1421–1458, 2022.

[122] Y. Achdou and I. Capuzzo-Dolcetta, "Mean field games: Numerical methods", *SIAM J. Numer. Anal.*, vol. 48, no. 3, pp. 1136–1162, 2010.

[123] O. Guéant, J.-M. Lasry, and P.-L. Lions, "Mean field games and applications", in *Paris-Princeton Lectures on Mathematical Finance 2010*, Springer Science & Business Media, 2011, pp. 205–266.

[124] A. Bensoussan, J. Frehse, and P. Yam, *Mean Field Games and Mean Field Type Control Theory*. Springer, 2013, vol. 101.

[125] P. E. Caines, "Mean field games", in *Encyclopedia of Systems and Control*, Springer, 2021, pp. 1197–1202.

[126] L.-P. Chaintron and A. Diez, "Propagation of chaos: A review of models, methods and applications. I. models and methods", *Kinet. Relat. Models*, vol. 15, no. 6, pp. 895–1015, 2022.

[127] S. Perrin, J. Pérolat, M. Laurière, M. Geist, R. Elie, and O. Pietquin, "Fictitious play for mean field games: Continuous time analysis and applications", in *Proc. NeurIPS*, vol. 33, 2020, pp. 13 199–13 213.

[128] Y. Guan, M. Afshari, and P. Tsiotras, "Zero-sum games between mean-field teams: Reachability-based analysis under mean-field sharing", in *Proc. AAAI*, vol. 38, 2024, pp. 9731–9739.

[129] S. Perrin, M. Laurière, J. Pérolat, M. Geist, R. Élie, and O. Pietquin, "Mean field games flock! The reinforcement learning way", in *Proc. IJCAI*, 2021, pp. 356–362.

[130] M. Lauriere, S. Perrin, S. Girgin, P. Muller, A. Jain, T. Cabannes, G. Piliouras, J. Perolat, R. Elie, O. Pietquin, and M. Geist, "Scalable deep reinforcement learning algorithms for mean field games", in *Proc. ICML*, 2022, pp. 12 078–12 095.

[131] X. Guo, A. Hu, and J. Zhang, "MF-OMO: An optimization formulation of mean-field games", *SIAM J. Control Optim.*, vol. 62, no. 1, pp. 243–270, 2024.

[132] X. Guo, A. Hu, M. Santamaria, M. Tajrobehkar, and J. Zhang, "MFGLib: A library for mean-field games", *arXiv:2304.08630*, 2023.

[133] M. Lanctot, E. Lockhart, J.-B. Lespiau, V. Zambaldi, S. Upadhyay, J. Pérolat, S. Srinivasan, F. Timbers, K. Tuyls, S. Omidshafiei, *et al.*, "OpenSpiel: A framework for reinforcement learning in games", *arXiv:1908.09453*, 2019.

[134] J. Huang, B. Yardim, and N. He, "On the statistical efficiency of mean-field reinforcement learning with general function approximation", in *Proc. AISTATS*, PMLR, 2024, pp. 289–297.

[135] B. Anahtarcı, C. D. Karıksız, and N. Saldi, "Value iteration algorithm for mean-field games", *Syst. Control Lett.*, vol. 143, p. 104 744, 2020.

[136] A. C. Kizilkale and R. P. Malhame, "Collective target tracking mean field control for Markovian jump-driven models of electric water heating loads", in *Control of Complex Systems*, Elsevier, 2016, pp. 559–584.

[137] M. Aziz and P. E. Caines, "A mean field game computational methodology for decentralized cellular network optimization", *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 2, pp. 563–576, 2016.

[138] L. Laguzet and G. Turinici, "Individual vaccination as Nash equilibrium in a SIR model with application to the 2009–2010 influenza A (H1N1) epidemic in France", *Bull. Math. Biol.*, vol. 77, no. 10, pp. 1955–1984, 2015.

[139] P. E. Caines and M. Huang, "Graphon mean field games and the GMFG equations: ε-Nash equilibria", in *Proc. CDC*, 2019, pp. 286–292.

[140] A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller, "Maximum a posteriori policy optimisation", in *Proc. ICML*, 2018.

[141] D. A. Gomes, J. Mohr, and R. R. Souza, "Discrete time, finite state space mean field games", *Journal de mathématiques pures et appliquées*, vol. 93, no. 3, pp. 308–328, 2010.

[142] B. Anahtarci, C. D. Kariksiz, and N. Saldi, "Q-learning in regularized mean-field games", *Dyn. Games and Appl.*, pp. 1–29, 2022.

[143] S. Huang, R. F. J. Dossa, C. Ye, J. Braga, D. Chakraborty, K. Mehta, and J. G. Araújo, "CleanRL: High-quality single-file implementations of deep reinforcement learning algorithms", *J. Mach. Learn. Res.*, vol. 23, no. 274, pp. 1–18, 2022.

[144] R. Vizuete, P. Frasca, and F. Garin, "Graphon-based sensitivity analysis of SIS epidemics", *IEEE L-CSS*, vol. 4, no. 3, pp. 542–547, 2020.

[145] F. Parise and A. Ozdaglar, "Graphon games", in *Proc. ACM Conf. Econ. Comput.*, 2019, pp. 457–458.

[146] R. Carmona, D. Cooney, C. Graves, and M. Lauriere, "Stochastic graphon games: I. the static case", *Math. Oper. Res.*, vol. 47, no. 1, pp. 750–778, 2021.

[147] S. Gao and P. E. Caines, "The control of arbitrary size networks of linear systems via graphon limits: An initial investigation", in *Proc. CDC*, 2017, pp. 1052–1057.

[148] S. Gao and P. E. Caines, "Graphon control of large-scale networks of linear systems", *IEEE Trans. Automat. Control*, vol. 65, no. 10, pp. 4090–4105, 2019.

[149] S. Gao and P. E. Caines, "Spectral representations of graphons in very large network systems control", in *Proc. CDC*, 2019, pp. 5068–5075.

[150] E. Bayraktar, S. Chakraborty, and R. Wu, "Graphon mean field systems", *Ann. Appl. Probab.*, vol. 33, no. 5, pp. 3587–3619, 2023.

[151] G. Bet, F. Coppini, and F. R. Nardi, "Weakly interacting oscillators on dense random graphs", *arXiv:2006.07670*, 2020.

[152] A. Aurell, R. Carmona, and M. Lauriere, "Stochastic graphon games: II. the linear-quadratic case", *Applied Mathematics & Optimization*, vol. 85, no. 3, p. 39, 2022.

[153] A. Aurell, R. Carmona, G. Dayanıklı, and M. Laurière, "Finite state graphon games with applications to epidemics", *Dyn. Games and Appl.*, vol. 12, no. 1, pp. 49–81, 2022.

[154] D. Vasal, R. Mishra, and S. Vishwanath, "Sequential decomposition of graphon mean field games", in *Proc. ACC*, 2021, pp. 730–736.

[155] M. A. Gkogkas and C. Kuehn, "Graphop mean-field limits for Kuramoto-type models", *SIAM J. Appl. Dyn. Syst.*, vol. 21, no. 1, pp. 248–283, 2022.

[156] D. Lacker and A. Soret, "A case study on stochastic games on large graphs in mean field and sparse regimes", *Math. Oper. Res.*, vol. 47, no. 2, pp. 1530–1565, 2021.

[157] L. Lovász, *Large Networks and Graph Limits*. Am. Math. Soc., 2012, vol. 60.

[158] C. Borgs, J. Chayes, L. Lovász, V. Sós, and K. Vesztergombi, "Limits of randomly grown graph sequences", *Eur. J. Comb.*, vol. 32, no. 7, pp. 985–999, 2011.

[159] G. Carmona, "Nash equilibria of games with a continuum of players", Universidade Nova de Lisboa, Nova School of Business and Economics, Tech. Rep., 2004.

[160] J. Xu, "Rates of convergence of spectral methods for graphon estimation", in *Proc. ICML*, 2018, pp. 5433–5442.

[161] G. Y. Weintraub, C. L. Benkard, and B. Van Roy, "Computational methods for oblivious equilibrium", *Oper. Res.*, vol. 58, no. 4-part-2, pp. 1247–1265, 2010.

[162] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor", in *Proc. ICML*, 2018, pp. 1861–1870.

[163] D. Glasscock, "What is... a graphon?", *Not. AMS*, vol. 62, no. 1, 2015.

[164] P. E. Caines and M. Huang, "Graphon mean field games and the GMFG equations", in *Proc. CDC*, 2018, pp. 4129–4134.

[165] Á. Bodó, G. Y. Katona, and P. L. Simon, "SIS epidemic propagation on hypergraphs", *Bull. Math. Biol.*, vol. 78, no. 4, pp. 713–735, 2016.

[166] N. W. Landry and J. G. Restrepo, "The effect of heterogeneity on hypergraph contagion models", *Chaos*, vol. 30, no. 10, p. 103 117, 2020.

[167] D. J. Higham and H.-L. de Kergorlay, "Mean field analysis of hypergraph contagion models", *SIAM J. Appl. Math.*, vol. 82, no. 6, pp. 1987–2007, 2022.

[168] J. Noonan and R. Lambiotte, "Dynamics of majority rule on hypergraphs", *Phys. Rev. E*, vol. 104, no. 2, p. 024 316, 2021.

[169] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks", *J. Complex Netw.*, vol. 2, no. 3, pp. 203–271, 2014.

[170] K. A. Jacobsen, M. G. Burch, J. H. Tien, and G. A. Rempała, "The large graph limit of a stochastic epidemic model on a dynamic multilayer network", *J. Biol. Dyn.*, vol. 12, no. 1, pp. 746–788, 2018.

[171] G. Elek and B. Szegedy, "A measure-theoretic approach to the theory of dense hypergraphs", *Adv. Math.*, vol. 231, no. 3-4, pp. 1731–1772, 2012.

[172] C. Borgs, J. T. Chayes, H. Cohn, and Y. Zhao, "An L^p theory of sparse graph convergence II: LD convergence, quotients and right convergence", *Ann. Probab.*, vol. 46, no. 1, pp. 337–396, 2018.

[173] C. Borgs, J. Chayes, H. Cohn, and Y. Zhao, "An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions", *Trans. Am. Math. Soc.*, vol. 372, no. 5, pp. 3019–3062, 2019.

[174] Y. Zhao, "Hypergraph limits: A regularity approach", *Random Structures & Algorithms*, vol. 47, no. 2, pp. 205–226, 2015.

[175] D. Maki and M. Thompson, *Mathematical Models and Applications: With EmphaSIS on the Social, Life, and Management Sciences*. Prentice-Hall, 1973.

[176] V. V. Junior, P. M. Rodriguez, and A. Speroto, "The maki-thompson rumor model on infinite cayley trees", *J. Stat. Phys.*, vol. 181, no. 4, pp. 1204–1217, 2020.

[177] F. Garbe, J. Hladkỳ, M. Šileikis, and F. Skerman, "From flip processes to dynamical systems on graphons", *arXiv:2201.12272*, 2022.

[178] R. Carmona, F. Delarue, and D. Lacker, "Mean field games with common noise", *Ann. Probab.*, vol. 44, no. 6, pp. 3740–3803, 2016.

[179] J. Huang and S. Wang, "Dynamic optimization of large-population systems with partial information", *J. Optim. Theory Appl.*, vol. 168, pp. 231–245, 2016.

[180] C. Harris, "On the rate of convergence of continuous-time fictitious play", *Games Econ. Behav.*, vol. 22, no. 2, pp. 238–259, 1998.

[181] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play", *Econometrica*, vol. 70, no. 6, pp. 2265–2294, 2002.

[182] M. Huang, "Large-population LQG games involving a major player: The Nash certainty equivalence principle", *SIAM J. Control Optim.*, vol. 48, no. 5, pp. 3318–3353, 2010.

[183] S. L. Nguyen and M. Huang, "Linear-quadratic-gaussian mixed games with continuum-parametrized minor players", *SIAM J. Control Optim.*, vol. 50, no. 5, pp. 2907–2937, 2012.

[184] A. Bensoussan, M. H. Chau, and S. C. Yam, "Mean field games with a dominating player", *Appl. Math. Optim.*, vol. 74, pp. 91–128, 2016.

[185] N. Şen and P. E. Caines, "Mean field game theory with a partially observed major agent", *SIAM J. Control Optim.*, vol. 54, no. 6, pp. 3174–3224, 2016.

[186] R. A. Carmona and X. Zhu, "A probabilistic approach to mean field games with major and minor players", *Ann. Appl. Probab.*, vol. 26, no. 3, pp. 1535–1580, 2016.

[187] R. Carmona and P. Wang, "An alternative approach to mean field game with major and minor players, and applications to herders impacts", *Appl. Math. Optim.*, vol. 76, pp. 5–27, 2017.

[188] J.-M. Lasry and P.-L. Lions, "Mean-field games with a major player", *Comptes Rendus Mathematique*, vol. 356, no. 8, pp. 886–890, 2018.

[189] P. Cardaliaguet, M. Cirant, and A. Porretta, "Remarks on Nash equilibria in mean field game models with a major player", *Proc. Am. Math. Soc.*, vol. 148, no. 10, pp. 4241–4255, 2020.

[190] R. Elie, T. Mastrolia, and D. Possamaï, "A tale of a principal and many, many agents", *Math. Oper. Res.*, vol. 44, no. 2, pp. 440–467, 2019.

[191] R. Carmona and P. Wang, "Finite-state contract theory with a principal and a field of agents", *Manag. Sci.*, vol. 67, no. 8, pp. 4725–4741, 2021.

[192] R. Carmona, G. Dayanıklı, and M. Laurière, "Mean field models to regulate carbon emissions in electricity production", *Dyn. Games Appl.*, vol. 12, no. 3, pp. 897–928, 2022.

[193] X. Guo, A. Hu, and J. Zhang, "Optimization frameworks and sensitivity analysis of Stackelberg mean-field games", *arXiv:2210.04110*, 2022.

[194] D. Vasal and R. Berry, "Master equation for discrete-time Stackelberg mean field games with a single leader", in *Proc. CDC*, 2022, pp. 5529–5535.

[195] M. Huang, P. E. Caines, and R. P. Malhamé, "Distributed multi-agent decision-making with partial observations: Asymptotic Nash equilibria", in *Proc. 17th Internat. Symp. MTNS*, 2006, pp. 2725–2730.

[196] R. Carmona and G. Dayanıklı, "Mean field game model for an advertising competition in a duopoly", *Int. Game Theory Rev.*, vol. 23, no. 04, p. 2 150 024, 2021.

[197] D. Andersson and B. Djehiche, "A maximum principle for SDEs of mean-field type", *Appl. Math. Optim.*, vol. 63, no. 3, pp. 341–356, 2011.

[198] B. Djehiche and H. Tembine, "Risk-sensitive mean-field type control under partial observation", in *Stochastics of Environmental and Financial Economics*, Springer, Cham, 2016, pp. 243–263.

[199] B. Djehiche, H. Tembine, and R. Tempone, "A stochastic maximum principle for risk-sensitive mean-field type control", *IEEE Trans. Automat. Control*, vol. 60, no. 10, pp. 2640–2649, 2015.

[200] M. F. Djete, D. Possamaï, and X. Tan, "McKean–Vlasov optimal control: The dynamic programming principle", *Ann. Probab.*, vol. 50, no. 2, pp. 791–833, 2022.

[201] H. Gu, X. Guo, X. Wei, and R. Xu, "Dynamic programming principles for mean-field controls with learning", *Oper. Res.*, 2023.

[202] P. E. Caines and A. C. Kizilkale, "ϵ-Nash equilibria for partially observed LQG mean field games with a major player", *IEEE Trans. Automat. Control*, vol. 62, no. 7, pp. 3225–3234, 2016.

[203] A. Mukhopadhyay and R. R. Mazumdar, "Analysis of randomized Join-the-Shortest-Queue (JSQ) schemes in large heterogeneous processor-sharing systems", *IEEE Trans. Control Netw. Syst.*, vol. 3, no. 2, pp. 116–126, 2016.

[204] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[205] M. Nourian, P. E. Caines, R. P. Malhame, and M. Huang, "Nash, social and centralized solutions to consensus problems via mean field control theory", *IEEE Trans. Automat. Control*, vol. 58, no. 3, pp. 639–653, 2012.

[206] D. F. Anderson and T. G. Kurtz, "Continuous time Markov chain models for chemical reaction networks", in *Design and Analysis of Biomolecular Circuits*, Springer, 2011, pp. 3–42.

[207] A. Dunyak and P. E. Caines, "Large scale systems and SIR models: A featured graphon approach", in *Proc. CDC*, 2021, pp. 6928–6933.

[208] H. Shiri, J. Park, and M. Bennis, "Massive autonomous UAV path planning: A neural network based mean-field game theoretic approach", in *Proc. GLOBECOM*, 2019, pp. 1–6.

[209] R. Carmona, Q. Cormier, and H. M. Soner, "Synchronization in a Kuramoto mean field game", *Communications in Partial Differential Equations*, vol. 48, no. 9, pp. 1214–1244, 2023.

[210] N. Bäuerle, "Mean field Markov decision processes", *Appl. Math. Optim.*, vol. 88, no. 1, p. 12, 2023.

[211] M. Motte and H. Pham, "Quantitative propagation of chaos for mean field Markov decision process with common noise", *Electronic Journal of Probability*, vol. 28, pp. 1–24, 2023.

[212] W. U. Mondal, V. Aggarwal, and S. Ukkusuri, "Mean-field control based approximation of multi-agent reinforcement learning in presence of a non-decomposable shared global state", *Trans. Mach. Learn. Res.*, 2023.

[213] N. Şen and P. E. Caines, "Mean field games with partially observed major player and stochastic mean field", in *Proc. CDC*, 2014, pp. 2709–2715.

[214] C. Villani, *Optimal Transport: Old and New*. Springer, 2009, vol. 338.

[215] P. Billingsley, *Convergence of Probability Measures*. John Wiley & Sons, 2013.

[216] K. R. Parthasarathy, *Probability Measures on Metric Spaces*. American Mathematical Soc., 2005, vol. 352.

[217] O. Kallenberg, *Random Measures, Theory and Applications*. Springer, 2017, vol. 1.

[218] A.-S. Sznitman, "Topics in propagation of chaos", in *Ecole d'été de probabilités de Saint-Flour XIX—1989*, Springer, 1991, pp. 165–251.

[219] C. Herrera, F. Krach, and J. Teichmann, "Local Lipschitz bounds of deep neural networks", *arXiv:2004.13135*, 2023.

[220] A. Araujo, A. J. Havens, B. Delattre, A. Allauzen, and B. Hu, "A unified algebraic perspective on Lipschitz neural networks", in *Proc. ICLR*, 2023, pp. 1–15.

[221] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation", in *Proc. NIPS*, 1999, pp. 1057–1063.

[222] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, "POT: Python Optimal Transport", *J. Mach. Learn. Res.*, vol. 22, no. 78, pp. 1–8, 2021.

[223] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym", *arXiv:1606.01540*, 2016.

[224] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage", *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1647–1650, 2016.

[225] Y. Yang, Y. Xiao, and T. Li, "Attacks on formation control for multiagent systems", *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 12 805–12 817, 2021.

[226] W. B. Arthur, "Inductive reasoning and bounded rationality", *Am. Econ. Rev.*, vol. 84, no. 2, pp. 406–411, 1994.

[227] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, "Swarm robotics: A review from the swarm engineering perspective", *Swarm Intell.*, vol. 7, no. 1, pp. 1–41, 2013.

[228] M. Marinelli, L. Caggiani, M. Ottomanelli, and M. Dell'Orco, "En route truck–drone parcel delivery for optimal vehicle routing strategies", *IET Intell. Transp. Syst.*, vol. 12, no. 4, pp. 253–261, 2018.

[229] D. Jin and L. Zhang, "Collective behaviors of magnetic active matter: Recent progress toward reconfigurable, adaptive, and multifunctional swarming micro/nanorobots", *Acc. Chem. Res.*, vol. 55, no. 1, pp. 98–109, 2021.

[230] N. Gast and B. Van Houdt, "A refined mean field approximation", *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 46, no. 1, pp. 113–113, 2018.

[231] T. Vicsek and A. Zafeiris, "Collective motion", *Phys. Rep.*, vol. 517, no. 3-4, pp. 71–140, 2012.

[232] C.-E. Hrabia, M. Lützenberger, and S. Albayrak, "Towards adaptive multi-robot systems: Self-organization and self-adaptation", *Knowl. Eng. Rev.*, vol. 33, e16, 2018.

[233] M. Schranz, G. A. Di Caro, T. Schmickl, W. Elmenreich, F. Arvin, A. Şekercioğlu, and M. Sende, "Swarm intelligence and cyber-physical systems: Concepts, challenges and future trends", *Swarm Evol. Comput.*, vol. 60, p. 100 762, 2021.

[234] Q. Zha, G. Kou, H. Zhang, H. Liang, X. Chen, C.-C. Li, and Y. Dong, "Opinion dynamics in finance and business: A literature review and research opportunities", *Financ. Innov.*, vol. 6, pp. 1–22, 2020.

[235] P. Yin, H. M. Choi, C. R. Calvert, and N. A. Pierce, "Programming biomolecular self-assembly pathways", *Nature*, vol. 451, no. 7176, pp. 318–322, 2008.

[236] F. Cichos, K. Gustavsson, B. Mehlig, and G. Volpe, "Machine learning for active matter", *Nat. Mach. Intell.*, vol. 2, no. 2, pp. 94–103, 2020.

[237] N. Kruk, J. A. Carrillo, and H. Koeppl, "Traveling bands, clouds, and vortices of chiral active matter", *Phys. Rev. E*, vol. 102, no. 2, p. 022 604, 2020.

[238] M. Nasiri and B. Liebchen, "Reinforcement learning of optimal active particle navigation", *New J. Phys.*, vol. 24, no. 7, p. 073 042, 2022.

[239] N. Narinder, C. Bechinger, and J. R. Gomez-Solano, "Memory-induced transition from a persistent random walk to circular motion for achiral microswimmers", *Phys. Rev. Lett.*, vol. 121, no. 7, p. 078 003, 2018.

[240] K. Zhang, Z. Yang, and T. Başar, "Decentralized multi-agent reinforcement learning with networked agents: Recent advances", *Front. Inf. Technol. Electron. Eng.*, vol. 22, no. 6, pp. 802–814, 2021.

[241] S. Ganapathi Subramanian, M. E. Taylor, M. Crowley, and P. Poupart, "Partially observable mean field reinforcement learning", in *Proc. AAMAS*, vol. 20, 2021, pp. 537–545.

[242] S. G. Subramanian, M. E. Taylor, M. Crowley, and P. Poupart, "Decentralized mean field games", in *Proc. AAAI*, vol. 36, 2022, pp. 9439–9447.

[243] N. Şen and P. E. Caines, "Mean field games with partial observation", *SIAM J. Control Optim.*, vol. 57, no. 3, pp. 2064–2091, 2019.

[244] T. Tottori and T. J. Kobayashi, "Memory-limited partially observable stochastic control and its mean-field control approach", *Entropy*, vol. 24, no. 11, p. 1599, 2022.

[245] W. Wang, J. Han, Z. Yang, and Z. Wang, "Global convergence of policy gradient for linear-quadratic mean-field control/game in continuous time", in *Proc. ICML*, 2021, pp. 10 772–10 782.

[246] D. Waelchli, P. Weber, and P. Koumoutsakos, "Discovering individual rewards in collective behavior through inverse multi-agent reinforcement learning", *arXiv:2305.10548*, 2023.

[247] M. Kwon, J. Agapiou, E. Duéñez-Guzmán, R. Elie, G. Piliouras, K. Bullard, and I. Gemp, "Auto-aligning multiagent incentives with global objectives", in *ALA Workshop, AAMAS*, 2023, pp. 1–9.

[248] Z. Fu, Z. Yang, Y. Chen, and Z. Wang, "Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games", in *Proc. ICLR*, 2019, pp. 1–82.

[249] R. Carmona, M. Laurière, and Z. Tan, "Linear-quadratic mean-field reinforcement learning: Convergence of policy gradient methods", *arXiv:1910.04295*, 2019.

[250] S. Witwicki and E. Durfee, "Influence-based policy abstraction for weakly-coupled Dec-POMDPs", in *Proc. ICAPS*, vol. 20, 2010, pp. 185–192.

[251] H. Hamann, *Swarm Robotics: A Formal Approach*. Springer, 2018.

[252] K. J. Åström, "Optimal control of Markov processes with incomplete state information", *J. Math. Anal. Appl.*, vol. 10, no. 1, pp. 174–205, 1965.

[253] O. Kallenberg, *Foundations of Modern Probability*. Springer, 2021.

[254] A. Mahajan, N. C. Martins, M. C. Rotkowitz, and S. Yüksel, "Information structures in optimal decentralized control", in *Proc. CDC*, 2012, pp. 1291–1306.

[255] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks", *arXiv:1609.09106*, 2016.

[256] R. Miculescu, "Approximation of continuous functions by Lipschitz functions", *Real Anal. Exch.*, pp. 449–452, 2000.

[257] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms", in *Proc. ICML*, 2014, pp. 387–395.

[258] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning.", in *Proc. ICLR*, 2016, pp. 1–14.

[259] J. Peters and S. Schaal, "Natural actor-critic", *Neurocomputing*, vol. 71, no. 7-9, pp. 1180–1190, 2008.

[260] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation", in *Proc. ICLR*, 2016, pp. 1–14.

[261] O. Soysal and E. Sahin, "Probabilistic aggregation strategies in swarm robotic systems", in *IEEE Swarm Intell. Symp.*, 2005, pp. 325–332.

[262] E. Bahgeçi and E. Sahin, "Evolving aggregation behaviors for swarm robotic systems: A systematic case study", in *IEEE Swarm Intell. Symp.*, 2005, pp. 333–340.

[263] J. A. Acebrón, L. L. Bonilla, C. J. Pérez, F. Ritort, and R. Spigler, "The Kuramoto model: A simple paradigm for synchronization phenomena", *Rev. Mod. Phys.*, vol. 77, no. 1, pp. 137–185, 2005.

[264] M. Breakspear, S. Heitmann, and A. Daffertshofer, "Generative models of cortical oscillations: Neurobiological implications of the Kuramoto model", *Front. Hum. Neurosci.*, vol. 4, p. 190, 2010.

[265] A. Diaz-Guilera, J. Gómez-Gardenes, Y. Moreno, and M. Nekovee, "Synchronization in random geometric graphs", *Int. J. Bifurcation Chaos*, vol. 19, no. 02, pp. 687–693, 2009.

[266] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet, "Novel type of phase transition in a system of self-driven particles", *Phys. Rev. Lett.*, vol. 75, no. 6, p. 1226, 1995.

[267] L. Zheng, J. Yang, H. Cai, W. Zhang, J. Wang, and Y. Yu, "MAgent: A many-agent reinforcement learning platform for artificial collective intelligence", in *Proc. AAAI*, 2018, pp. 8222–8223.

[268] G. Dayanikli, M. Laurière, and J. Zhang, "Deep learning for population-dependent controls in mean field control problems", *arXiv:2306.04788*, 2023.

[269] J. Huang, B. Yardim, and N. He, "On the statistical efficiency of mean field reinforcement learning with general function approximation", *arXiv:2305.11283*, 2023.

[270] P. Li, X. Wang, S. Li, H. Chan, and B. An, "Population-size-aware policy optimization for mean-field games", in *Proc. ICLR*, 2023, pp. 1–32.

[271] H. X. Pham, H. M. La, D. Feil-Seifer, and A. Nefian, "Cooperative and distributed reinforcement learning of drones for field coverage", *arXiv:1803.07250*, 2018.

[272] F. Bagagiolo and D. Bauso, "Mean-field games and dynamic demand management in power grids", *Dyn. Games Appl.*, vol. 4, no. 2, pp. 155–176, 2014.

[273] R. A. Banez, L. Li, C. Yang, L. Song, and Z. Han, "A mean-field-type game approach to computation offloading in mobile edge computing networks", in *Proc. ICC*, 2019, pp. 1–6.

[274] L. Horstmeyer and C. Kuehn, "Adaptive voter model on simplicial complexes", *Phys. Rev. E*, vol. 101, no. 2, p. 022 305, 2020.

[275] I. Iacopini, G. Petri, A. Barrat, and V. Latora, "Simplicial models of social contagion", *Nature Commun.*, vol. 10, no. 1, pp. 1–9, 2019.

[276] X.-J. Xu, S. He, and L.-J. Zhang, "Dynamics of the threshold model on hypergraphs", *Chaos*, vol. 32, no. 2, p. 023 125, 2022.

[277] P. S. Skardal, L. Arola-Fernández, D. Taylor, and A. Arenas, "Higher-order interactions can better optimize network synchronization", *Phys. Rev. Res.*, vol. 3, p. 043 193, 4 Dec. 2021.

[278] M. S. Anwar and D. Ghosh, "Intralayer and interlayer synchronization in multiplex network with higher-order interactions", *Chaos*, vol. 32, no. 3, p. 033 125, 2022.

[279] C. Ziegler, P. S. Skardal, H. Dutta, and D. Taylor, "Balanced Hodge Laplacians optimize consensus dynamics over simplicial complexes", *Chaos*, vol. 32, no. 2, p. 023 128, 2022.

[280] M. A. Porter, "Nonlinearity+ networks: A 2020 vision", in *Emerging Frontiers in Nonlinear Science*, Springer, 2020, pp. 131–159.

[281] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, "Networks beyond pairwise interactions: Structure and dynamics", *Phys. Rep.*, vol. 874, pp. 1–92, 2020.

[282] C. Bick, E. Gross, H. A. Harrington, and M. T. Schaub, "What are higher-order networks?", *SIAM Rev.*, vol. 65, no. 3, pp. 686–731, 2023.

[283] A. Angiuli, J.-P. Fouque, and M. Lauriere, "Reinforcement learning for mean field games, with applications to economics", *arXiv:2106.13755*, 2021.

[284] N. Brown and T. Sandholm, "Superhuman AI for multiplayer poker", *Science*, vol. 365, no. 6456, pp. 885–890, 2019.

[285] J. Suarez, Y. Du, I. Mordach, and P. Isola, "Neural MMO v1. 3: A massively multiagent game environment for training and evaluating neural networks", in *Proc. AAMAS*, vol. 19, 2020, pp. 2020–2022.

[286] P. Golle, K. Leyton-Brown, I. Mironov, and M. Lillibridge, "Incentives for sharing in peer-to-peer networks", in *Proc. Int. Workshop Electron. Commerce*, 2001, pp. 75–87.

[287] Z. Jiang and J. Liang, "Cryptocurrency portfolio management with deep reinforcement learning", in *Proc. IntelliSys*, 2017, pp. 905–913.

[288] A. Schaerf, Y. Shoham, and M. Tennenholtz, "Adaptive load balancing: A study in multi-agent learning", *J. Artif. Intell. Res.*, vol. 2, pp. 475–500, 1994.

[289] D. Lipshutz, "Open problem—load balancing using delayed information", *Stoch. Syst.*, vol. 9, no. 3, pp. 305–306, 2019.

[290] R. Zheng, H. Wang, and M. De Mari, "Optimal computation offloading with a shared mec center: A mean field game approach", in *Proc. GLOBECOM Workshops*, 2019, pp. 1–6.

[291] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics", *J. Intell. Robot. Syst.*, vol. 86, no. 2, pp. 153–173, 2017.

[292] I. Kovalev, A. Voroshilova, and M. Karaseva, "Analysis of the current situation and development trend of the international cargo UAVs market", in *Proc. J. Phys.: Conf. Ser.*, IOP Publishing, vol. 1399, 2019, p. 055 095.

[293] D. Câmara, "Cavalry to the rescue: Drones fleet to help rescuers operations over disasters scenarios", in *Proc. IEEE CAMA*, 2014, pp. 1–4.

[294] Y. Karaca, M. Cicek, O. Tatli, A. Sahin, S. Pasli, M. F. Beser, and S. Turedi, "The potential use of unmanned aircraft systems (drones) in mountain search and rescue operations", *Am. J. Emerg. Med.*, vol. 36, no. 4, pp. 583–588, 2018.

[295] H. Shakhatreh, A. H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N. S. Othman, A. Khreishah, and M. Guizani, "Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges", *IEEE Access*, vol. 7, pp. 48 572–48 634, 2019.

[296] G. Chmaj and H. Selvaraj, "Distributed processing applications for UAV/drones: A survey", in *Progress in Systems Engineering*, Springer, 2015, pp. 449–454.

[297] Z. Zhang, D. Zhang, and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview", *CSEE J. Power Energy Syst.*, vol. 6, no. 1, pp. 213–225, 2019.

[298] A. Castelletti, G. Corani, A. Rizzolli, R. Soncinie-Sessa, and E. Weber, "Reinforcement learning in the operational management of a water system", in *Proc. IFAC Workshop on Model. Contr. in Environmental Issues*, 2002, pp. 325–330.

[299] S. Brandi, M. S. Piscitelli, M. Martellacci, and A. Capozzoli, "Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings", *Energy Build.*, vol. 224, p. 110 225, 2020.

[300] Z. Ma, D. S. Callaway, and I. A. Hiskens, "Decentralized charging control of large populations of plug-in electric vehicles", *IEEE Trans. Contr. Syst. Technol.*, vol. 21, no. 1, pp. 67–78, 2011.

[301] Z. Ma, N. Yang, S. Zou, and Y. Shao, "Charging coordination of plug-in electric vehicles in distribution networks with capacity constrained feeder lines", *IEEE Trans. Contr. Syst. Technol.*, vol. 26, no. 5, pp. 1917–1924, 2017.

[302] Y. Wang, C. Chen, J. Wang, and R. Baldick, "Research on resilience of power systems under natural disasters—a review", *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1604–1613, 2015.

[303] D. Troullinos, G. Chalkiadakis, I. Papamichail, and M. Papageorgiou, "Collaborative multiagent decision making for lane-free autonomous driving", in *Proc. AAMAS*, vol. 20, 2021, pp. 1335–1343.

[304] C. Yu, X. Wang, X. Xu, M. Zhang, H. Ge, J. Ren, L. Sun, B. Chen, and G. Tan, "Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs", *IEEE Trans. Transp. Syst.*, vol. 21, no. 2, pp. 735–748, 2019.

[305] C. Wu, K. Kumekawa, and T. Kato, "Distributed reinforcement learning approach for vehicular ad hoc networks", *IEICE Trans. Commun.*, vol. 93, no. 6, pp. 1431–1442, 2010.

[306] A. Tampuu, T. Matiisen, M. Semikin, D. Fishman, and N. Muhammad, "A survey of end-to-end driving: Architectures and training methods", *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1364–1384, 2020.

[307] L. M. Schmidt, J. Brosig, A. Plinge, B. M. Eskofier, and C. Mutschler, "An introduction to multi-agent reinforcement learning and review of its application to autonomous mobility", *arXiv:2203.07676*, 2022.

[308] K. Braekers, K. Ramaekers, and I. Van Nieuwenhuyse, "The vehicle routing problem: State of the art classification and review", *Comput. Ind. Eng.*, vol. 99, pp. 300–313, 2016.

[309] J. S. Juul and M. A. Porter, "Hipsters on networks: How a minority group of individuals can lead to an antiestablishment majority", *Phys. Rev. E*, vol. 99, p. 022 313, 2 Feb. 2019.

[310] R. van der Hofstad, *Random Graphs and Complex Networks: Volume 1*. USA: Cambridge University Press, 2016.

[311] M. Gribaudo, D. Cerotti, and A. Bobbio, "Analysis of on-off policies in sensor networks using interacting Markovian agents", in *Proc. PerCom*, 2008, pp. 300–305.

[312] P. Van Mieghem, J. Omic, and R. Kooij, "Virus spread in networks", *IEEE/ACM Trans. Netw.*, vol. 17, no. 1, pp. 1–14, 2009.

[313] Y. Xie, Z. Wang, J. Lu, and Y. Li, "Stability analysis and control strategies for a new SIS epidemic model in heterogeneous networks", *Appl. Math. Comput.*, vol. 383, p. 125 381, 2020.

[314] K. Tran and G. Yin, "Optimal control and numerical methods for hybrid stochastic SIS models", *Nonlinear Analysis: Hybrid Systems*, vol. 41, p. 101 051, 2021.

[315] Z. Abbasi, I. Zamani, A. H. A. Mehra, M. Shafieirad, and A. Ibeas, "Optimal control design of impulsive SQEIAR epidemic models with application to COVID-19", *Chaos, Solitons & Fractals*, vol. 139, p. 110 054, 2020.

[316] C. Liu, "A microscopic epidemic model and pandemic prediction using multi-agent reinforcement learning", *arXiv:2004.12959*, 2020.

[317] A. Q. Ohi, M. Mridha, M. M. Monowar, M. Hamid, *et al.*, "Exploring optimal control of epidemic spread using reinforcement learning", *Sci. Rep.*, vol. 10, no. 1, pp. 1–19, 2020.

[318] V. Kompella, R. Capobianco, S. Jong, J. Browne, S. Fox, L. Meyers, P. Wurman, and P. Stone, "Reinforcement learning for optimization of COVID-19 mitigation policies", *arXiv:2010.10560*, 2020.

[319] R. Capobianco, V. Kompella, J. Ault, G. Sharon, S. Jong, S. Fox, L. Meyers, P. R. Wurman, and P. Stone, "Agent-based Markov modeling for improved COVID-19 mitigation policies", *J. Artif. Intell. Res.*, vol. 71, pp. 953–992, 2021.

[320] M. Schranz, M. Umlauft, M. Sende, and W. Elmenreich, "Swarm robotic behaviors and current applications", *Front. Robot. AI*, vol. 7, p. 36, 2020.

[321] S.-J. Chung, A. A. Paranjape, P. Dames, S. Shen, and V. Kumar, "A survey on aerial swarm robotics", *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 837–855, 2018.

[322] M. Dorigo, G. Theraulaz, and V. Trianni, "Reflections on the future of swarm robotics", *Sci. Robot.*, vol. 5, no. 49, eabe4385, 2020.

[323] N. Correll and A. Martinoli, "System identification of self-organizing robotic swarms", in *Distributed Autonomous Robotic Systems 7*, Springer, 2006, pp. 31–40.

[324] E. Tuci, M. H. Alkilabi, and O. Akanyeti, "Cooperative object transport in multi-robot systems: A review of the state-of-the-art", *Front. Robot. AI*, vol. 5, p. 59, 2018.

[325] R. Gross and M. Dorigo, "Evolution of solitary and group transport behaviors for autonomous robots capable of self-assembling", *Adapt. Behav.*, vol. 16, no. 5, pp. 285–305, 2008.

[326] R. Gross and M. Dorigo, "Towards group transport by swarms of robots", *Int. J. Bio-Inspired Comput.*, vol. 1, no. 1/2, pp. 1–13, 2009.

[327] D. Albani, T. Manoni, D. Nardi, and V. Trianni, "Dynamic UAV swarm deployment for non-uniform coverage", in *Proc. AAMAS*, 2018, pp. 523–531.

[328] D. Xing, Z. Zhen, and H. Gong, "Offense–defense confrontation decision making for dynamic UAV swarm versus UAV swarm", *Proc. Inst. Mech. Eng. G*, vol. 233, no. 15, pp. 5689–5702, 2019.

[329] P. Vincent and I. Rubin, "A framework and analysis for cooperative search using UAV swarms", in *Proc. ACM Symp. Appl. Comput.*, 2004, pp. 79–86.

[330] H. A. Al-Rawi, M. A. Ng, and K.-L. A. Yau, "Application of reinforcement learning to routing in distributed wireless networks: A review", *Artif. Intell. Rev.*, vol. 43, no. 3, pp. 381–416, 2015.

[331] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control", *IEEE Trans. Transp. Syst.*, vol. 21, no. 3, pp. 1086–1095, 2019.

[332] K. Elamvazhuthi and S. Berman, "Mean-field models in swarm robotics: A survey", *Bioinspir. Biomim.*, vol. 15, no. 1, p. 015 001, 2019.

[333] K. Elamvazhuthi, M. Kawski, S. Biswal, V. Deshmukh, and S. Berman, "Mean-field controllability and decentralized stabilization of Markov chains", in *Proc. CDC*, 2017, pp. 3131–3137.

[334] V. Deshmukh, K. Elamvazhuthi, S. Biswal, Z. Kakish, and S. Berman, "Mean-field stabilization of Markov chain models for robotic swarms: Computational approaches and experimental results", *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1985–1992, 2018.

[335] K. Elamvazhuthi, S. Biswal, and S. Berman, "Mean-field stabilization of robotic swarms to probability distributions with disconnected supports", in *Proc. ACC*, 2018, pp. 885–892.

[336] S. Mayya, P. Pierpaoli, G. Nair, and M. Egerstedt, "Localization in densely packed swarms using interrobot collisions as a sensing modality", *IEEE Trans. Robot.*, vol. 35, no. 1, pp. 21–34, 2018.

[337] S. Mayya, S. Wilson, and M. Egerstedt, "Closed-loop task allocation in robot swarms using inter-robot encounters", *Swarm Intell.*, vol. 13, no. 2, pp. 115–143, 2019.

[338] K. Lerman, A. Galstyan, A. Martinoli, and A. Ijspeert, "A macroscopic analytical model of collaboration in distributed robotic systems", *Artif. Life*, vol. 7, no. 4, pp. 375–393, 2001.

[339] U. Eren and B. Açıkmeşe, "Velocity field generation for density control of swarms using heat equation and smoothing kernels", *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 9405–9411, 2017.

[340] T. Zheng, Q. Han, and H. Lin, "Transporting robotic swarms via mean-field feedback control", *IEEE Trans. Automat. Control*, vol. 67, no. 8, pp. 4170–4177, 2021.

[341] D. Milutinović and P. Lima, "Modeling and optimal centralized control of a large-size robotic population", *IEEE Trans. Robot.*, vol. 22, no. 6, pp. 1280–1285, 2006.

[342] H. Hamann and H. Wörn, "A framework of space–time continuous models for algorithm design in swarm robotics", *Swarm Intell.*, vol. 2, no. 2, pp. 209–239, 2008.

[343] L. Campi and M. Fischer, "Correlated equilibria and mean field games: A simple model", *Math. Oper. Res.*, 2022.

[344] P. Muller, R. Elie, M. Rowland, M. Lauriere, J. Perolat, S. Perrin, M. Geist, G. Piliouras, O. Pietquin, and K. Tuyls, "Learning correlated equilibria in mean-field games", *arXiv:2208.10138*, 2022.

[345] H. Gao, W. Lee, Y. Kang, W. Li, Z. Han, S. Osher, and H. V. Poor, "Energy-efficient velocity control for massive numbers of UAVs: A mean field game approach", *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6266–6278, 2022.

[346] G. Wang, W. Yao, X. Zhang, and Z. Li, "A mean-field game control for large-scale swarm formation flight in dense environments", *Sensors*, vol. 22, no. 14, p. 5437, 2022.

[347] A. Šošić, A. M. Zoubir, and H. Koeppl, "Reinforcement learning in a continuum of agents", *Swarm Intell.*, vol. 12, no. 1, pp. 23–51, 2018.

[348] M. Hüttenrauch, S. Adrian, and G. Neumann, "Deep reinforcement learning for swarm systems", *J. Mach. Learn. Res.*, vol. 20, no. 54, pp. 1–31, 2019.

[349] P. Fiorini and Z. Shiller, "Motion planning in dynamic environments using velocity obstacles", *Int. J. Robot. Res.*, vol. 17, no. 7, pp. 760–772, 1998.

[350] M. Hamer, L. Widmer, and R. D'andrea, "Fast generation of collision-free trajectories for robot swarms using GPU acceleration", *IEEE Access*, vol. 7, pp. 6679–6690, 2018.

[351] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey", *IEEE Sig. Proc. Mag.*, vol. 34, no. 6, pp. 26–38, 2017.

[352] W. U. Mondal, V. Aggarwal, and S. Ukkusuri, "On the near-optimality of local policies in large cooperative multi-agent reinforcement learning", *Trans. Mach. Learn. Res.*, 2022.

[353] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. Springer Science & Business Media, 1993, vol. 303.

[354] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots", in *Proc. ICRA*, vol. 2, 1985, pp. 500–505.

[355] W. Giernacki, M. Skwierczyński, W. Witwicki, P. Wroński, and P. Kozierski, "Crazyflie 2.0 quadrotor as a platform for research and education in robotics and control engineering", in *Proc. MMAR*, 2017, pp. 37–42.

[356] M. Greiff, A. Robertsson, and K. Berntorp, "Performance bounds in positioning with the VIVE lighthouse system", in *Proc. FUSION*, 2019, pp. 1–8.

[357] C. Duan, T. Nishikawa, and A. E. Motter, "Prevalence and scalable control of localized networks", *PNAS*, vol. 119, no. 32, e2122566119, 2022.

[358] Cisco, "Cisco annual internet report (2018–2023)", Tech. Rep. C11-741490-01, 2020.

[359] R. A. Banez, H. Tembine, L. Li, C. Yang, L. Song, Z. Han, and H. V. Poor, "Mean-field-type game-based computation offloading in multi-access edge computing networks", *IEEE Trans. Wirel. Commun.*, vol. 19, no. 12, pp. 8366–8381, 2020.

[360] X. Wang, J. Ye, and J. C. Lui, "Joint D2D collaboration and task offloading for edge computing: A mean field graph approach", in *Proc. IEEE/ACM 29th Int. Symp. Qual. Serv. (IWQOS)*, IEEE, 2021, pp. 1–10.

[361] H. Gao, W. Li, R. A. Banez, Z. Han, and H. V. Poor, "Mean field evolutionary dynamics in dense-user multi-access edge computing systems", *IEEE Trans. Wirel. Commun.*, vol. 19, no. 12, pp. 7825–7835, 2020.

[362] R. Zheng, H. Wang, M. De Mari, M. Cui, X. Chu, and T. Q. Quek, "Dynamic computation offloading in ultra-dense networks based on mean field games", *IEEE Trans. Wirel. Commun.*, vol. 20, no. 10, pp. 6551–6565, 2021.

[363] L. Li, Q. Cheng, X. Tang, T. Bai, W. Chen, Z. Ding, and Z. Han, "Resource allocation for NOMA-MEC systems in ultra-dense networks: A learning aided mean-field game approach", *IEEE Trans. Wirel. Commun.*, vol. 20, no. 3, pp. 1487–1500, 2020.

[364] R. A. Banez, L. Li, C. Yang, and Z. Han, "A survey of mean field game applications in wireless networks", in *Mean Field Game and its Applications in Wireless Networks*, Springer, 2021, pp. 61–82.

[365] B. Yılmaz, A. Ortiz, and A. Klein, "Delay minimization for edge computing with dynamic server computing capacity: A learning approach", in *Proc. GLOBECOM*, 2020, pp. 1–6.

[366] P. M. Pardalos and S. A. Vavasis, "Quadratic programming with one negative eigenvalue is NP-hard", *J. Global Optim.*, vol. 1, no. 1, pp. 15–22, 1991.

[367] B. Yardim, S. Cayci, M. Geist, and N. He, "Policy mirror ascent for efficient and independent learning in mean field games", in *Proc. ICML*, PMLR, 2023, pp. 39 722–39 754.

[368] D. Lacker, K. Ramanan, and R. Wu, "Local weak convergence for sparse networks of interacting processes", *Ann. Appl. Probab.*, vol. 33, no. 2, pp. 843–888, 2023.

[369] P. Muller, M. Rowland, R. Elie, G. Piliouras, J. Perolat, M. Lauriere, R. Marinier, O. Pietquin, and K. Tuyls, "Learning equilibria in mean-field games: Introducing mean-field PSRO", in *Proc. AAMAS*, vol. 20, 2021, pp. 926–934.

[370] X. He, K. Zhao, and X. Chu, "Automl: A survey of the state-of-the-art", *Knowl.-Based Syst.*, vol. 212, p. 106 622, 2021.

[371] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks", *Proc. Natl. Acad. Sci.*, vol. 115, no. 33, E7665–E7671, 2018.

[372] J. Sirignano and K. Spiliopoulos, "Mean field analysis of neural networks: A central limit theorem", *Stoch. Process. Their Appl.*, vol. 130, no. 3, pp. 1820–1852, 2020.

[373] T. R. Jensen and B. Toft, *Graph Coloring Problems*. John Wiley & Sons, 2011.

[374] R. J. Fowler, M. S. Paterson, and S. L. Tanimoto, "Optimal packing and covering in the plane are NP-complete", *Inf. Process. Lett.*, vol. 12, no. 3, pp. 133–137, 1981.

[375] S. Grassi, H. Huang, L. Pareschi, and J. Qiu, "Mean-field particle swarm optimization", in *Modeling and Simulation for Collective Dynamics*, World Scientific, 2023, pp. 127–193.

[376] G. Neu, A. Jonsson, and V. Gómez, "A unified view of entropy-regularized Markov decision processes", *arXiv:1705.07798*, 2017.

[377] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies", in *Proc. ICML*, 2017, pp. 1352–1361.

[378] B. Belousov and J. Peters, "Entropic regularization of Markov decision processes", *Entropy*, vol. 21, no. 7, p. 674, 2019.

[379] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning", in *Proc. ICML*, 2016, pp. 1995–2003.

[380] L. Shapley, "Some topics in two-person games", *Adv. Game Theory*, vol. 52, pp. 1–29, 1964.

[381] J. J. Yeh, *Real Analysis: Theory of Measure and Integration*. World Scientific Publishing Company, 2014.

[382] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, "Array programming with NumPy", *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[383] R. F. Tchuendom, P. E. Caines, and M. Huang, "Critical nodes in graphon mean field games", in *Proc. CDC*, 2021, pp. 166–170.

[384] Y. Hu, X. Wei, J. Yan, and H. Zhang, "Graphon mean-field control for cooperative multi-agent reinforcement learning", *Journal of the Franklin Institute*, vol. 360, no. 18, pp. 14 783–14 805, 2023.

[385] O. Bonesini, L. Campi, and M. Fischer, "Correlated equilibria for mean field games with progressive strategies", *arXiv:2212.01656*, 2022.

[386] S. Sanjari and S. Yüksel, "Optimal solutions to infinite-player stochastic teams and mean-field teams", *IEEE Trans. Automat. Control*, vol. 66, no. 3, pp. 1071–1086, 2020.

[387] N. Gast and B. Gaujal, "A mean field approach for optimization in discrete time", *Discrete Event Dyn. Syst.*, vol. 21, no. 1, pp. 63–101, 2011.

[388] S. Ganapathi Subramanian, P. Poupart, M. E. Taylor, and N. Hegde, "Multi type mean field reinforcement learning", in *Proc. AAMAS*, vol. 19, 2020, pp. 411–419.

[389] X. Liu, H. Wei, and L. Ying, "Scalable and sample efficient distributed policy gradient algorithms in multi-agent networked systems", *arXiv:2212.06357*, 2022.

[390] O. Hernández-Lerma and M. Muñoz de Ozak, "Discrete-time Markov control processes with discounted unbounded costs: Optimality criteria", *Kybernetika*, vol. 28, no. 3, pp. 191–212, 1992.

[391] W. Rudin, *Principles of Mathematical Analysis*. McGraw-hill New York, 1976, vol. 3.

[392] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics", *Int. Stat. Rev.*, vol. 70, no. 3, pp. 419–435, 2002.

[393] Ş. Cobzaş, R. Miculescu, and A. Nicolae, *Lipschitz Functions*. Springer, 2019.

[394] L. Barberis, "Emergence of a single cluster in Vicsek's model at very low noise", *Phys. Rev. E*, vol. 98, no. 3, 2017.

[395] J. L. Zapotecatl, A. Munoz-Meléndez, and C. Gershenson, "Performance metrics of collective coordinated motion in flocks", in *Proc. ALIFE XV*, 2016, pp. 322–329.

ERKLÄRUNG LAUT PROMOTIONSORDNUNG

§ 8 Abs. 1 lit. c PromO
Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§ 8 Abs. 1 lit. d PromO
Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§ 9 Abs. 1 PromO
Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§ 9 Abs. 2 PromO
Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 24. Juni 2024