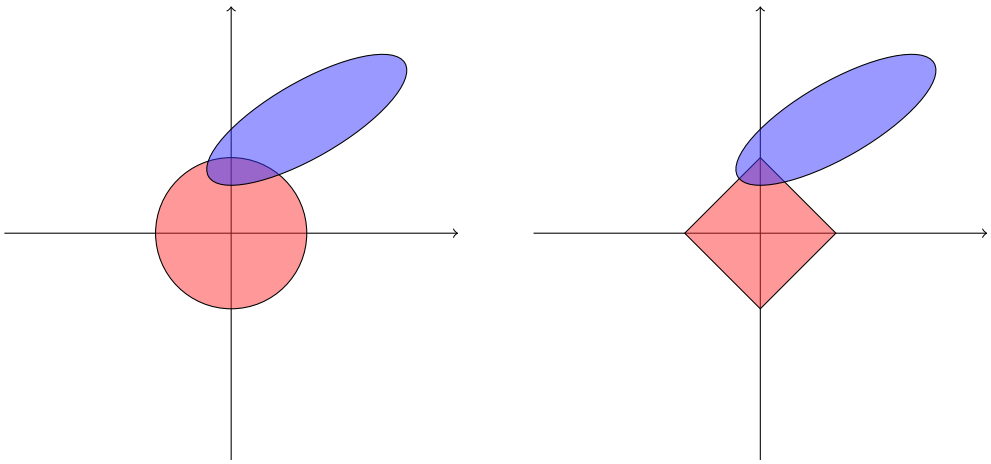
PAC-Bayesian Bandit Algorithms With Guarantees



Vom Fachbereich Informatik

an der Technischen Universität Darmstadt

Genehmigte Dissertation von Hamish Flynn aus Großbritannien

Zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)

Gutachten: Prof. Jan Peters, Prof. Melih Kandemir, Prof. Benjamin Guedj

Darmstadt, November 2023

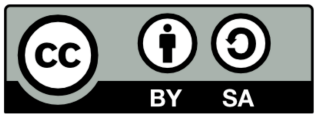
PAC-Bayesian Bandit Algorithms With Guarantees

Accepted doctoral thesis by Hamish Flynn

Date of submission: 31st August 2023

Date of thesis defense: 18th October 2023

Darmstadt, Technische Universität Darmstadt



Erklärungen 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.

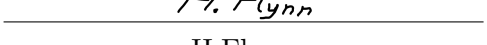
§9 Abs. 1 PromO

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, 08.11.2023



H.Flynn

Abstract

PAC-Bayes is a mathematical framework that can be used to provide performance guarantees for machine learning algorithms, explain why specific machine learning algorithms work well, and design new machine learning algorithms. Since the first PAC-Bayesian theorems were proven in the late 1990's, several impressive milestones have been achieved. PAC-Bayes generalisation bounds have been used to prove tight error bounds for deep neural networks. In addition, PAC-Bayes bounds have been used to explain why machine learning principles such as large margin classification and preference for flat minima of a loss function work well. However, these milestones were achieved in simple supervised learning problems.

In this thesis, inspired by the success of the PAC-Bayes framework in supervised learning settings, we investigate the potential of the PAC-Bayes framework as a tool for designing and analysing bandit algorithms.

First, we provide a comprehensive overview of PAC-Bayes bounds for bandit problems and an experimental comparison of these bounds. Previous works focused on PAC-Bayes bounds for martingales and their application to importance sampling-based estimates of the reward or regret of a policy. On the one hand, we found that these PAC-Bayes bounds are a useful tool for designing offline policy search algorithms with performance guarantees. In our experiments, a PAC-Bayesian offline policy search algorithm was able to learn randomised neural network polices with competitive expected reward and non-vacuous performance guarantees. On the other hand, the PAC-Bayesian online policy search algorithms that we tested had underwhelming performance and loose cumulative regret bounds.

Next, we present novel PAC-Bayes-style algorithms with worst-case regret bounds for linear bandit problems. We combine PAC-Bayes bounds with the "optimism in the face of uncertainty" principle, which reduces a stochastic bandit problem to the construction of a confidence sequence for the unknown reward function. We use a novel PAC-Bayes-style tail bound for adaptive martingale mixtures to construct convex PAC-Bayes-style confidence sequences for (sparse) linear bandits. We show that (sparse) linear bandit algorithms based on our PAC-Bayes-style confidence sequences are guaranteed to achieve competitive worst-case regret. We also show that our confidence sequences yield confidence bounds that are tighter than competitors, both empirically and theoretically. Finally, we demonstrate that our tighter PAC-Bayes-style confidence bounds result in bandit algorithms with improved cumulative regret.

Zusammenfassung

PAC-Bayes ist ein mathematisches Framework, das verwendet werden kann, um Leistungsgarantien für Algorithmen für maschinelles Lernen bereitzustellen, zu erklären, warum bestimmte Algorithmen für maschinelles Lernen gut funktionieren, und um neue Algorithmen für maschinelles Lernen zu entwerfen. Seit der Beweis der ersten PAC-Bayes'schen Theoreme Ende der 1990er Jahre wurden mehrere beeindruckende Meilensteine erreicht. PAC-Bayes-Generalisierungsgrenzen wurden verwendet, um enge Fehlergrenzen für tiefe neuronale Netze zu beweisen. Darüber hinaus wurden PAC-Bayes-Schranken verwendet, um zu erklären, warum Prinzipien des maschinellen Lernens wie die Klassifizierung mit großen Margen und die Bevorzugung flacher Minima einer Verlustfunktion gut funktionieren. Diese Meilensteine wurden jedoch bei einfachen überwachten Lernproblemen erreicht.

In dieser Arbeit untersuchen wir, inspiriert vom Erfolg des PAC-Bayes-Frameworks in überwachten Lernumgebungen, das Potenzial des PAC-Bayes-Frameworks als Werkzeug zum Entwerfen und Analysieren von Bandit-Algorithmen.

Zunächst bieten wir einen umfassenden Überblick über die PAC-Bayes-Schranken für Bandit-Probleme und einen experimentellen Vergleich dieser Schranken. Frühere Arbeiten konzentrierten sich auf PAC-Bayes-Grenzen für Martingale und deren Anwendung auf auf Wichtigkeitsstichproben basierende Schätzungen der Belohnung oder des Bedauerns einer Richtlinie. Einerseits haben wir festgestellt, dass diese PAC-Bayes-Grenzen ein nützliches Werkzeug zum Entwerfen von Offline-Richtliniensuchalgorithmen mit Leistungsgarantien sind. In unseren Experimenten war ein PAC-Bayes'scher Offline-Richtliniensuchalgorithmus in der Lage, randomisierte neuronale Netzwerkrichtlinien mit wettbewerbsfähiger erwarteter Belohnung und nicht leeren Leistungsgarantien zu erlernen. Andererseits zeigten die von uns getesteten PAC-Bayes'schen Online-Richtliniensuchalgorithmen eine enttäuschende Leistung und schwache kumulative Bedauernsgrenzen.

Als nächstes stellen wir neuartige Algorithmen im PAC-Bayes-Stil mit Worst-Case-Bedauernsgrenzen für lineare Bandit-Probleme vor. Wir kombinieren PAC-Bayes-Schranken mit dem "Optimismus angesichts der Unsicherheit"-Prinzip, das ein stochastisches Banditenproblem auf die Konstruktion einer Konfidenzfolge für die unbekannte Belohnungsfunktion reduziert. Wir verwenden eine neuartige Schwanzgrenze im PAC-Bayes-Stil für adaptive Martingalmischungen, um konvexe Konfidenzsequenzen im PAC-Bayes-Stil für (spärliche) lineare Banditen zu konstruieren. Wir zeigen, dass (spärliche) lineare Banditenalgorithmen, die auf unseren Konfidenzsequenzen im PAC-Bayes-Stil basieren, garantiert einen kompetitiven Worst-Case-Bedauern erzielen. Wir zeigen auch, dass unsere Konfidenzsequenzen sowohl empirisch als auch theoretisch engere Konfidenzgrenzen als die Konkurrenz ergeben. Schließlich zeigen wir, dass unsere engeren Vertrauensgrenzen im PAC-Bayes-Stil zu Bandit-Algorithmen mit verbessertem kumulativen Bedauern führen.

Acknowledgements

First of all, I would like to thank my university supervisor, Jan Peters, for his guidance and encouragement. I'm grateful for all the research and career advice that he gave me, and for granting me tremendous freedom to work on the research questions that interested me the most.

I would also like to thank Melih Kandemir, who supervised me during the first year and a half of my PhD when we were both still working at Bosch. I benefitted greatly from his knowledge of Bayesian inference and his creativity in shaping the research questions that we investigated.

Next, I would like to thank David Reeb, who very kindly agreed to take over as my Bosch supervisor for the last two years of my PhD. I'm grateful for the time and effort that he invested in helping me to develop the technical material in this thesis and present it in the most elegant fashion. In total, we spent many hours going through hundreds of pages of notes and calculations, which had a huge impact on both the quality of the thesis and my understanding of the underlying mathematics.

I would like to thank Benjamin Guedj, for agreeing to examine my thesis, and the other committee members, Zsolt István, Matthias Hollick and Stefan Roth, for reading my thesis and participating in my defence.

Finally, I would like to thank the many people that I had to pleasure of working with during my PhD. Thank you to the members of my Bosch research group for lots of exciting technical and lunchtime discussions. Thanks also to the members of the IAS group at TU Darmstadt, who welcomed me when I was able to visit and made the annual IAS retreats very enjoyable for me.

Contents

Chapter 1

Introduction

PAC-Bayesian (or just PAC-Bayes) theory is a branch of statistical learning theory that can be used to provide generalisation bounds for machine learning algorithms. The first PAC-Bayesian bounds were developed by McAllester [121], and took inspiration from Shawe-Taylor and Williamson [165]. Since then, Seeger [158], Catoni [36] and Maurer [118] (amongst many others) have refined the resulting PAC-Bayesian theory.

PAC-Bayes has recently grown in popularity for several reasons. Compared to classical generalisation bounds based on uniform laws of large numbers (e.g. [180, 96]), PAC-Bayes bounds are often very tight. Recently, PAC-Bayes has emerged as one of the few ways to provide non-vacuous generalisation bounds for deep neural networks [60, 61, 107, 146, 59, 135, 133, 134]. In addition, PAC-Bayes bounds can be used to design learning algorithms. A natural way to do this is to optimise a PAC-Bayes bound as a learning objective, which will return the model/hypothesis/predictor with the best generalisation guarantee. This general principle has yielded PAC-Bayesian algorithms that perform competitively with traditional algorithms (e.g. [14, 67, 170, 143]). Finally, PAC-Bayes bounds have been used to explain why specific learning strategies, such as large margin classification [31, 78, 101, 119] and preference for flat minima [80, 60, 188, 175], lead to good generalisation. However, these impressive milestones have been achieved in simple supervised learning problems, such as classification from an i.i.d. sample. It is unclear whether PAC-Bayesian theory can be used to achieve similarly impressive results in other kinds of learning problems.

Bandits, first introduced by Thompson [171] and later formalised by Robbins [147], are a relatively simple way to model the problem of decision-making under uncertainty. A bandit algorithm must learn to choose actions that maximise a reward signal. The uncertainty comes from the fact that the reward associated with each action is unknown and must be estimated based on previously observed actions and rewards. Bandit problems are frequently encountered in real-world problems, such as clinical trials [56], [26], dynamic pricing [124], [127] and recommendation [117], to name just a few. In many of these applications, one cares about algorithms that can be guaranteed to work well. There is a rich literature on (analysis of) bandit algorithms (see for instance [33, 104]). However, PAC-Bayesian theory has barely been used. This observation, and the success of PAC-Bayesian theory in supervised learning problems, provides the motivation for this thesis.

The main research question of this thesis is:

> *How can PAC-Bayesian theory be used to design bandit algorithms with performance guarantees?*

To address this question, we first establish how PAC-Bayes bounds have already been used to design bandit algorithms or provide performance guarantees for these algorithms. We review the literature on PAC-Bayesian bandit algorithms and conduct an experimental comparison to identify the strengths and weaknesses of existing approaches.

Having established what has already been done, we aim to use PAC-Bayes bounds to design new bandit algorithms that have three characteristics: a) they should have rigorous and non-vacuous performance guarantees; b) they should perform well in practice, not just in theory; c) they should be computationally efficient.

1.1 Contributions

This thesis contributes to the fields of PAC-Bayes and bandits. On the one hand, we review and compare efforts to provide PAC-Bayesian analyses of bandit problems before developing novel PAC-Bayes bounds and algorithms for various bandit problems. On the other hand, we contribute to the literature on bandits by developing PAC-Bayes as a tool that can be used to design and analyse bandit algorithms.

1.1.1 Review and Comparison of Existing PAC-Bayesian Bandit Algorithms

In Chapter 3, we provide a comprehensive overview of existing PAC-Bayes bounds and algorithms for bandit problems. We perform an experimental comparison in which we compare the tightness of various bounds, the relative merits of different reward estimates and methods for selecting the prior in a distribution or data-dependent fashion. As well as providing an overview, we provide some improved versions of existing results, such as a slightly tighter version of an Efron-Stein PAC-Bayes bound by Kuzborskij and Szepesvári [97], which holds under weaker assumptions.

1.1.2 Novel PAC-Bayesian Bounds With Adaptive Priors

One of the main theoretical results of this thesis is a general-purpose PAC-Bayes-style tail bound for martingale mixtures. Like some recently proposed PAC-Bayes bounds by Haddouche and Guedj [74, 75] and Chugg et al. [43], our new general-purpose PAC-Bayes bound is time-uniform, which means it holds simultaneously for every sample size. This makes it very well-suited to sequential learning problems, such as bandits, in which an algorithm must learn by interacting with a stream of data. Unlike in previous works, we develop PAC-Bayes-style bounds for martingale mixtures which are indexed by growing (over time) vectors of function values, as opposed to being indexed by entire functions. A key insight is that at time t, we often only need to place a mixture distribution/prior over the first t function values rather than over an entire function. Using this approach, we are able to prove PAC-Bayes bounds where the prior is refined over time as more data is observed.

1.1.3 Novel PAC-Bayesian Bandit Algorithms With Performance Guarantees

We develop several novel PAC-Bayesian bandit algorithms by combining PAC-Bayes bounds with the optimism in the face of uncertainty (OFU) principle. The OFU principle reduces a bandit problem to the problem of constructing a confidence sequence for an unknown reward function. The performance of the resulting bandit algorithm depends on the size of the confidence sequence, with smaller confidence sets yielding better empirical performance and stronger regret guarantees.

We use our general-purpose PAC-Bayes-style tail bound for martingale mixtures to construct PAC-Bayes-style confidence sequences for several bandit problems. These confidence sequences inherit several nice properties of PAC-Bayes bounds. Firstly, they are quite tight relative to other confidence sequences. Secondly, they allow one to incorporate prior knowledge about the reward function in the form of a (prior) probability distribution. Regardless of how the prior is chosen, our confidence sets contain the unknown reward function and bandit algorithms that use our confidence sets enjoy valid frequentist performance guarantees. If the prior is chosen well, then our confidence sets get smaller, our bandit algorithms perform better and their regret bounds get tighter.

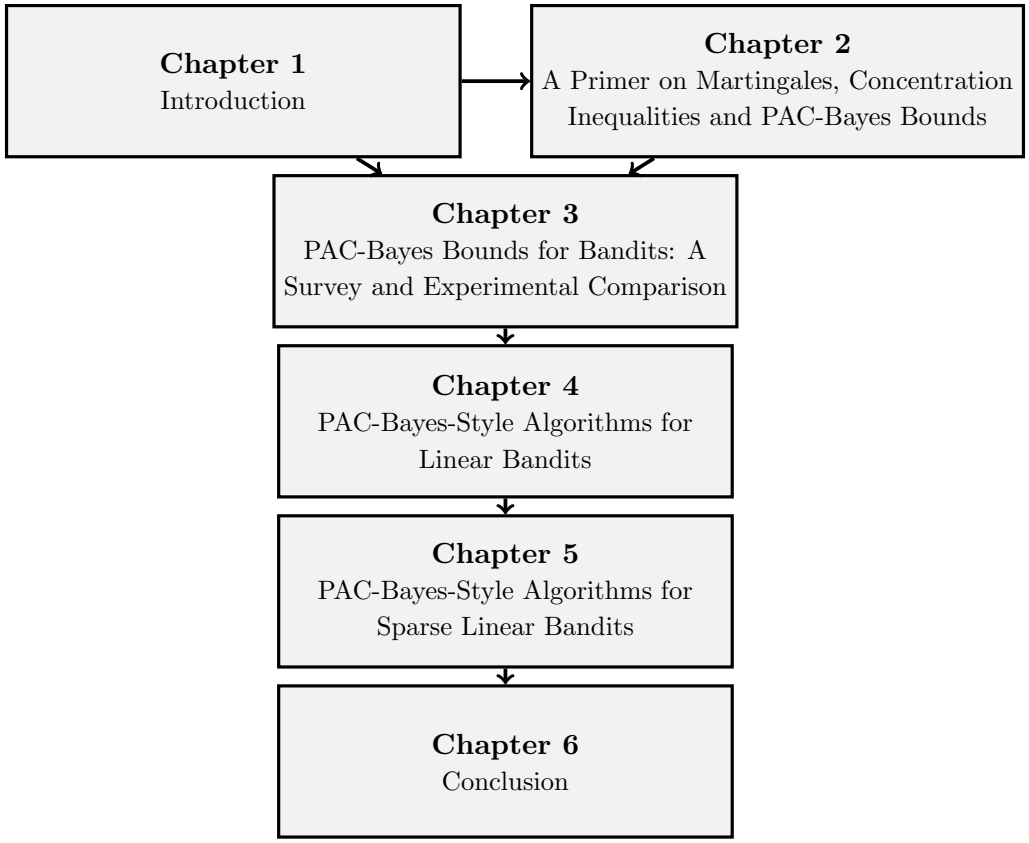


Figure 1.1: The structure of this thesis and the relation of the thesis chapters.

We specialise our general results to (sparse) linear bandit problems. In these problems our confidence sets are convex, which means the corresponding upper confidence bounds (UCBs) are found by solving a convex optimisation problem (maximisation of a linear function over the convex confi-

dence set). We either calculate the UCBs numerically using convex solvers from the CVXPY library [52] or resort to analytic upper bounds on the UCBs, which are obtained via Lagrangian duality. The resulting bandit algorithms have improved empirical performance and come with competitive cumulative regret bounds.

1.2 Thesis Outline

This thesis is organised into 6 chapters as shown in Figure 1.1. Chapter 2 covers background material on martingales, concentration inequalities and PAC-Bayes bounds, which are all recurring themes of this thesis.

In Chapter 3, we review and compare existing PAC-Bayesian bounds and PAC-Bayesian bandit algorithms. In Chapter 4, we present our general-purpose PAC-Bayes-style tail bound for martingale mixtures and use it to design upper confidence bound algorithms for linear bandits. In Chapter 5, we re-use our general-purpose tail bound to design algorithms for sparse linear bandits. In Chapter 6, we summarise our results and the contributions of the thesis before describing some directions for future research.

Chapter 2

A Primer on Martingales, Concentration Inequalities and PAC-Bayes Bounds

This chapter provides the necessary background on PAC-Bayes bounds and martingales. In Section 2.1, we introduce martingales and some definitions from measure theory. In Section 2.2, we introduce Ville's inequality for non-negative supermartingales and use it to derive a tail bound for a random walk as an example. In Section 2.3, we introduce PAC-Bayes bounds and the Donsker-Varadhan change of measure inequality. As an example, we derive a PAC-Bayes bound for mixtures of random walks.

2.1 Some Probability and Martingale Theory

In this thesis, we will often avoid using measure-theoretic notions, but we will sometimes refer to measurable functions, σ-algebras, filtrations and sequences of adapted or predictable random variables. This section aims to: (a) provide some background on these terms; (b) introduce martingales. The definitions in this section come from [187], [58] and [24].

2.1.1 Probability Spaces

We begin by introducing the notion of a *probability space* $(\Omega, \mathcal{D}, \mathbb{P})$. Ω is a set of outcomes and is called the sample space. $\mathcal{D}$ is a *σ-algebra* (on Ω), which means it is a collection of subsets of Ω that satisfies: (a) $\Omega \in \mathcal{D}$; (b) for every set $A \in \mathcal{D}$, $A^{\mathsf{c}} \in \mathcal{D}$ (where $A^{\mathsf{c}} := \Omega \setminus A$); (c) for every countable collection of sets $A_1, A_2, \cdots \in \mathcal{D}$, $\cup_n A_n \in \mathcal{D}$. $\mathbb{P}$ is a *probability measure* (on $\mathcal{D}$), which means it is a set function $\mathbb{P} : \mathcal{D} \to [0, 1]$ that satisfies: (a) $\mathbb{P}(\Omega) = 1$; (b) for every countable sequence of disjoint sets $A_1, A_2, \ldots, \in \mathcal{D}$, $\mathbb{P}(\cup_n A_n) = \sum_n \mathbb{P}(A_n)$.

Probability spaces can be more easily understood through examples. Suppose we conduct a random experiment, in which we toss a fair coin once. The sample space is $\Omega = \{H, T\}$, which contains all outcomes of the experiment (i.e. heads (H) or tails (T)). We can take the σ-algebra to be either

$\mathcal{D}_0 = \{\emptyset, \{H, T\}\}$ or $\mathcal{D}_1 = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$. Since the coin is fair, the probability measure satisfies $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}$ and $\mathbb{P}(\Omega) = 1$.

Based on this example, we can think of a σ-algebra as a representation of the information available at some time. Before we toss the coin, we only know that the outcome ω will be H or T. We can say whether ω is in $\emptyset$ or $\{H, T\}$, but we cannot say whether ω is in $\{H\}$ or $\{T\}$. Said another way, for every set in $\mathcal{D}_0$, we can say whether ω is in that set. Therefore, we can think of $\mathcal{D}_0$ as representing the information available before we toss the coin. After tossing the coin, we can also say whether ω is in each set in $\mathcal{D}_1$. We can therefore think of $\mathcal{D}_1$ as representing the information available after we toss the coin.

For a given sample space Ω, we call $\{\emptyset, \Omega\}$ the *trivial σ-algebra*, which is the smallest σ-algebra on Ω. In the coin toss example, $\mathcal{D}_0$ is the trivial σ-algebra on $\Omega = \{H, T\}$.

For a given sample space Ω and a collection $\mathcal{E}$ of subsets of Ω, one can always construct a smallest σ-algebra $\sigma(\mathcal{E})$ that contains all sets in $\mathcal{E}$. See, for example, Lemma 1.8. of [24] for proof. We call $\sigma(\mathcal{E})$ the *σ-algebra generated by $\mathcal{E}$*. If $\mathcal{E} = \{\Omega\}$, (or $\mathcal{E} = \{\emptyset\}$), then $\sigma(\mathcal{E})$ is the trivial σ-algebra. If $\mathcal{E}$ is the set of open intervals of the form (a, b), where $a < b \in \mathbb{R}$, then $\sigma(\mathcal{E})$ is the *Borel σ-algebra* on $\mathbb{R}$. We denote the Borel σ-algebra on $\mathbb{R}$ by $\mathcal{B}(\mathbb{R})$. We call the elements of $\mathcal{B}(\mathbb{R})$ the *Borel sets*.

2.1.2 Random Variables

We are now ready to introduce measurable sets, measurable functions and random variables. If a set $A \subseteq \Omega$ is in the σ-algebra $\mathcal{D}$, then we say that A is *$\mathcal{D}$-measurable*. Since the domain of the probability measure $\mathbb{P}$ is $\mathcal{D}$, the probability $\mathbb{P}(A)$ is only defined when $A \in \mathcal{D}$. In other words, one can only measure (the probability of) sets in $\mathcal{D}$, hence the term measurable.

Suppose that we have a sample space Ω and a σ-algebra $\mathcal{D}$. The pair $(\Omega, \mathcal{D})$ is called a *measurable space*. For a function $X : \Omega \to \mathbb{R}$ and a Borel set $B \in \mathcal{B}(\mathbb{R})$, the *pre-image* of B is $X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\}$, which is the set of outcomes that get mapped into B by the function X. The function X is called a *$\mathcal{D}/\mathcal{B}(\mathbb{R})$-measurable function* (from $(\Omega, \mathcal{D})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$) if for every Borel set $B \in \mathcal{B}(\mathbb{R})$, $X^{-1}(B) \in \mathcal{D}$. A *random variable* on a probability space $(\Omega, \mathcal{D}, \mathbb{P})$ is a measurable function $X : \Omega \to \mathbb{R}$.

To see why this definition makes sense, consider the probability of a random variable X being less than or equal to 0. Letting $B = (-\infty, 0] \in \mathcal{B}(\mathbb{R})$, we have

$$\mathbb{P}(X \leq 0) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}) = \mathbb{P}(X^{-1}(B)).$$

This probability is only defined when $X^{-1}(B)$ is $\mathcal{D}$-measurable, i.e. when $X^{-1}(B)$ is in the domain of the probability measure $\mathbb{P}$. If we view $\mathbb{P}(X \leq x)$ as a function of $x \in \mathbb{R}$, then this is the usual cumulative distribution function for the random variable X.

When we work with σ-algebras, we will typically use a σ-algebra generated by a random variable. For a given random variable $X : \Omega \to \mathbb{R}$, $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$ is the *σ-algebra generated by X*. This is the smallest σ-algebra such that $X : \Omega \to \mathbb{R}$ is a measurable function. In the coin toss example from earlier, if the random variable X is defined as $X(H) = 1$ and $X(T) = 0$, then the σ-algebra generated by X is $\mathcal{D}_1 = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$.

2.1.3 Expectation

We now build up to the definition of conditional expectation. A random variable is called *simple* if there exist $n > 0$, $x_1, \ldots, x_n \in \mathbb{R}$ and $A_1, A_2, \ldots, A_n \in \mathcal{D}$ such that

$$X(\omega) = \sum_{i=1}^{n} x_i \mathbb{I}\{\omega \in A_i\}.$$

The *expectation* of a random variable X is: a) If X is simple with $X(\omega) = \sum_{i=1}^{n} x_i \mathbb{I}\{\omega \in A_i\}$, then $\mathbb{E}[X] := \sum_{i=1}^{n} x_i \mathbb{P}(A_i)$; b) If $X \geq 0$, then $\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n]$, where $X_1, X_2, \ldots$ is a non-decreasing sequence of simple random variables ($X_1, X_2, \ldots$ can be guaranteed to exist and the limit is unique (see e.g. Theorem 1.14 of [24]), but we may have $\mathbb{E}[X] = \infty$); c) If X is a random variable, then $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$, where $X^+ = \max(X, 0)$ and $X^- = -\min(X, 0)$.

Now, assume we have a probability space $(\Omega, \mathcal{D}_0, \mathbb{P})$ and a sub-σ-algebra $\mathcal{D}$, which is a σ-algebra that satisfies $\mathcal{D} \subset \mathcal{D}_0$. For a random variable that is $\mathcal{D}_0$-measurable and satisfies $\mathbb{E}[|X|] < \infty$, the *conditional expectation* $\mathbb{E}[X|\mathcal{D}]$ is any random variable V that satisfies: a) V is $\mathcal{D}$-measurable; b) for all $A \in \mathcal{D}$, $\mathbb{E}[X \cdot \mathbb{I}\{\omega \in A\}] = \mathbb{E}[V \cdot \mathbb{I}\{\omega \in A\}]$.

The conditional expectation can be more easily understood through examples. First, suppose that $\mathbb{E}[|X|] < \infty$ and let $\mathcal{D} = \sigma(X)$. One can verify that $V = X$ satisfies conditions a) and b). Clearly $V = X$ is $\mathcal{D}$-measurable and $\mathbb{E}[V \cdot \mathbb{I}\{\omega \in A\}] = \mathbb{E}[X \cdot \mathbb{I}\{\omega \in A\}]$ for all $A \in \mathcal{D}$ when $V = X$. Therefore, we have

$$\mathbb{E}[X|\sigma(X)] = \mathbb{E}[X|X] = X.$$

Now, let $\mathcal{D} = \{\emptyset, \Omega\}$ be the trivial σ-algebra. The requirement that V is $\mathcal{D}$-measurable must mean that V is constant. Since we must have $\mathbb{E}[X \cdot \mathbb{I}\{\omega \in \Omega\}] = \mathbb{E}[V \cdot \mathbb{I}\{\omega \in \Omega\}] \implies \mathbb{E}[X] = \mathbb{E}[V]$, we conclude that $\mathbb{E}[X|\{\emptyset, \Omega\}] = \mathbb{E}[X]$.

2.1.4 Martingales

Put simply, martingales are sequences of random variables, for which at any point in the sequence, the conditional expectation of the present value is equal to the previous value.

To define a martingale properly, we need to first define a *filtered space* $(\Omega, \mathcal{D}, (\mathcal{D}_t | t \in \mathbb{N}), \mathbb{P})$. As with a probability space, Ω is the sample space, $\mathcal{D}$ is a σ-algebra and $\mathbb{P}$ is a probability measure. $(\mathcal{D}_t | t \in \mathbb{N})$ is a *filtration*, which means it is a sequence of σ-algebras that satisfies $\mathcal{D}_0 \subseteq \mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \cdots \subseteq \mathcal{D}$. We also need to define an adapted random process. A random process (a sequence of random variables) $(X_t | t \in \mathbb{N})$ is *adapted* to the filtration $(\mathcal{D}_t | t \in \mathbb{N})$ if for every $t \in \mathbb{N}$, X_t is $\mathcal{D}_t$-measurable. This means that $\mathcal{D}_t$ contains all information about X_t (given $\mathcal{D}_t$, X_t is not random anymore). A random process $(X_t | t \in \mathbb{N})$ is *predictable* if for every $t \in \mathbb{N}$, X_t is $\mathcal{D}_{t-1}$-measurable. A very common way to define a filtration is to set $\mathcal{D}_t = \sigma(X_1, \ldots, X_t)$ for a random process $(X_t | t \in \mathbb{N})$.

A random process $(M_t | t \in \mathbb{N})$ in a probability space $(\Omega, \mathcal{D}, \mathbb{P})$ is a *martingale* with respect to the filtration $(\mathcal{D}_t | t \in \mathbb{N})$ if: (a) $(M_t | t \in \mathbb{N})$ is adapted to $(\mathcal{D}_t | t \in \mathbb{N})$; (b) $\mathbb{E}[|M_t|] < \infty$; (c); $\mathbb{E}[M_t | \mathcal{D}_{t-1}] = M_{t-1}$ for all t. We call property (c) the *martingale property.*

If instead of the martingale property, we have $\mathbb{E}[M_t | \mathcal{D}_{t-1}] \leq M_{t-1}$, then $(M_t | t \in \mathbb{N})$ is called a *supermartingale.* Similarly, if instead the martingale property, we have $\mathbb{E}[M_t | \mathcal{D}_{t-1}] \geq M_{t-1}$,

then $(M_t|t \in \mathbb{N})$ is called a *submartingale*. If an adapted random process $(X_t|t \in \mathbb{N})$ satisfies $\mathbb{E}[X_t|\mathcal{D}_{t-1}] = 0$ for every t (and $\mathbb{E}[|X_t|] < \infty$), then $(X_t|t \in \mathbb{N})$ is called a *martingale difference sequence*, because its sum $M_t = \sum_{k=1}^{t} X_k$ is a martingale.

2.1.5 Basic Definition

In this thesis, we will always assume that there is a (filtered) probability space in background, but we will not state what the probability space is. Similarly, we typically prefer to use "simpler" expressions where appropriate. With that in mind, we present a basic definition of a martingale.

A sequence of random variables $(M_n|n \in \mathbb{N})$ is a martingale with respect to another sequence of random variables $(X_n|n \in \mathbb{N})$ if, for every $n \in \mathbb{N}$: (a) M_n is fully determined by $X_1, \ldots, X_n$ (i.e. M_n conditioned on $X_1, \ldots, X_n$ is non-random); (b) $\mathbb{E}[|M_n|] < \infty$; (c) $\mathbb{E}[M_n|X_1, \ldots, X_{n-1}] = M_{n-1}$.

2.2 Concentration Inequalities via Martingale Methods

In this thesis, we are interested martingales as a tool for deriving tail bounds and concentration inequalities for random variables and random processes. A *tail bound* for a random variable Z is an upper bound on the tail probability $\mathbb{P}(Z > z)$, for some $z \in \mathbb{R}$. Tail bounds for Z can be stated as equivalent high probability upper bounds on the value of Z. For instance, if we are given the tail bound $\mathbb{P}(Z > z) \leq \delta$, then we can say that, with probability at least $1 - \delta$ (over the random draw of Z), $Z \leq z$. We will usually work with statements of the second kind.

A *concentration inequality* for a random variable Z is a high probability upper bound on the difference between Z and its expected value $\mathbb{E}[Z]$ (or another quantity such as its median). In this thesis, we will typically work with concentration inequalities stated in the form: with probability at least $1 - \delta$, $|Z - \mathbb{E}[Z]| \leq z$.

A *time-uniform tail bound* for a random process $(Z_t|t \in \mathbb{N})$ is an upper bound on the probability that at *any* time $t \geq 1$, Z_t is greater than a fixed value $z_t \in \mathbb{R}$, i.e. $\mathbb{P}(\exists t \geq 1 : Z_t > z_t)$. If we are given the time-uniform tail bound $\mathbb{P}(\exists t \geq 1 : Z_t > z_t) \leq \delta$, then we can say that, with probability $1 - \delta$ (over the random draw of $(Z_t|t \in \mathbb{N})$), for all $t \geq 1$ simultaneously, $Z_t \leq z_t$. A *time-uniform concentration inequality* for a random process $(Z_t|t \in \mathbb{N})$ is a time-uniform tail bound for $(|Z_t - \mathbb{E}[Z_t]||t \in \mathbb{N})$.

2.2.1 Ville's Inequality for Non-Negative Supermartingales

In this section, we introduce Ville's inequality for non-negative supermartingales [181], which can be thought of as a time-uniform version of Markov's inequality.

Lemma 2.1 (Ville's inequality for non-negative supermartingales [181])**.** *Let $(M_t|t \in \mathbb{N})$ be any non-negative supermartingale with respect to a filtration $(\mathcal{D}_t|t \in \mathbb{N})$, which satisfies $M_0 \leq 1$. For any $\delta \in (0, 1]$, we have*

$$\mathbb{P}(\exists t \geq 1, M_t > 1/\delta) \leq \delta.$$

Equivalently, with probability at least $1 - \delta$,

$$\forall t \geq 1 : \quad M_t \leq 1/\delta.$$

Ville's inequality states that with high probability (at least $1-\delta$), the random process $(M_t | t \in \mathbb{N})$ will *never* exceed the value $1/\delta$. In other words, it provides a time-uniform tail bound for $(M_t | t \in \mathbb{N})$. Streamlined proofs of Ville's inequality can be found in [81] and [75].

As an example, consider the following game. At time $t = 0$, we are given £1. At each time $t = 1, 2, \ldots$, with equal probability we either increase or decrease our total money by 5%. The amount of money we have at time t can be modelled by a non-negative supermartingale. Let

$$X_t = \begin{cases} 1.05 & \text{with prob. } 0.5 \\ 0.95 & \text{with prob. } 0.5 \end{cases}, \quad M_t = \prod_{k=1}^{t} X_k, \quad M_0 := 1.$$

$(M_t | t \in \mathbb{N})$ is a martingale with respect to the filtration $(\mathcal{D}_t | t \in \mathbb{N})$, where $\mathcal{D}_t$ is the σ-algebra generated by $X_1, X_2, \ldots, X_t$, since

$$\mathbb{E}\left[\prod_{k=1}^{t} X_k \middle| \mathcal{D}_{t-1}\right] = \prod_{k=1}^{t-1} X_k \mathbb{E}[X_t | \mathcal{D}_{t-1}] = \prod_{k=1}^{t-1} X_k.$$

Clearly $(M_t | t \in \mathbb{N})$ is non-negative and $M_0 = 1$ by definition. We can therefore use Ville's inequality to obtain a time-uniform tail bound for the amount of money we have in this game. In Fig. 2.1, we show 100 random draws of the amount of money in this game over 10000 time steps. We also plot the tail bound from Ville's inequality at the confidence level $\delta = 0.01$

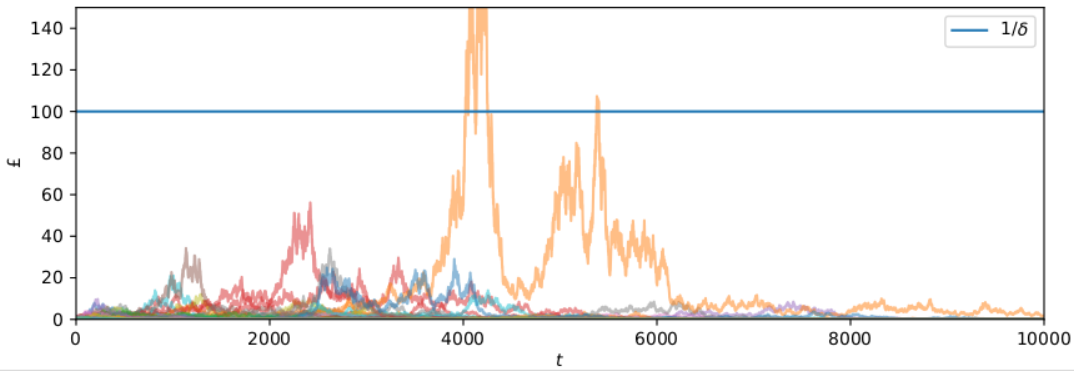


Figure 2.1: A time-uniform bound for the amount of money in the game and 100 draws of the amount of money in the game.

In this simulation, the empirical failure-rate of the tail bound is actually equal to $\delta = 0.01$ (i.e. only 1 out of the 100 random draws ever exceeds £100). Ville's inequality tells us that as the number of simulations approaches ∞, this failure-rate will approach a number less than or equal to 0.01.

Suppose we can stop playing the game at any point. The tail bound from Ville's inequality tells us something about our chances of making money in this game. If we want to finish the game with at least £100, we might decide to stop playing as soon as our amount of money reaches or exceeds £100. The tail bound (at the level $\delta = 0.01$) tells us that, with probability at least 0.99,

this strategy will fail: our amount of money will never reach or exceed £100 and we will never stop playing the game.

2.2.2 Tail Bounds For Random Walks

We will now show how to use Ville's inequality to obtain time-uniform tail bounds for some random processes that are not necessarily non-negative supermartingales. We present an example, which is an instance of a time-uniform Chernoff bound [81].

We have a random process $(Z_t|t \in \mathbb{N})$, which is adapted to a filtration $(\mathcal{D}_t|t \in \mathbb{N})$. The only restriction on $(Z_t|t \in \mathbb{N})$ is that $M_t = \exp(Z_t)$ must be a non-negative supermartingale with $M_0 = 1$. This requirement, combined with Jensen's inequality, implies that $(Z_t|t \in \mathbb{N})$ is itself a (not necessarily non-negative) supermartingale, since

$$\mathbb{E}[Z_t|\mathcal{D}_{t-1}] = \ln(\exp(\mathbb{E}[Z_t|\mathcal{D}_{t-1}])) \leq \ln(\mathbb{E}[\exp(Z_t)|\mathcal{D}_{t-1}]) \leq Z_{t-1}.$$

To apply Ville's inequality to $(Z_t|t \in \mathbb{N})$, we need to relate the tail probabilities of $(Z_t|t \in \mathbb{N})$ to those of $(M_t|t \in \mathbb{N})$. Since e^x is monotonically increasing in x, for any $\delta \in (0, 1]$, we have

$$\mathbb{P}(\exists t \geq 1, Z_t > \ln(1/\delta)) = \mathbb{P}(\exists t \geq 1, \exp(Z_t) > 1/\delta) \leq \delta. \tag{2.1}$$

Equivalently, with probability at least $1 - \delta$,

$$\forall t \geq 0 : \quad Z_t \leq \ln(1/\delta).$$

Let $(X_t|t \in \mathbb{N})$ be a sequence of independent standard Gaussian random variables and let $(\mathcal{D}_t|t \in \mathbb{N})$ be the filtration where $\mathcal{D}_t = \sigma(X_1, \ldots, X_t)$. We call the sum $\sum_{k=1}^t X_k$ a random walk. To obtain an upper tail bound for the random walk, we could try the choice $Z_t = \sum_{k=1}^t X_k$. Unfortunately, $M_t = \exp(\sum_{k=1}^t X_k)$ is a submartingale instead of a supermartingale, since

$$\mathbb{E}\left[\exp\left(\sum_{k=1}^{t} X_k\right) \middle| \mathcal{D}_{t-1}\right] = \mathbb{E}[\exp(X_t)] \exp\left(\sum_{k=1}^{t-1} X_k\right) = \exp(1/2) \exp\left(\sum_{k=1}^{t-1} X_k\right) \geq \exp\left(\sum_{k=1}^{t-1} X_k\right). \tag{2.2}$$

From Eq. (2.2), we can see that if we subtract $1/2$ from each X_t (i.e. we choose $Z_t = \sum_{k=1}^t (X_k - 1/2)$, then $M_t = \exp(\sum_{k=1}^t (X_k - 1/2))$ is a (super)martingale with $M_0 = 1$. We will actually make a more general choice $Z_t = \sum_{k=1}^t (\lambda_t X_k - \lambda_t^2/2)$, where $(\lambda_t|t \in \mathbb{N})$ is a sequence of fixed positive real numbers, which also has the property that $M_t = \exp(Z_t)$ is a (super)martingale with $M_0 = 1$. One can think of $\lambda_1, \lambda_2, \ldots$ as parameters of the resulting tail bounds, which we will optimise to achieve tighter bounds.

The combination of Eq. (2.1) and Ville's inequality gives us a tail bound. For any fixed sequence of positive real numbers $(\lambda_t|t \in \mathbb{N})$ and any $\delta \in (0, 1]$, with probability at least $1 - \delta$, we have

$$\forall t \geq 1; \quad \sum_{k=1}^{t} (\lambda_t X_k - \lambda_t^2/2) \leq \ln(1/\delta).$$

For each t, this can be rearranged to give an upper bound for the random walk $\sum_{k=1}^{t} X_k$, which is

$$\sum_{k=1}^{t} X_k \leq \frac{\lambda_t}{2} + \frac{\ln(1/\delta)}{\lambda_t}.$$

Since the right-hand-side does not depend on the random process $(X_t | t \in \mathbb{N})$, we can freely optimise each λ_t to obtain

$$\sum_{k=1}^{t} X_k \leq \min_{\lambda_t < 0} \left\{ \frac{\lambda_t}{2} + \frac{\ln(1/\delta)}{\lambda_t} \right\} = \sqrt{2t \ln(1/\delta)}. \tag{2.3}$$

In Fig. 2.2, we plot 100 random draws of the random walk from $t = 1, \ldots, 1000$ and the tail bound in Eq. (2.3) at the level $\delta = 0.01$.

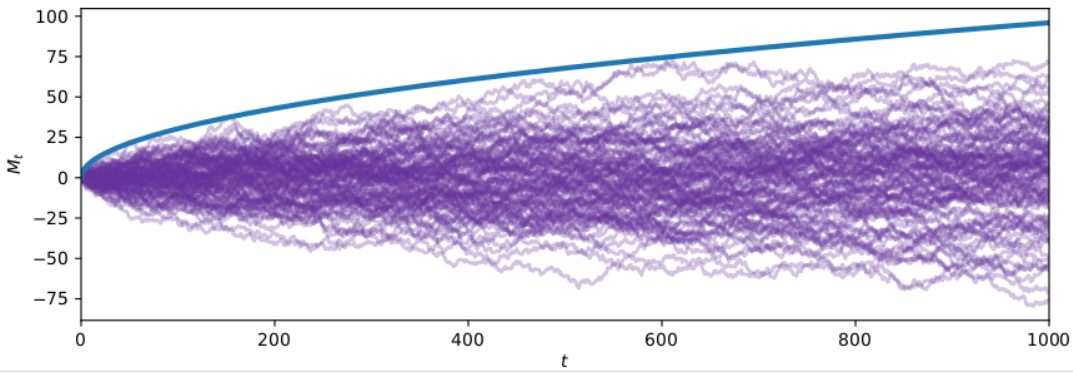


Figure 2.2: A time-uniform upper tail bound for a random walk (blue) and 100 draws of the random walk (purple). With probability at least 0.01 (over the random draw of the purple line), the purple line should never be above the blue line.

2.3 PAC-Bayes Bounds via Martingale Methods

We are now ready to introduce PAC-Bayes bounds. PAC-Bayes bounds [165, 121, 70, 11] can be defined as Probably Approximately Correct (PAC) [176] performance bounds for (generalised) Bayesian learning algorithms. A PAC bound states that, with high probability (*probably*), the error-rate of the hypothesis returned by a learning algorithm is upper bounded. If this upper bound on the error-rate is small, then the hypothesis returned by the learning algorithm is *approximately correct*. When PAC bounds are applied to Bayesian learning algorithms, the result is called a PAC-Bayes bound.

In this thesis, we will not work in the classical PAC learning setting and we will use a broader definition of what constitutes a PAC-Bayes bound. In the context of this thesis, PAC-Bayes bounds are *uniform* tail bounds/concentration inequalities for *mixtures* of random variables/processes. Here, the "uniform" part means that a PAC-Bayes bound holds uniformly over mixture distributions.

Suppose that for each value of an integer-valued parameter $\theta \in \{1, \ldots, 100\}$, $(X_t(\theta) | t \in \mathbb{N})$ is a sequence of independent standard Gaussian random variables. Let $(\mathcal{D}_t | t \in \mathbb{N})$ be the filtration

where $\mathcal{D}_t = \sigma(X_1(1), \dots, X_1(100), \dots, X_t(1), \dots, X_t(100))$. For each θ and any sequence of positive real numbers $(\lambda_t | t \in \mathbb{N})$, $\sum_{k=1}^t X_k(\theta)$ is a random walk and $M_t(\theta) = \exp(\sum_{k=1}^t (\lambda_t X_k(\theta) - \lambda_t^2/2))$ is a non-negative supermartingale with $M_0 = 1$.

To introduce PAC-Bayes bounds, we will show how to derive a PAC-Bayesian upper tail bound for the mixture of random walks $\mathbb{E}_{\theta\sim Q}[\sum_{k=1}^t X_k(\theta)]$, which is an upper bound on $\mathbb{E}_{\theta\sim Q}[\sum_{k=1}^t X_k(\theta)]$ that holds with high probability for all mixture distributions $Q \in \mathcal{P}(\{1, \dots, 100\})$ simultaneously (including mixture distributions that depend the values of $X_1(\theta), X_2(\theta), \dots$). The bound presented here is an example of a time-uniform PAC-Bayes bound [74, 75, 43]. $\mathcal{P}(\{1, \dots, 100\})$ denotes the set of probability distributions on the set $\{1, \dots, 100\}$. To derive the bound, we will use a very useful result, which is now standard in the PAC-Bayesian literature.

Lemma 2.2 (Donsker-Varadhan Change of Measure [53, 36])**.** *For any set $\mathcal{X}$, any measurable function $h : \mathcal{X} \to \mathbb{R}$ and any probability distribution $P \in \mathcal{P}(\mathcal{X})$ (i.e. any distribution on $\mathcal{X}$), such that $\mathbb{E}_{x\sim P}[\exp(h(x))] < \infty$, we have*

$$\sup_{Q\in\mathcal{P}(\mathcal{X})} \left\{ \mathop{\mathbb{E}}_{x\sim Q}[h(x)] - D_{\mathrm{KL}}(Q||P) \right\} = \ln\left(\mathop{\mathbb{E}}_{x\sim P}[\exp(h(x))] \right). \tag{2.4}$$

If h is upper bounded on the support of P, then the supremum is achieved when Q is the Gibbs distribution P_h, which is defined as

$$P_h(x) = \frac{P(x)\exp(h(x))}{\mathbb{E}_{x\sim P}[\exp(h(x))]}. \tag{2.5}$$

By rearranging (2.4), we also have

$$\inf_{Q\in\mathcal{P}(\mathcal{X})} \left\{ \mathop{\mathbb{E}}_{x\sim Q}[h(x)] + D_{\mathrm{KL}}(Q||P) \right\} = -\ln\left(\mathop{\mathbb{E}}_{x\sim P}[\exp(-h(x))] \right). \tag{2.6}$$

It is now fairly straightforward to obtain a PAC-Bayesian tail bound for $\mathbb{E}_{\theta\sim Q}[\sum_{k=1}^t X_k(\theta)]$. One can verify that for any distribution P that does not depend on $X_t(\theta)$ for any t or θ, the mixture of martingales (or martingale mixture) $(\mathbb{E}_{\theta\sim P}[M_t(\theta)] | t \in \mathbb{N})$ is also a non-negative supermartingale with $M_0 = 1$. Using the Donsker-Varadhan change of measure inequality, we have

$$\exp\left(\sup_Q \left\{ \mathop{\mathbb{E}}_{\theta\sim Q}\left[\sum_{k=1}^t (\lambda_t X_k(\theta) - \lambda_t^2/2) \right] - D_{\mathrm{KL}}(Q||P) \right\} \right) = \mathop{\mathbb{E}}_{\theta\sim P}[M_t(\theta)]. \tag{2.7}$$

The left-hand-side of (2.7) must also be a non-negative supermartingale. Using Ville's inequality, we have that for any $\delta \in (0, 1]$, with probability at least $1 - \delta$,

$$\forall t \geq 1; \quad \exp\left(\sup_Q \left\{ \mathop{\mathbb{E}}_{\theta\sim Q}\left[\sum_{k=1}^t (\lambda_t X_k(\theta) - \lambda_t^2/2) \right] - D_{\mathrm{KL}}(Q||P) \right\} \right) \leq 1/\delta. \tag{2.8}$$

In other words, this inequality holds simultaneously for all distributions Q. After rearranging (2.8), we obtain a PAC-Bayes upper tail bound for mixtures of random walks.

$$\mathop{\mathbb{E}}_{\theta\sim Q}\left[\sum_{k=1}^{t} X_k(\theta)\right] \leq \frac{t\lambda_t}{2} + \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda_t}.$$

Note that this tail bound holds for a *fixed* (before observing $X_1(\theta), X_2(\theta), \dots$) sequence $(\lambda_t | t \in \mathbb{N})$. We cannot optimise the right-hand-side of this tail bound with respect to λ_t because the optimal value of λ_t depends on Q (through $D_{\mathrm{KL}}(Q||P)$), which may depend on the observations $X_1(\theta), X_2(\theta), \dots$. In the interest of obtaining a deterministic upper bound (to plot in Figure 2.3), we choose P to be a uniform distribution, upper bound the KL divergence by $\ln(100)$ and set $\lambda_t = \sqrt{2\ln(100/\delta)/t}$, to obtain

$$\mathop{\mathbb{E}}_{\theta\sim Q}\left[\sum_{k=1}^{t} X_k(\theta)\right] \leq \sqrt{2t\ln(100/\delta)}.$$

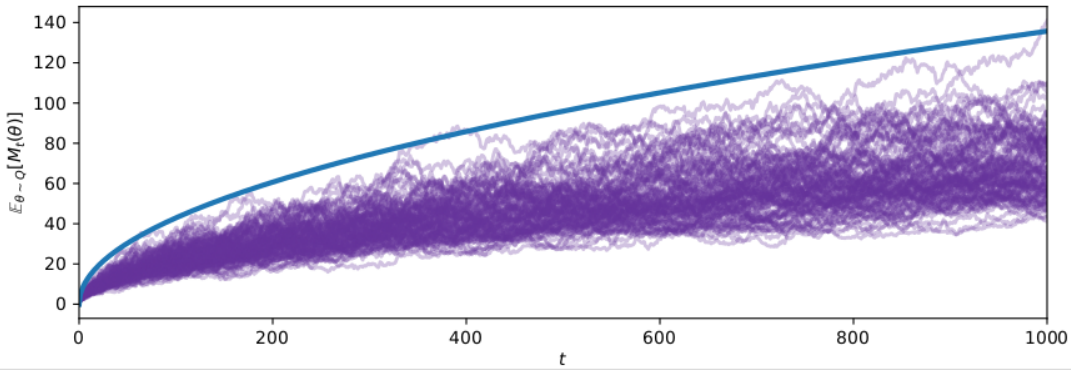


Figure 2.3: A time-uniform PAC-Bayes upper tail bound for mixtures of random walks (blue) and 100 draws of the Gibbs mixture of random walks (purple). With probability at least 0.01 (over the random draw of the purple line), the purple line should never be above the blue line.

In Figure 2.3, we plot 100 random draws of the mixture of random walks $\mathbb{E}_{\theta\sim Q}\left[\sum_{k=1}^{t} X_k(\theta)\right]$, where Q is the Gibbs distribution with density $Q(\theta) \propto \exp(\lambda_t \sum_{k=1}^{t} X_k(\theta))$. Note that the draws of this mixture of random walks look quite different to the individual random walks in Figure 2.2. For instance, the Gibbs mixture of random walks is clearly neither a martingale nor a supermartingale, since its expected value increases over time.

Chapter 3

PAC-Bayes Bounds for Bandits: A Survey and Experimental Comparison

This chapter provides an overview of PAC-Bayes bounds for bandit problems and an experimental comparison of these bounds. Following the literature on PAC-Bayes bounds and algorithms for bandits, we focus on policy search algorithms that learn a policy from data using reward estimates based on importance sampling. On the one hand, we found that PAC-Bayes bounds are a useful tool for designing offline bandit algorithms with performance guarantees. In our experiments, a PAC-Bayesian offline contextual bandit algorithm was able to learn randomised neural network polices with competitive expected reward and non-vacuous performance guarantees. On the other hand, the PAC-Bayesian online bandit algorithms that we tested had loose cumulative regret bounds. We conclude this chapter by discussing some open research questions about PAC-Bayesian bandit algorithms, some of which are the subject of subsequent chapters.

3.1 Introduction

At the time of writing, there is neither a detailed overview of PAC-Bayes bounds for bandit problems nor an experimental comparison of these bounds. It is therefore difficult to know which PAC-Bayes bandit bounds give the best guarantees or how tight the best bounds are. There are two main reasons why we believe that now is the right time to review PAC-Bayesian approaches to bandits. First, PAC-Bayes bounds have recently been used to design effective offline bandit algorithms with performance guarantees [115, 154]. Second, PAC-Bayes has been growing in popularity due to numerous successful applications to deep learning. In parallel, there has been growing interest in bandit algorithms that use deep neural network function approximation. We believe that it is worth investigating whether PAC-Bayes would be a useful tool for studying these deep bandit algorithms.

The scope of this survey is determined by the selection of PAC-Bayesian approaches to bandits that can be found in the literature. Consequently, we focus on policy search algorithms that directly learn a policy from data using reward estimates based on importance sampling. We found that there were no model-based PAC-Bayesian bandit algorithms, which first model the reward function and then use this model to learn a policy, so we do not cover these approaches. However, we discuss the compatibility of PAC-Bayes with other approaches to bandits in Sec.

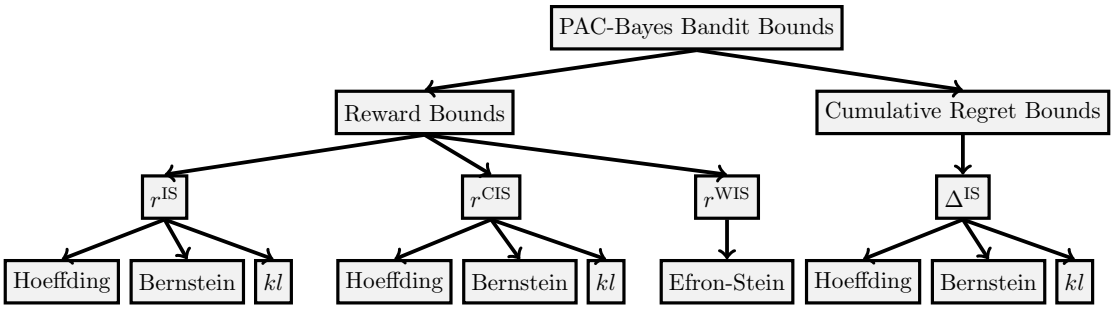


Figure 3.1: A taxonomy of existing PAC-Bayes bandit bounds. The bounds are first separated into lower bounds on reward and upper bounds on cumulative regret. At the next level, the bounds are categorised by the empirical reward/regret estimate that they use. The reward estimates r^{IS}, r^{CIS}, and r^{WIS}, and the regret estimate Δ^{IS}, are defined in Section 3.4, Appendix 3.4.3 and Section 3.5. Finally, the bounds are divided according to the concentration inequality that they use in their proofs. kl is the Binary KL divergence, defined in Section 3.4.

We cover offline and online variants of both multi-armed and contextual bandit problems. We consider two types of PAC-Bayes bounds: one for offline bandits and one for online bandits. For offline bandits, we consider lower bounds on the expected reward of a policy learned from historical data. For online bandits, we consider upper bounds on the cumulative regret suffered by playing a sequence of policies. The bounds considered in this survey are categorised further in Fig. 3.1.

We only consider stationary, stochastic bandit problems, where the rewards are sampled from fixed distributions. We do not cover extensions such as restless bandits [185] or adversarial bandits [22]. We also do not cover bandit problems with additional structural assumptions, such as linear bandits [20].

Findings. We compared the values of the bounds, as well as the performance of bandit algorithms motivated by the bounds. On the one hand, we found that some of the PAC-Bayes lower bounds on the expected reward are surprisingly tight, particularly when data-dependent priors are used. Moreover, we found that directly optimising PAC-Bayes reward bounds can yield effective offline bandit algorithms. PAC-Bayes appears to be a useful tool for designing offline bandit algorithms with performance guarantees. On the other hand, we found that the few existing PAC-Bayes cumulative regret bounds are all loose, and that the algorithms motivated by these bounds are noticeably worse than state-of-the-art methods. The reason for this is that both the bounds and algorithms rely on loose upper bounds on the variance of importance sampling-based reward estimates.

Related work. PAC-Bayes bounds have been the subject of several tutorials [120], [178], [106], [11], surveys [70] and monographs [38]. McAllester [120] describes 3 different types of PAC-Bayes bounds and presents a new application of PAC-Bayes bounds to dropout. Van Erven [178] describes the relationship between PAC-Bayes bounds and some classical concentration inequalities. Laviolette [106] describes the history of PAC-Bayes bounds as well as some recent developments. Alquier [11] gives an overview of PAC-Bayes bounds for supervised learning and an introduction to localised bounds, fast-rate bounds and bounds for non i.i.d. data and unbounded losses. Guedj [70] surveys the PAC-Bayes framework, its links to Bayesian methods, and some theoretical and algorithmic developments. Catoni [38] provides a rich analysis of supervised classification using PAC-Bayes

bounds. There have been a few experimental comparisons of some PAC-Bayes bounds [63], [135] in supervised learning problems. There are several books [33], [167], [104] about bandit algorithms and their performance guarantees. However, none of these resources on bandits cover PAC-Bayes.

Chapter Overview. First, we formally describe the online and offline variants of multi-armed and contextual bandit problems in Sec. 3.2. In Sec. 3.3, we describe the PAC-Bayesian approach to the bandit problems introduced in Sec. 3.2. We then provide a structured overview of PAC-Bayes bounds for bandit problems and some techniques for achieving the tightest bound values. Sec. 3.4 reviews PAC-Bayes lower bounds on the expected reward, Sec. 3.5 reviews PAC-Bayes upper bounds on the cumulative regret, and Sec. 3.6 reviews techniques for optimising PAC-Bayes bandit bounds with respect to the prior and other parameters. In Sec. 3.7, we compare the PAC-Bayes bandit bounds in several experiments. Finally, in Sec. 3.8, we discuss our findings and comment on some open problems.

Contributions. Our first contribution is a comprehensive overview of existing PAC-Bayes bounds for bandit problems. Our second contribution is an experimental comparison of PAC-Bayes bounds and algorithms for bandit problems. We also provide a slightly tighter version of the Efron-Stein PAC-Bayes bound by Kuzborskij and Szepesvári [97], which holds under slightly weaker conditions.

3.2 Problem Formulation

The goal of all the bandit problems we consider is to select the best policy π from a set of policies Π, which we call the policy class. In this paper, we are interested in bandit algorithms that return a probability distribution over the policy class rather than a single policy. $\mathcal{P}(\Pi)$ denotes the set of all probability distributions over the policy class.

The choice of policy is informed by data. We use $\mathcal{Z}$ to denote the observation space. A bandit algorithm observes or collects a data set of observations $D_T = \{z_t\}_{t=1}^T$. Each z_t is drawn from a distribution $\mathbb{P}_t$ over $\mathcal{Z}$. In bandit problems, we may have non-identically distributed data, where $\mathbb{P}_t \neq \mathbb{P}_{t'}$ for $t \neq t'$. We may also have dependent data, where z_t depends on $z_1, \ldots, z_{t-1}$. Usually, we will make $\mathcal{Z}$ more explicit. In the simplest case, we observe pairs of actions and rewards, so $\mathcal{Z} = \mathcal{A} \times \mathcal{R}$ where $\mathcal{A}$ is a set of actions and $\mathcal{R}$ is a set values that the rewards can take.

3.2.1 Policy Search for Multi-Armed Bandits

A multi-armed bandit (MAB) problem is a tuple $\langle \mathcal{A}, \mathcal{R}, \mathbb{P}_R \rangle$. $\mathcal{A}$ is a set of actions (or arms), $\mathcal{R}$ is a set of values that the rewards can take and $\mathbb{P}_R(\cdot|a)$ is a distribution over rewards conditioned on the action a. $\mathcal{A}$ and $\mathcal{R}$ are known, but $\mathbb{P}_R$ is unknown. Throughout this paper, we assume that the rewards are bounded between 0 and 1, so $\mathcal{R} \subseteq [0, 1]$.

A bandit algorithm selects actions through a policy π. In a MAB problem, a policy is a (possibly degenerate) probability distribution over the set of actions $\mathcal{A}$. $\pi(a)$ denotes the probability of selecting action a under the policy π. In the offline MAB problem, an algorithm is given a data set $D_T = \{(a_t, r_t)\}_{t=1}^T$. We let $D_{t-1} = \{(a_k, r_k)\}_{k=1}^{t-1}$ denote the first $t-1$ elements of D_T. Each action a_t is sampled from a behaviour policy b_t ($b_t(a)$ denotes the probability of selecting action a under the policy b_t). Each reward r_t is sampled from the reward distribution, given a_t. In the most general setting, the sequence of behaviour policies $b_1, b_2, \ldots$ can be arbitrary, as long as b_t only

depends on the previously observed data D_{t-1}. Some of the PAC-Bayes bounds we will encounter hold only when the data set D_T consists of i.i.d. samples. For these bounds to hold, we require that the entire data set is drawn using a fixed behaviour policy b, which must be independent of all the data D_T. We always assume that the behaviour policies are known. The expected reward for a policy π is defined as:

$$R(\pi) = \mathop{\mathbb{E}}_{a\sim\pi(\cdot), r\sim\mathbb{P}_R(\cdot|a)}[r]\,. \tag{3.1}$$

For a probability distribution $Q \in \mathcal{P}(\Pi)$ over the policy class, the expected reward is $R(Q) = \mathbb{E}_{\pi\sim Q}[R(\pi)]$. Given a policy class Π and a data set D_T, the goal of policy search in the offline MAB problem is to return a distribution $Q^* \in \mathcal{P}(\Pi)$ that maximises the expected reward:

$$Q^* \in \operatorname*{argmax}_{Q\in\mathcal{P}(\Pi)} \{R(Q)\}\,.$$

In the online MAB problem, an algorithm must learn and act simultaneously. Policy search in the online MAB problem proceeds in rounds. At round t, the algorithm selects a distribution $Q_t \in \mathcal{P}(\Pi)$ to be played. A policy π_t is drawn from Q_t and then an action a_t is drawn from the policy π_t. The algorithm observes a reward r_t drawn from the reward distribution $\mathbb{P}_R(\cdot|a_t)$. To guide the selection of Q_t at each round, the algorithm can use the action-reward pairs gathered from previous rounds. In other words, the choice of Q_t can depend on the data D_{t-1}.

The goal of policy search in the online MAB problem is to select a sequence of distributions (over the policy class) $Q_1, \ldots, Q_T$ that minimises the cumulative regret. For a sequence of policies $\pi_1, \ldots, \pi_T$, the regret for round t and the cumulative regret are defined as:

$$\Delta(\pi_t) = R(\pi^*) - R(\pi_t), \qquad \Delta(\pi_{1:T}) = \sum_{t=1}^{T} \Delta(\pi_t),$$

where $\pi^* \in \operatorname{argmax}_{\pi\in\Pi}\{R(\pi)\}$ is an optimal policy. The per-round regret and cumulative regret for a sequence of distributions $Q_1, \ldots, Q_T$ are defined as:

$$\Delta(Q_t) = R(\pi^*) - R(Q_t), \qquad \Delta(Q_{1:T}) = \sum_{t=1}^{T} \Delta(Q_t). \tag{3.2}$$

The goal of minimising cumulative regret brings about a dilemma known as the exploration-exploitation trade-off. To achieve low cumulative regret, an algorithm must try out lots of policies to identify which ones have the highest expected reward. However, while it identifies which policies are the best, it must also limit the number of times it selects a sub-optimal policy.

Example 3.1 (Clinical trial)**.** There are two flu treatments and we are given the results of a clinical trial where 100 patients have been randomly given either treatment A or treatment B. We want to decide which treatment is better. This can be modelled as an offline multi-armed bandit problem, where the actions are the treatment types and the rewards are the outcomes of the treatments. A PAC-Bayes reward bound could give a lower bound on the success rate of each treatment.

If we wanted to assign treatments to patients sequentially, with the goal of handing out the better treatment as often as possible, this could be modelled as an online bandit problem. A PAC-Bayes

cumulative regret bound could tell us (before handing out any treatments) an upper bound on the gap between the optimal expected number of successful treatments and the expected number of successful treatments of our allocation strategy.

3.2.2 Policy Search for Contextual Bandits

A contextual bandit (CB) problem is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbb{P}_S, \mathbb{P}_R \rangle$. $\mathcal{S}$ is a set of states (or contexts), $\mathcal{A}$ is a set of actions, $\mathcal{R}$ is a set of values that the rewards can take, $\mathbb{P}_S(\cdot)$ is a distribution over the set of states and $\mathbb{P}_R(\cdot|s,a)$ is a distribution over rewards conditioned on the state s and the action a. $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$ are known, but $\mathbb{P}_S$ and $\mathbb{P}_R$ are unknown. As in the MAB problem, we assume that $\mathcal{R} \subseteq [0,1]$ throughout this paper.

In a CB problem, a policy is a function that maps states to probability distributions over the set of actions $\mathcal{A}$. $\pi(a|s)$ denotes the probability of selecting action a, given the state s, under the policy π. The expected reward for a policy π is defined as:

$$R(\pi) = \mathop{\mathbb{E}}_{s\sim\mathbb{P}_S(\cdot), a\sim\pi(\cdot|s), r\sim\mathbb{P}_R(\cdot|s,a)} [r] .$$

As before, the expected reward for a distribution $Q \in \mathcal{P}(\Pi)$ over the policy class is $R(Q) = \mathbb{E}_{\pi\sim Q}[R(\pi)]$. The distinction between offline and online CB problems is very similar to the MAB case. In the offline CB problem, a data set $D_T = \{(s_t, a_t, r_t)\}_{t=1}^T$ of state-action-reward triples is available. The states s_t are sampled from the state distribution P_S, the actions a_t are sampled from behaviour policies $b_t(\cdot|s_t)$ and the rewards r_t are sampled from the reward distribution $\mathbb{P}_R(\cdot|s_t, a_t)$. Whenever we require an i.i.d. data set, we will assume each action a_t is drawn from the same, fixed behaviour policy $b(\cdot|s_t)$. Given a policy class Π and a data set D_T, the goal of policy search in the offline CB problem is to return a distribution $Q^* \in \mathcal{P}(\Pi)$ that maximises the expected reward.

In round t of the online CB problem, an algorithm selects a distribution Q_t. A state s_t is drawn from P_S, a policy π_t is drawn from Q_t, an action a_t is drawn from $\pi_t(\cdot|s_t)$, and then a reward r_t is drawn from $\mathbb{P}_R(\cdot|s_t, a_t)$. The choice of Q_t can depend on the data D_{t-1}. The goal is to select a sequence of distributions $Q_1, \ldots, Q_T$ that minimises the cumulative regret. Per-round regret and cumulative regret are defined in the same way as in (3.2).

3.3 PAC-Bayesian Policy Search Algorithms

A PAC-Bayesian approach to the policy search problems described in Sec. 3.2 proceeds as follows. First, we fix a reference distribution or prior $P \in \mathcal{P}(\Pi)$ over the policy class Π. Then we observe data D_T, either a batch of historical data or the data collected in previous rounds, which helps us to learn another distribution $Q \in \mathcal{P}(\Pi)$, which we will call a posterior distribution.

In the context of these policy search problems, a PAC-Bayes bound is an upper bound on either the difference between $R(Q)$ and an empirical estimate of the reward of Q or the difference between $\Delta(Q)$ and an empirical estimate of the regret of Q, which holds uniformly over all posteriors Q. One of the empirical reward estimates we consider is the importance sampling (IS) estimate, which is defined as

$$r^{\mathrm{IS}}(\pi, D_T) = \frac{1}{T} \sum_{t=1}^{T} \frac{\pi(a_t)}{b_t(a_t)} r_t. \tag{3.3}$$

The IS estimate is an average of the observed rewards weighted by the importance weights $\pi(a_t)/b_t(a_t)$. When upper bounding the difference between $R(Q)$ and $r^{\mathrm{IS}}(Q, D_T) = \mathbb{E}_{\pi\sim Q}\left[r^{\mathrm{IS}}(\pi, D_T)\right]$, we face challenges that are not present in typical PAC-Bayesian learning settings. For example, the data D_T are often not independent or identically distributed. This challenge can be dealt with using the martingale methods presented in Chapter 2.

3.3.1 PAC-Bayesian Offline Bandit Algorithms

We continue our introduction to PAC-Bayes bounds for bandits with an example. We present a PAC-Bayes bound for the expected reward $R(Q)$ in the MAB setting, which was originally proposed by Seldin et al. [162]. Then, we present an offline bandit algorithm that is motivated by this bound. In Sec. 3.2.1, we stated that the goal of the offline policy search problem is to choose a distribution $Q^* \in \mathcal{P}(\Pi)$ that maximises the expected reward, i.e.

$$Q^* \in \operatorname*{argmax}_{Q\in\mathcal{P}(\Pi)} \left\{ \mathop{\mathbb{E}}_{\pi\sim Q}\left[R(\pi)\right] \right\}.$$

Since the reward distribution $\mathbb{P}_R$ is unknown, we cannot directly maximise $R(\pi)$. However, $R(\pi)$ can be estimated from historical data D_T by the importance sampling estimate $r^{\mathrm{IS}}(\pi, D_T)$. We will assume that for all the behaviour policies $b_1, \ldots, b_T$, the importance weights $\pi(a)/b_t(a)$ are uniformly (over π, a) bounded above by $1/\epsilon_T$. We can maximise $r^{\mathrm{IS}}(Q, D_T) = \mathbb{E}_{\pi\sim Q}\left[r^{\mathrm{IS}}(\pi, D_T)\right]$ with respect to Q. However, if the estimate $r^{\mathrm{IS}}(Q, D_T)$ greatly overestimates the expected reward $R(Q)$ for even a single choice of Q, simply maximising the reward estimate may result in overfitting. When can we guarantee that $r^{\mathrm{IS}}(Q, D_T)$ does not greatly overestimate $R(Q)$? PAC-Bayes bounds can provide an answer.

Theorem 3.2 (PAC-Bayes Hoeffding-Azuma bound for r^{IS} [162])**.** *For any $\lambda > 0$, any $\delta \in (0, 1)$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1-\delta$ (over the sampling of D_T), for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously:*

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - \frac{\lambda}{8T\epsilon_T^2} - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda}.$$

Thm. 3.2 states that if Q is close to the prior P (as measured by the KL divergence) and $r^{\mathrm{IS}}(Q, D_T)$ is high, then with high probability it is guaranteed that $R(Q)$ is also high. We can define an offline policy search algorithm that returns the distribution $\widehat{Q}$ that maximises the lower bound in Thm. 3.2, and hence has the best performance guarantee. The resulting optimisation problem is

$$\widehat{Q} \in \operatorname*{argmax}_{Q\in\mathcal{P}(\Pi)} \left\{ \mathop{\mathbb{E}}_{\pi\sim Q}[r^{\mathrm{IS}}(\pi, D_T)] - \frac{D_{\mathrm{KL}}(Q||P)}{\lambda} \right\}. \tag{3.4}$$

The change of measure inequality in Lemma 2.2 shows that the optimisation problem in (3.4) has a closed-form solution: $\widehat{Q}(\pi) \propto P(\pi)e^{\lambda r^{\mathrm{IS}}(\pi, D_T)}$. When the policy class Π is finite, the normalisation constant of $\widehat{Q}$ can be calculated by summing over all $\pi \in \Pi$. When Π is infinite, one can design algorithms that approximate $\widehat{Q}$ with variational inference [182], [30] or algorithms that sample from $\widehat{Q}$ using Monte Carlo methods [16], [25]. Of course, if Π is a complicated (e.g. high-dimensional)

policy class, then approximating or sampling from $\widehat{Q}$ may be challenging. However, these challenges are beyond the scope of this survey.

Offline bandit algorithms that learn a policy by maximising a PAC-Bayesian lower bound on the expected reward are not new. London and Sandler [115] and Sakhi et al. [154] have used better PAC-Bayes reward bounds, which we will encounter in Section 3.4, to design PAC-Bayesian offline contextual bandit algorithms that learn policies with performance guarantees.

3.3.2 PAC-Bayesian Online Bandit Algorithms

One can also use PAC-Bayes bounds to design online bandit algorithms. In Section 3.2.1, we stated that the goal of the online policy search problem is to choose a sequence of distributions $Q_1, \ldots, Q_T \in \mathcal{P}(\Pi)$ that minimise the cumulative regret $\Delta(Q_{1:T})$ (see Equation (3.2)). We can estimate the regret at round t using the IS regret estimate, which is defined as

$$\Delta^{\mathrm{IS}}(\pi, D_t) = r^{\mathrm{IS}}(\pi^*, D_t) - r^{\mathrm{IS}}(\pi, D_t).$$

We can also obtain a PAC-Bayes Hoeffding-Azuma bound for the IS regret estimate [162] (see Theorem 3.12), which states that (with high probability)

$$\forall Q \in \mathcal{P}(\Pi), \quad \Delta(Q) \leq \Delta^{\mathrm{IS}}(Q, D_t) + \frac{\lambda}{2t\epsilon_t^2} + \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda}. \tag{3.5}$$

We could set each distribution Q_t to the minimiser of the bound in (3.5). However, we also need to consider the effect of each $\pi_t \sim Q_t$ on the subsequent importance weights, and on ϵ_t in particular. In Section 3.5, we will see how use PAC-Bayes bounds to obtain PAC-Bayesian online bandit algorithms that come with cumulative regret bounds.

PAC-Bayes bounds on regret estimates have already been used to design online bandit algorithms with cumulative regret bounds. Seldin et al. [162, 159] have derived PAC-Bayes bounds for the IS regret estimate and used them to design online (contextual) bandit algorithms. As we will see, these algorithms closely resemble the EXP3 and EXP4 algorithms [22].

3.3.3 Relation To Existing Methods

The basic algorithm in (3.4) can provide a new perspective on some well-known principles for policy search.

Example 3.3 (Relative Entropy Regularisation [90], [137], [23])**.** Let the policy class be the set of all deterministic policies. Then $\Pi = \mathcal{A}$ and both Q and P are now individual stochastic policies. Suppose there is a single behaviour policy b, and set $P = b$. The optimisation problem in (3.4) becomes

$$\widehat{Q} \in \operatorname*{argmax}_{Q \in \mathcal{P}(\mathcal{A})} \left\{ \mathop{\mathbb{E}}_{a \sim Q}[r^{\mathrm{IS}}(a, D_T)] - \frac{D_{\mathrm{KL}}(Q||b)}{\lambda} \right\}.$$

This motivates maximising the IS reward estimate subject to a penalty on the relative entropy between Q and the behaviour policy b. Relative Entropy Policy Search [136], Trust Region Policy Optimization [156] and Proximal Policy Optimization [157] are all based upon this principle of relative entropy regularisation.

Example 3.4 (Maximum Entropy [86, 87, 190])**.** Let $\Pi = \mathcal{A}$. This time, choose the prior P to be a uniform distribution over $\mathcal{A}$. The KL divergence between Q and a uniform distribution is equal a constant minus the Shannon entropy $H(Q)$ of Q. The optimisation problem in (3.4) becomes

$$\widehat{Q} \in \operatorname*{argmax}_{Q \in \mathcal{P}(\mathcal{A})} \left\{ \mathop{\mathbb{E}}_{a \sim Q}[r^{\mathrm{IS}}(a, D_T)] + \frac{H(Q)}{\lambda} \right\}.$$

This motivates maximisation of a weighted sum of the reward estimate and the entropy of Q, or alternatively, choosing the policy Q with the highest entropy subject to a constraint that the reward estimate is sufficiently high. This is essentially equivalent to a classical strategy known as Boltzmann exploration [89]. In addition, several modern deep reinforcement learning algorithms, such as Soft Q-learning [72] and Soft Actor-Critic [73], follow the maximum entropy principle.

3.4 PAC-Bayes Reward Bounds

In this section, we give an overview of PAC-Bayes bounds for the expected reward, organised by the reward estimate used in the bound.

3.4.1 Importance Sampling

We have already encountered the importance sampling (IS) estimate, which was defined in (3.3). We remind the reader of the assumption that the importance weights $\pi(a)/b_t(a)$ are uniformly bounded above by $1/\epsilon_T$ for every $t = 1, \ldots, T$. This can be achieved by constraining the behaviour policies and/or the policy class Π.

One of the most well-known PAC-Bayes bounds is the PAC-Bayes *kl* bound, which was proposed by Seeger [158] and improved by Maurer [118]. The binary KL divergence is defined as:

$$kl(p||q) := p \ln\left(\frac{p}{q}\right) + (1-p)\ln\left(\frac{1-p}{1-q}\right).$$

This is the KL divergence between a Bernoulli distribution with parameter p and a Bernoulli distribution with parameter q, and is defined for $p, q \in [0, 1]$ (although it is infinite if $q = 0$ or $q = 1$). Seldin et al. [162] derived the PAC-Bayes *kl* bound for the IS estimate:

Theorem 3.5 (PAC-Bayes *kl* bound for r^{IS}[162])**.** *For any $\delta \in (0, 1)$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1 - \delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously,*

$$kl\big(\epsilon_T r^{\mathrm{IS}}(Q, D_T) \,\big\|\, \epsilon_T R(Q)\big) \leq \frac{D_{\mathrm{KL}}(Q||P) + \ln(2\sqrt{T}/\delta)}{T}.$$

The original PAC-Bayes *kl* bound holds only for i.i.d. data, yet the bound in Thm. 3.5 holds even when the behaviour policies are dependent on previous observations. Seldin et al. [162] achieve this extra generality by using a comparison inequality (Lem. 1 of [161]) that bounds expectations of convex functions of certain martingale-like sequences by expectations of the same functions of independent Bernoulli random variables. In this form, the PAC-Bayes *kl* bound is not so useful;

we would prefer a lower bound on $R(Q)$. Following Seeger [158], the lower inverse of the binary KL divergence can be defined as:

$$kl^{-1}(p, B) := \min\{q \in [0, 1] : kl(p||q) \leq B\}.$$

With this definition, the PAC-Bayes kl bound for the r^{IS} estimate can be rewritten as:

$$R(Q) \geq \frac{1}{\epsilon_T} kl^{-1}\left(\epsilon_T r^{\text{IS}}(Q, D_T), \frac{D_{\text{KL}}(Q||P) + \ln(2\sqrt{T}/\delta)}{T}\right). \tag{3.6}$$

We refer to this bound as the PAC-Bayes kl^{-1} bound. This is the tightest possible lower bound on $R(Q)$ that can be derived from the PAC-Bayes kl bound. From the definition of kl^{-1}, it is apparent that this bound is never vacuous (less than 0). Unfortunately, kl^{-1} has no closed-form solution. However, it can be calculated numerically to arbitrary accuracy using, for example, the bisection method. Instead of inverting the binary KL divergence, one can use Pinsker's inequality [139], which states that

$$|p - q| \leq \sqrt{kl(p||q)/2}.$$

We can then obtain a (looser) high probability lower bound on the expected reward:

$$R(Q) \geq r^{\text{IS}}(Q, D_T) - \frac{1}{\epsilon_T}\sqrt{\frac{D_{\text{KL}}(Q||P) + \ln(2\sqrt{T}/\delta)}{2T}}. \tag{3.7}$$

We refer to this bound as the PAC-Bayes Pinsker bound. Several authors [119], [174], [170], [146] have used tighter (than Pinsker's inequality) bounds on the binary KL divergence to obtain better, more explicit PAC-Bayes bounds from the PAC-Bayes kl bound. Similar bounds for the IS reward estimate can be obtained by combining the same techniques with Theorem 3.5.

Seldin et al. [160] provide a PAC-Bayes bound for the IS estimate that depends on the variance of the reward estimate. The (conditional) average variance of the IS estimate for the policy π is defined as

$$V^{\text{IS}}(\pi, D_T) = \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{a'_t \sim b_t, r'_t \sim \mathbb{P}_R(\cdot|a'_t)} \left[\left(\frac{\pi(a'_t)}{b_t(a'_t)} r'_t - R(\pi)\right)^2\right].$$

Note that this is a data-dependent quantity. $V^{\text{IS}}(\pi, D_T)$ is the average variance of the IS estimate given the sequence of behaviour policies that are selected based on the data D_T. We write $V^{\text{IS}}(Q, D_T) = \mathbb{E}_{\pi \sim T}\left[V^{\text{IS}}(\pi, D_T)\right]$. The bound is derived by using Bernstein's inequality for martingales instead of the Hoeffding-Azuma inequality.

Theorem 3.6 (PAC-Bayes Bernstein Bound for r^{IS} [160])**.** *For any $\lambda \in [0, T\epsilon_T]$, any $\delta \in (0, 1)$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1 - \delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously:*

$$R(Q) \geq r^{\text{IS}}(Q, D_T) - \frac{\lambda(e-2)V^{\text{IS}}(Q, D_T)}{T} - \frac{D_{\text{KL}}(Q||P) + \ln(1/\delta)}{\lambda}.$$

Seldin et al. [160] show that the variance for any policy π satisfies $V^{\mathrm{IS}}(\pi, D_T) \leq 1/\epsilon_T$. This bound on the variance leads to the lower bound

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - \frac{\lambda(e-2)}{T\epsilon_T} - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda}. \tag{3.8}$$

The PAC-Bayes bounds for the IS estimate can be compared by examining their rates in T and ϵ_T and $D_{\mathrm{KL}}(Q||P)$. The rate at which they degrade as ϵ_T approaches 0 becomes particularly important as the action set $\mathcal{A}$ grows. For example, if $\mathcal{A} = \{1, \ldots, K\}$ and the behaviour policies are all uniform, then $\epsilon_T \leq 1/K$. Alternatively, if $\mathcal{A}$ is a bounded subset of $\mathbb{R}^d$ and the behaviour policies are all uniform, then $\epsilon_T \leq \mathcal{O}(1/\mathrm{vol}(\mathcal{A}))$. These examples suggest that if a PAC-Bayes bound degrades rapidly as ϵ_T decreases, then the bound may be loose when $\mathcal{A}$ is large.

The Pinsker bound in (3.7) has a rate of $\mathcal{O}(\sqrt{D_{\mathrm{KL}}(Q||P)}/(\epsilon_T\sqrt{T}))$, ignoring the $\ln(\sqrt{T})$ term. For the PAC-Bayes Hoeffding-Azuma bound in Theorem 3.2, it can be shown that the optimal value of λ is proportional to $\epsilon_T\sqrt{T}$. With this choice of λ, this bound also has a rate of $\mathcal{O}(\sqrt{D_{\mathrm{KL}}(Q||P)}/(\epsilon_T\sqrt{T}))$. The PAC-Bayes Bernstein bound in (3.8) has an improved rate in ϵ_T. For a suitable choice of λ, this bound has a rate of $\mathcal{O}(\sqrt{D_{\mathrm{KL}}(Q||P)}/\sqrt{\epsilon_T T})$. A Taylor expansion reveals that the PAC-Bayes kl^{-1} bound in (3.6) decays approximately exponentially in $D_{\mathrm{KL}}(Q||P)/(\epsilon_T T)$ as $\frac{r^{\mathrm{IS}}(Q,D_T)}{e} \exp \frac{-D_{\mathrm{KL}}(Q||P) - \ln(2\sqrt{T}/\delta)}{T\epsilon_T r^{\mathrm{IS}}(Q,D_T)}$. Based on these rates, we can expect the PAC-Bayes Bernstein and PAC-Bayes kl^{-1} bounds to scale better to large action sets.

Finally, we discuss PAC-Bayes bounds for the IS estimate in the contextual bandit setting. In the CB setting, the IS estimate is defined as

$$r^{\mathrm{IS}}(\pi, D_T) = \frac{1}{T}\sum_{t=1}^{T} \frac{\pi(a_t|s_t)}{b_t(a_t|s_t)} r_t. \tag{3.9}$$

We still require that $1/\epsilon_T$ is a uniform bound on the importance weights, though now the importance weights are $\pi(s_t|a_t)/b_t(a_t|s_t)$. As in the MAB setting, one can construct martingales containing the IS estimate that are compatible with the Hoeffding-Azuma, Bernstein and Seldin et al.'s comparison inequality [161]. Therefore, the same PAC-Bayes Hoeffding-Azuma, PAC-Bayes kl and PAC-Bayes Bernstein bounds as in Theorems 3.2, 3.5 and 3.6 can be derived. In the CB versions of these bounds, ϵ_T and the reward estimate $r^{\mathrm{IS}}(\pi, D_T)$ are just defined slightly differently. A PAC-Bayes Bernstein bound for the IS estimate in the CB setting was first derived by Seldin et al. [159].

3.4.2 Clipped Importance Sampling

In Section 3.4.1, we saw that all the PAC-Bayes bounds for the IS reward estimate degrade as the uniform bound $1/\epsilon_T$ on the importance weights increases. One way to ensure that $1/\epsilon_T$ is never too large is to clip the importance weights. Clipped (or truncated) importance sampling was first proposed by Ionides [82]. The clipped importance sampling (CIS) reward estimate for MAB problems is defined as

$$r^{\mathrm{CIS}}(\pi, D_T) = \frac{1}{T}\sum_{t=1}^{T} \min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t. \tag{3.10}$$

By definition, the clipped importance weights are bounded above by $1/\tau$. However, clipping the importance weights introduces bias. Let $R^{\mathrm{CIS}}(\pi) = \mathbb{E}_{D_T}[r^{\mathrm{CIS}}(\pi, D_T)]$ denote the expected value of the CIS estimate, and let $R^{\mathrm{CIS}}(Q) = \mathbb{E}_{\pi\sim Q}[R^{\mathrm{CIS}}(\pi)]$. It can be shown (see Lemma A.2) that the CIS estimate is biased to underestimate the expected reward, i.e. $R^{\mathrm{CIS}}(Q) \leq R(Q)$.

Therefore, any lower bound on $R^{\mathrm{CIS}}(Q)$ is also a lower bound on $R(Q)$, which means the we can essentially ignore the bias of the CIS estimate if we only require a lower bound on the expected reward. It is possible to derive PAC-Bayes Hoeffding-Azuma, Bernstein and *kl* bounds for the CIS estimate that are almost the same as those for the IS estimate, except that ϵ_T is replaced by τ. Under the assumption that there is a single, fixed behaviour policy, Wang et al. [183] have proved PAC-Bayes Pinsker, Hoeffding-Azuma and Bernstein bounds for the CIS risk (one minus reward) estimate.

In Appendix A.1.3, we prove the following PAC-Bayes Hoeffding-Azuma bound for the CIS reward estimate, which holds in the most general setting where the data are drawn from an arbitrary sequence of behaviour policies.

Theorem 3.7 (PAC-Bayes Hoeffding-Azuma bound for r^{CIS})**.** *For any $\tau \in (0, 1]$, any $\lambda > 0$, any $\delta \in (0, 1)$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1 - \delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously,*

$$R(Q) \geq r^{\mathrm{CIS}}(Q, D_T) - \frac{\lambda}{8T\tau^2} - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda}. \tag{3.11}$$

Like λ, the clipping parameter τ must be independent of the data D_T. We don't claim that this bound is new, since it is essentially a corollary of the PAC-Bayes Hoeffding-Azuma bound for martingales by Seldin et al. [161]. A PAC-Bayes Bernstein bound for the CIS estimate can also be derived in the general case where the data are drawn from an arbitrary sequence of behaviour policies. We define the average variance of the CIS estimate as

$$V^{\mathrm{CIS}}(\pi, D_T) = \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\substack{a'_t\sim b_t \\ r'_t\sim\mathbb{P}_R(\cdot|a'_t)}} \left[\left(\min\left(\frac{\pi(a'_t)}{b_t(a'_t)}, \frac{1}{\tau}\right) r'_t - \mathop{\mathbb{E}}_{\substack{a'_t\sim b_t \\ r'_t\sim\mathbb{P}_R(\cdot|a'_t)}} \left[\min\left(\frac{\pi(a'_t)}{b_t(a'_t)}, \frac{1}{\tau}\right) r'_t\right]\right)^2\right], \tag{3.12}$$

One can show (see Lemma A.4) that the average variance of the CIS estimate satisfies $V^{\mathrm{CIS}}(\pi, D_T) \leq 1/\tau$. In Appendix A.1.5, we prove the following PAC-Bayes Bernstein bound for the CIS estimate.

Theorem 3.8 (PAC-Bayes Bernstein Bound for r^{CIS})**.** *For any $\tau \in (0, 1]$, any $\lambda \in [0, T\tau]$, any $\delta \in (0, 1)$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1 - \delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously,*

$$R(Q) \geq r^{\mathrm{CIS}}(Q, D_T) - \frac{\lambda(e-2)V^{\mathrm{CIS}}(Q, D_T)}{T} - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda}.$$

Applying the variance bound $V^{\mathrm{CIS}}(\pi, D_T) \leq 1/\tau$ gives

$$R(Q) \geq r^{\mathrm{CIS}}(Q, D_T) - \frac{\lambda(e-2)}{T\tau} - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda}. \tag{3.13}$$

This bound is essentially a corollary of the PAC-Bayes Bernstein bound for martingales by Seldin et al. [161]. A PAC-Bayes kl bound for the CIS estimate has so far only been proven for the case where the data are all drawn from a fixed behaviour policy. In this case, the CIS estimate is a sum of independent random variables bounded in $[0, 1/\tau]$. Therefore, one can apply Seeger's original PAC-Bayes kl bound [158] to the CIS estimate, scaled by a factor of τ.

Theorem 3.9 (PAC-Bayes kl bound for r^{CIS})**.** *If the data set D_T is drawn from a single, fixed behaviour policy, then for any $\tau \in (0, 1]$, any $\delta \in (0, 1)$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1 - \delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously,*

$$kl\big(\tau r^{\mathrm{CIS}}(Q, D_T) \,\big\|\, \tau R^{\mathrm{CIS}}(Q)\big) \leq \frac{D_{\mathrm{KL}}(Q||P) + \ln(2\sqrt{T}/\delta)}{T}.$$

Since $R(Q) \geq R^{\mathrm{CIS}}(Q)$, one can still use Pinsker's inequality or the inverse of kl to obtain lower bounds on $R(Q)$. If we invert the Binary KL divergence, we obtain

$$R(Q) \geq \frac{1}{\tau} kl^{-1}\left(\tau r^{\mathrm{CIS}}(Q, D_T), \frac{D_{\mathrm{KL}}(Q||P) + \ln(2\sqrt{T}/\delta)}{T}\right). \tag{3.14}$$

If we use Pinsker's inequality, we obtain:

$$R(Q) \geq r^{\mathrm{CIS}}(Q, D_T) - \frac{1}{\tau}\sqrt{\frac{D_{\mathrm{KL}}(Q||P) + \ln(2\sqrt{T}/\delta)}{2T}}. \tag{3.15}$$

The discussion about rates in Section 3.4.1 applies to the bounds for the CIS estimate, although the rates in ϵ_T are now rates in τ. The PAC-Bayes Bernstein and kl^{-1} bounds both have improved rates in τ and should therefore be preferred when τ is close to 0.

Next, we will discuss PAC-Bayes bounds for the CIS estimate in the contextual bandit setting. In the CB setting, the CIS estimate is defined as

$$r^{\mathrm{CIS}}(\pi, D_T) = \frac{1}{T}\sum_{t=1}^{T} \min\left(\frac{\pi(a_t|s_t)}{b_t(a_t|s_t)}, \frac{1}{\tau}\right) r_t. \tag{3.16}$$

As in the MAB setting, the CIS estimate is biased to underestimate $R(Q)$ (meaning $R^{\mathrm{CIS}}(Q) \leq R(Q)$), so we can still essentially ignore it in the one-sided PAC-Bayes reward bounds. The PAC-Bayes Hoeffding-Azuma and Bernstein bounds for the CIS estimate, in Theorem 3.7 and Theorem 3.8, also hold in the CB setting. When there is a fixed behaviour policy, the CB CIS estimate is still an average of i.i.d. random variables bounded in $[0, 1/\tau]$, so the PAC-Bayes kl bound in Theorem 3.9 also holds in the CB setting.

3.4.3 Weighted Importance Sampling

We present PAC-Bayes reward bounds for the weighted (or self-normalised) importance sampling (WIS) estimator. The PAC-Bayes bounds presented in this section are only valid when the data

are i.i.d.; for example, when the data set is drawn from a fixed behaviour policy. For the rest of Section 3.4.3, we assume that this is the case. On the bright side, the bounds in this section do not require the importance weights to be bounded or clipped. For MAB problems, the WIS estimator can be defined as

$$r^{\mathrm{WIS}}(\pi, D_T) = \frac{\sum_{t=1}^{T} \frac{\pi(a_t)}{b(a_t)} r_t}{\sum_{t=1}^{T} \frac{\pi(a_t)}{b(a_t)}}.$$

The WIS estimate has some pleasing properties. Firstly, when the rewards are bounded between 0 and 1, it always takes values in the range $[0, 1]$, even when the importance weights $\pi(a_t)/b(a_t)$ are unbounded. Secondly, it is invariant to constant shifts in the importance weights. Therefore, we only need to know the unnormalised probability mass/density functions of the policies π and b.

The WIS estimator is biased but consistent, meaning its bias decays to 0 as T tends to infinity. Liu [113] shows that the bias decays to 0 with rate $\mathcal{O}(1/T)$, so we can expect it to be close to 0 as long as T is reasonably large. One can obtain PAC-Bayes bounds on the difference between $r^{\mathrm{WIS}}(Q, D_T)$ and $R(Q)$ by upper bounding both terms in the following bias-concentration decomposition.

$$R(Q) - r^{\mathrm{WIS}}(Q, D_T) = \underbrace{R(Q) - R^{\mathrm{WIS}}(Q)}_{\text{bias}} + \underbrace{R^{\mathrm{WIS}}(Q) - r^{\mathrm{WIS}}(Q, D_T)}_{\text{concentration}}. \tag{3.17}$$

We are not aware of any empirical upper bounds on this bias term that don't require additional assumptions on the reward distribution $\mathbb{P}_R$. Kuzborskij et al. [98] proved a bound on the bias term, although it only holds when the rewards are one-hot; there is always one action with reward 1 and all remaining actions have reward 0.

The concentration term can be bounded using PAC-Bayes bounds. Since the WIS estimate is not a sum of i.i.d. random variables or even the sum of a martingale difference sequence, we cannot use any of the previously seen PAC-Bayes bounds to bound the concentration term. Kuzborskij and Szepesvári [97] derived a very general Efron-Stein (ES) PAC-Bayes bound and showed that it can be used to upper bound the concentration term in Equation 3.17. This bound contains the semi-empirical ES variance proxy of the WIS estimate. For any real-valued function $f(\pi, D_T)$, the corresponding semi-empirical ES variance proxy is defined as

$$V^{\mathrm{ES}}(\pi, D_T) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{D_T, D'_T} \left[\left(f(\pi, D_T) - f(\pi, D_T^{(t)}) \right)^2 \middle| D_t \right]. \tag{3.18}$$

D'_T is an independently sampled copy of D_T. $D_T^{(t)}$ is the data set D_T, except the t^{th} element is replaced with the t^{th} element of D'_T. For example, in the MAB setting, (a_t, r_t) is replaced by an independent copy (a'_t, r'_t). This variance proxy is semi-empirical since it depends on both the observed data and the distribution of the data. Kuzborskij and Szepesvári [97] derived a PAC-Bayes bound on the absolute difference between $f(Q, D_T)$ and its expected value $F(Q) = \mathbb{E}_{D_T}[f(Q, D_T)]$. When $f = r^{\mathrm{WIS}}$, we obtain the following result.

Theorem 3.10 (Efron-Stein PAC-Bayes Bound for r^{WIS} [97])**.** *If the data set D_T is drawn from a single, fixed behaviour policy, then for any $y > 0$, any $\delta \in (0, 1)$ and any probability distribution*

$P \in \mathcal{P}(\Pi)$, with probability at least $1-\delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously, we have

$$\left|r^{\mathrm{WIS}}(Q, D_T) - R^{\mathrm{WIS}}(Q)\right| \leq \sqrt{2\left(y + V^{\mathrm{ES}}(Q, D_T)\right)\left(D_{\mathrm{KL}}(Q||P) + \frac{1}{2}\ln\left(1 + V^{\mathrm{ES}}(Q, D_T)/y\right) + \ln(1/\delta)\right)}.$$

Theorem 3.10 is actually a slightly tighter version of the second inequality in Theorem 3 of [97], which holds under weaker assumptions. In the original bound, the factor of $1/2$ in front of $\ln\left(1 + V^{\mathrm{ES}}(Q, D_T)/y\right)$ is replaced with $\ln(1/\delta)/2$, which is larger than $1/2$ when $\delta \leq e^{-1}$. The original bound of Kuzborskij and Szepesvári is only valid when $\delta \leq e^{-2}$, so the bound in Theorem 3.10 is always slightly tighter. Moreover, this bound holds simultaneously for all distributions Q, whereas in the original bound of Kuzborskij and Szepesvári, Q must be given by a fixed probability kernel that maps any data set D_T to a distribution ρ_{D_T}. We provide a proof of Theorem 3.10 for general functions $f(\pi, D_T)$ in Appendix A.1.6.

The semi-empirical ES variance proxy for $r^{\mathrm{WIS}}(\pi, D_T)$ depends on the unknown reward distribution $\mathbb{P}_R$, which means that $V^{\mathrm{ES}}(Q, D_T)$ is also unknown. However, Kuzborskij and Szepesvári [97] show that it can be upper bounded by a quantity that can be computed without knowledge of $\mathbb{P}_R$.

Lemma 3.11 (r^{WIS} ES Variance Proxy Upper Bound [97])**.** *For $f = r^{\mathrm{WIS}}$ and any $\pi \in \Pi$, we have that:*

$$V^{\mathrm{ES}}(\pi, D_T) \leq 2V^{\mathrm{WIS}}(\pi, D_T) = 2\sum_{t=1}^{T} \mathop{\mathbb{E}}_{D_T, D_T'}\left[\tilde{w}_{\pi,t}^2 + \tilde{u}_{\pi,t}^2 | D_t\right],$$

where

$$\tilde{w}_{\pi,t} = \frac{\frac{\pi(a_t)}{b(a_t)}}{\sum_{k=1}^{T}\frac{\pi(a_k)}{b(a_k)}}, \qquad \tilde{u}_{\pi,i} = \frac{\frac{\pi(a_t')}{b(a_t')}}{\frac{\pi(a_t')}{b(a_t')} + \sum_{k\neq t}\frac{\pi(a_k)}{b(a_k)}}.$$

Though $V^{\mathrm{WIS}}(\pi, D_T)$ is still semi-empirical, it does not depend on the reward distribution $\mathbb{P}_R$. Therefore, it can be estimated with arbitrary accuracy if π and b are known. We can combine the bias-concentration decomposition in Equation 3.17, the Efron-Stein PAC-Bayes bound in Theorem 3.10 and the bound on the ES variance proxy in Lemma 3.11 to obtain the following PAC-Bayes bound on the expected reward, which holds with probability greater than $1-\delta$ and for all $Q \in \mathcal{P}(\Pi)$ simultaneously.

$$\begin{aligned} R(Q) \geq\ & r^{\mathrm{WIS}}(Q, D_T) - \left|R^{\mathrm{WIS}}(Q) - R(Q)\right| \\ & - \sqrt{2\left(y + 2V^{\mathrm{WIS}}(Q, D_T)\right)\left(D_{\mathrm{KL}}(Q||P) + \frac{1}{2}\ln\left(1 + \frac{2V^{\mathrm{WIS}}(Q, D_T)}{y}\right) + \ln(1/\delta)\right)}. \end{aligned} \tag{3.19}$$

In order to use this bound, we would need to upper bound the bias term $\left|R^{\mathrm{WIS}}(Q) - R(Q)\right|$. Ignoring the bias term, the rate of this bound in T depends on the values of $V^{\mathrm{WIS}}(Q, D_T)$ and y. At one extreme, when all policies in the support of Q result in approximately equal importance weights for every action, we have $V^{\mathrm{WIS}}(Q, D_T) = \mathcal{O}(1/T)$. At the other extreme, when policies in the support of Q result in one importance weight dominating all the others, we have $V^{\mathrm{WIS}}(Q, D_T) =$

$\mathcal{O}(1)$. Therefore, $V^{\mathrm{WIS}}(Q, D_T) = \mathcal{O}(1/T^\alpha)$ for some $\alpha \in [0, 1]$. If we choose $y = \mathcal{O}(1/T^\alpha)$, then the bound in Equation 3.19 has rate $\mathcal{O}(\sqrt{D_{\mathrm{KL}}(Q||P)}/T^{\alpha/2})$.

Now, we discuss the ES PAC-Bayes bound for the WIS estimate in the contextual bandit setting. In the CB setting, the WIS estimate can be defined as

$$r^{\mathrm{WIS}}(\pi, D_T) = \frac{\sum_{t=1}^T \frac{\pi(a_t|s_t)}{b(a_t|s_t)} r_t}{\sum_{t=1}^T \frac{\pi(a_t|s_t)}{b(a_t|s_t)}}.$$

The ES PAC-Bayes bound in Theorem 3.10 can still be used and one can derive an equivalent to the bound in Equation 3.19. However, the upper bound on the semi-empirical ES variance proxy $V^{\mathrm{WIS}}(\pi, D_T)$, as defined in Lemma 3.11, now depends on the unknown state distribution $\mathbb{P}_S$. To rectify this, one can use an alternative bias-concentration decomposition suggested by Kuzborskij et al. [98]:

$$\begin{aligned} R(Q) - r^{\mathrm{WIS}}(Q, D_T) = &\underbrace{R(Q) - R(Q; s_{1:T})}_{\text{concentration of contexts}} + \underbrace{R(Q; s_{1:T}) - R^{\mathrm{WIS}}(Q; s_{1:T})}_{\text{bias}} \\ &+ \underbrace{R^{\mathrm{WIS}}(Q; s_{1:T}) - r^{\mathrm{WIS}}(Q, D_T)}_{\text{concentration}}, \end{aligned}$$

where

$$R(Q; s_{1:T}) = \mathop{\mathbb{E}}_{\pi\sim Q}\left[\frac{1}{T}\sum_{t=1}^T \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s_t), r\sim\mathbb{P}_R(\cdot|s_t,a)}[r]\right], \; R^{\mathrm{WIS}}(Q; s_{1:T}) = \mathop{\mathbb{E}}_{\pi\sim Q}\left[\mathop{\mathbb{E}}_{D_T}\left[r^{\mathrm{WIS}}(\pi, D_T)|s_1, \ldots, s_T\right]\right].$$

The concentration of contexts term can be bounded by, for example, the PAC-Bayes Hoeffding-Azuma bound. The concentration term can be bounded using a conditional version of the Efron-Stein PAC-Bayes bound in Theorem 3.10, which holds with high probability over the sampling of D_T given the observed states $s_1, \ldots, s_T$. The upper bound $2V^{\mathrm{WIS}}(\pi, D_T)$ on the ES variance proxy, given the observed states, no longer depends on the state distribution, so it can be estimated with knowledge of only π and b. Finally, we note that replacing r^{WIS} in Theorem 3.10 with r^{IS} or r^{CIS} would lead to new ES PAC-Bayes bounds for the IS or CIS estimates. However, this has not yet been explored in the literature.

3.5 PAC-Bayes Regret Bounds

In this section, we review PAC-Bayes bounds for the cumulative regret $\Delta(Q_{1:T})$ associated with a sequence of distributions $Q_1, \ldots, Q_T$ over the policy class Π. First, we state some PAC-Bayes bounds on the expected regret for a single round. Then, we present some PAC-Bayes bounds on the cumulative regret.

In the MAB setting, we consider the case when the set of actions is finite: $\mathcal{A} = \{1, \ldots, K\}$. We set the policy class to be the set of all deterministic policies, so $\Pi = \mathcal{A}$. In this case, any distribution Q

over the policy class $\mathcal{A}$ is a single stochastic policy. The IS estimate of the reward for a deterministic policy (an action) a can be defined as

$$r^{\mathrm{IS}}(a, D_T) = \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{I}\{a_t = a\}}{b_t(a_t)} r_t. \tag{3.20}$$

This coincides with the earlier definition for general policy classes. For this choice of policy class, a uniform upper bound on the importance weights $\mathbb{I}\{a_t = a\}/b_t(a_t) \leq 1/\epsilon_T$ for all $t = 1, \ldots, T$ can be achieved with a uniform lower bound on the behaviour policy probabilities $b_t(a_t)$, i.e. $b_t(a_t) \geq \epsilon_T$. The regret for an action a is defined as

$$\Delta(a) = R(a^*) - R(a),$$

where a^* is an action that maximises the expected reward. The IS regret estimate for an action a is defined as

$$\Delta^{\mathrm{IS}}(a, D_T) = r^{\mathrm{IS}}(a^*, D_T) - r^{\mathrm{IS}}(a, D_T). \tag{3.21}$$

Seldin et al. [162] [160] showed that a martingale, which is compatible with both the Hoeffding-Azuma inequality and Bernstein's inequality, can be constructed from the IS regret estimate. Consequently, one can obtain a PAC-Bayes Hoeffding-Azuma bound and a PAC-Bayes Bernstein bound on the difference between the expected regret and the IS regret estimate.

Theorem 3.12 (PAC-Bayes Hoeffding-Azuma bound for Δ^{IS} [162, 43])**.** *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all distributions $Q \in \mathcal{P}(\mathcal{A})$ and all $t \geq 1$ simultaneously:*

$$\Delta(Q) - \Delta^{\mathrm{IS}}(Q, D_t) \leq \sqrt{\frac{2\ln(K/\delta)}{t\epsilon_t^2}}.$$

One can observe several differences between this bound and the PAC-Bayes Hoeffding-Azuma bound in Theorem 3.2. This bound uses a uniform prior P, and since both Q and P are distributions over a finite set with K elements, $D_{\mathrm{KL}}(Q||P) \leq \ln(K)$. Hence, the KL divergence has been replaced with $\ln(K)$ (and then added to $\ln(1/\delta)$). This bound holds with probability at least $1 - \delta$ for all $t \geq 1$ simultaneously, rather than for a single $t \geq 1$. This is achieved by using the time-uniform extension of Seldin et al.'s [162] original PAC-Bayes Hoeffding-Azuma bound, which can be found in Corollary 6.5 of [43]. Finally, this bound does not contain λ. This is because, for each t, we have set $\lambda_t = \sqrt{2t\epsilon_t^2 \ln(K/\delta)}$.

The PAC-Bayes Bernstein bound for the IS regret estimate contains (an upper bound on) the average variance of the IS regret estimate $V^{\mathrm{IS}}(a, D_T)$, which is defined as:

$$V^{\mathrm{IS}}(a, D_T) = \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{a'_t \sim b_t, r'_t \sim p_R(\cdot|a'_t)} \left[\left(\frac{\mathbb{I}\{a'_t = a^*\}}{b_t(a'_t)} r'_t - \frac{\mathbb{I}\{a'_t = a\}}{b(a'_t)} r'_t - \Delta(a) \right)^2 \right],$$

Seldin et al. show that in both the MAB [160] and CB settings [159], this average variance satisfies the bound $V^{\mathrm{IS}}(a, D_t) \leq 2/\epsilon_t$. Using this bound on the variance, the following PAC-Bayes bound can be derived.

Theorem 3.13 (PAC-Bayes Bernstein bound for Δ^{IS} [160, 43])**.** *For any $\delta \in (0,1)$, with probability at least $1-\delta$, for all distributions $Q \in \mathcal{P}(\mathcal{A})$ and all $t \geq 1$ simultaneously, where*

$$\frac{\ln(K/\delta)}{2(e-2)t} \leq \epsilon_t, \tag{3.22}$$

we have

$$\Delta(Q) \leq \Delta^{\mathrm{IS}}(Q, D_t) + \sqrt{\frac{8(e-2)\ln(K/\delta)}{\epsilon_t t}}.$$

This bound comes from the time-uniform extension (see Corollary 6.6 of [43]) of Seldin et al.'s PAC-Bayes Bernstein bound [160]. In this bound we have set $\lambda_t = \sqrt{t\epsilon_t \ln(K/\delta)/(2(e-2))}$. The requirement that $\lambda_t \in [0, t\epsilon_t]$ becomes the requirement on t in Equation 3.22. Following Seldin et al. [159], [160], we present cumulative regret bounds for a family of MAB algorithms. We define

$$Q_t^{\exp}(a) \propto P(a) e^{\gamma_t r^{\mathrm{IS}}(a, D_t)}, \quad \tilde{Q}_t^{\exp}(a) = (1 - K\epsilon_{t+1}) Q_t^{\exp}(a) + \epsilon_{t+1}. \tag{3.23}$$

If $\epsilon_t = \min(1/\sqrt{tK}, 1/K)$ and $\gamma_t = \sqrt{t\ln(K)/K}$, this strategy is known as EXP3 [22]. Alternatively, in the limit as γ_t tends to infinity, we obtain the ϵ-greedy algorithm [21]. Note that ϵ_t cannot be greater than $1/K$, which is why we truncate it to be less than or equal to $1/K$. The first step towards a cumulative regret bound for the policies $\tilde{Q}_1^{\exp}, \tilde{Q}_2^{\exp}, \ldots$ is to re-write the regret for a single round as

$$\Delta(\tilde{Q}_t^{\exp}) = \Delta(Q_t^{\exp}) - \Delta^{\mathrm{IS}}(Q_t^{\exp}, D_t) + \Delta^{\mathrm{IS}}(Q_t^{\exp}, D_t) + R(Q_t^{\exp}) - R(\tilde{Q}_t^{\exp}). \tag{3.24}$$

Seldin et al. [160] show that $\Delta^{\mathrm{IS}}(Q_t^{\exp}, D_t) \leq \ln(K)/\gamma_t$ and that $R(Q_t^{\exp}) - R(\tilde{Q}_t^{\exp}) \leq K\epsilon_{t+1}$. If the PAC-Bayes Hoeffding-Azuma bound in Theorem 3.12 is used to bound $\Delta(Q_t^{\exp}) - \Delta^{\mathrm{IS}}(Q_t^{\exp}, D_t)$, then we obtain the following cumulative regret bound.

Theorem 3.14 (PAC-Bayes Hoeffding-Azuma cumulative regret bound [162], [160])**.** *Let $\epsilon_t = \min(t^{-1/4}K^{-1/2}, 1/K)$ and take any γ_t such that $\gamma_t \geq t^{1/4}K^{-1/2}\sqrt{\ln(K)}$. For any $\delta \in (0,1]$, with probability at least $1-\delta$, for all $T > K^2$, we have*

$$\sum_{t=1}^{T} \Delta(\tilde{Q}_t^{\exp}) \leq \frac{4}{3}T^{3/4}K^{1/2}(\sqrt{2\ln(K/\delta)} + \sqrt{\ln(K)} + 1) - \frac{4}{3}K^2(\sqrt{2\ln(K/\delta)} + \sqrt{\ln(K)} + \frac{1}{4}), \tag{3.25}$$

and for all $1 \leq T \leq K^2$, we have

$$\sum_{t=1}^{T} \Delta(\tilde{Q}_t^{\exp}) \leq T.$$

Proof. From the assumption that the rewards are bounded between 0 and 1, we have $\Delta(Q) \leq 1$ for all $Q \in \mathcal{P}(\Pi)$. Whenever $t^{-1/4}K^{-1/2} \leq 1/K$, we have $\epsilon_t = t^{-1/4}K^{-1/2}$. By rearranging this

inequality, we can deduce that $\epsilon_t = t^{-1/4}K^{-1/2}$ when $t \geq K^2$. Now, for all $T > K^2$, we have

$$\begin{aligned}
\sum_{t=1}^{T} \Delta(\tilde{Q}_t^{\exp}) &= \sum_{t=1}^{K^2} \Delta(\tilde{Q}_t^{\exp}) + \sum_{t=K^2+1}^{T} \Delta(\tilde{Q}_t^{\exp}) \\
&\leq K^2 + \sum_{t=K^2+1}^{T} \left(\Delta(Q_t^{\exp}) - \Delta^{\mathrm{IS}}(Q_t^{\exp}, D_t) + \Delta^{\mathrm{IS}}(Q_t^{\exp}, D_t) + R(Q_t^{\exp}) - R(\tilde{Q}_t^{\exp}) \right) \\
&\leq K^2 + \sum_{t=K^2+1}^{T} \left(\sqrt{\frac{2\ln(K/\delta)}{t\epsilon_t^2}} + \frac{\ln(K)}{\gamma_t} + K\epsilon_{t+1} \right) \\
&\leq K^2 + \sum_{t=K^2+1}^{T} t^{-1/4}K^{1/2}(\sqrt{2\ln(K/\delta)} + \sqrt{\ln(K)} + 1) \\
&\leq K^2 + K^{1/2}(\sqrt{2\ln(K/\delta)} + \sqrt{\ln(K)} + 1) \int_{K^2}^{T} t^{-1/4}\mathrm{d}t \\
&= K^2 + K^{1/2}(\sqrt{2\ln(K/\delta)} + \sqrt{\ln(K)} + 1)\frac{4}{3}(T^{3/4} - K^{3/2}) \\
&= \frac{4}{3}T^{3/4}K^{1/2}(\sqrt{2\ln(K/\delta)} + \sqrt{\ln(K)} + 1) - \frac{4}{3}K^2(\sqrt{2\ln(K/\delta)} + \sqrt{\ln(K)} + 1/4).
\end{aligned}$$

We can upper bound the sum $\sum_{t=K^2+1}^{T} t^{-1/4}$ by the integral $\int_{K^2}^{T} t^{-1/4}\mathrm{d}t$ because $t^{-1/4}$ is monotonically decreasing in t. □

This cumulative regret bound is of order $\mathcal{O}(T^{3/4}K^{1/2}\ln(K)^{1/2})$. If the PAC-Bayes Bernstein bound in Theorem 3.13 is used to bound $\Delta(\rho_n^{\exp}) - \Delta^{\mathrm{IS}}(\rho_n^{\exp}, D_n)$, then we obtain the following cumulative regret bound.

Theorem 3.15 (PAC-Bayes Bernstein cumulative regret bound [160])**.** *Let* $\epsilon_t = \min(t^{-1/3}K^{-2/3}, 1/K)$ *and take any* γ_t *such that* $\gamma_t \geq t^{1/3}K^{-1/3}\sqrt{\ln(K)}$*. Define*

$$\widetilde{T} = \left\lceil K \left(\frac{\ln(K/\delta)}{2(e-2)} \right)^{3/2} \right\rceil.$$

For any $\delta \in (0, 1]$*, with probability at least* $1 - \delta$*, for all* $T > \widetilde{T}$*, we have*

$$\begin{aligned}
\sum_{t=1}^{T} \Delta(\tilde{Q}_t^{\exp}) &\leq \frac{3}{2}(T^{2/3} - \widetilde{T}^{2/3})K^{1/3}(\sqrt{8(e-2)\ln(K/\delta)} + \sqrt{\ln(K)} + 1) + \widetilde{T}, \\
&\leq \frac{3}{2}T^{2/3}K^{1/3}(\sqrt{8(e-2)\ln(K/\delta)} + \sqrt{\ln(K)} + 1), \qquad (3.26)
\end{aligned}$$

and for all $1 \leq T \leq \widetilde{T}$*, we have*

$$\sum_{t=1}^{T} \Delta(\tilde{Q}_t^{\exp}) \leq T.$$

The proof of this regret bound is very similar to the proof of Theorem 3.14, so we omit it. The definition of $\widetilde{T}$ and the condition $T > \widetilde{T}$ come from Equation (3.22) and the choice of ϵ_t. This cumulative regret bound is of order $\mathcal{O}(T^{2/3}K^{1/3}\ln(K)^{1/2})$. The improved scaling with T and K is due to the PAC-Bayes Bernstein bound having improved dependence on ϵ_t. Sadly, this regret bound has a sub-optimal growth rate in T. Audibert and Bubeck [19] have shown that the cumulative regret for EXP3 can be upper bounded by a term of order $\mathcal{O}(\sqrt{TK\ln(K)})$. Moreover, Audibert and Bubeck show that the best possible worst-case regret bound that any algorithm can achieve is $\mathcal{O}(\sqrt{TK})$.

Seldin et al. [160] hypothesise that the PAC-Bayes cumulative regret bound in Theorem 3.15 can be improved for the EXP3 algorithm with $\epsilon_t = \min(1/\sqrt{tK}, 1/K)$ and $\gamma_t = \sqrt{t\ln(K)/K}$. They suggest, and verify empirically, that for this choice of ϵ_t and γ_t, the bound on the average variance can be tightened to $V^{\mathrm{IS}}(\tilde{Q}_t^{\exp}, D_t) \leq 2K$. Using this bound on the average variance, the cumulative regret bound in Theorem 3.15 would become $\mathcal{O}(\sqrt{TK\ln(K)})$.

To conclude this section, we present a PAC-Bayes cumulative regret bound for contextual bandits by Seldin et al. [159]. We consider the case when the set of actions is finite with K elements ($\mathcal{A} = \{1, \ldots, K\}$) and the set of states is finite with N elements ($\mathcal{S} = \{1, \ldots, N\}$). The policy class Π is the set of all deterministic policies, which for this problem is the set of all functions from $\mathcal{S}$ to $\mathcal{A}$, of which there are K^N. For any distribution Q over Π, there is a corresponding stochastic policy $Q(a|s)$, where

$$Q(a|s) = \mathop{\mathbb{E}}_{\pi\sim Q}\left[\mathbb{I}\{\pi(s) = a\}\right].$$

In this setting, the IS reward estimate for a single policy $\pi \in \Pi$ can be defined as

$$r^{\mathrm{IS}}(\pi, D_T) = \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{I}\{a_t = \pi(s_t)\}}{b_t(a_t|s_t)}r_i. \tag{3.27}$$

A lower bound $b_t(a|s) \geq \epsilon_t$ for all s, a ensures that the importance weights are upper bounded by $1/\epsilon_t$. Let $T(s) = \sum_{t=1}^{T}\mathbb{I}\{s_t = s\}$ denote the number of times that state s appears in the data set D_T. We define the IS reward estimate for a single state and action as

$$r^{\mathrm{IS}}(s, a, D_T) = \frac{1}{T(s)}\sum_{t=1,\ldots,T:s_t=s}\frac{\mathbb{I}\{a_t = a\}}{b_t(a_t|s_t)}r_t. \tag{3.28}$$

If $T(s) = 0$, then $r^{\mathrm{IS}}(s, a, D_T) = 0$. The expected regret for a policy π can be defined as

$$\Delta(\pi) = R(\pi^*) - R(\pi),$$

where π^* is a policy in Π that maximises the expected reward. The IS regret estimate for a policy π is defined as:

$$\Delta^{\mathrm{IS}}(\pi, D_T) = r^{\mathrm{IS}}(\pi^*, D_T) - r^{\mathrm{IS}}(\pi, D_T). \tag{3.29}$$

Seldin et al. [159] show that, as in the MAB setting, a martingale compatible with Bernstein's inequality can be constructed from the CB IS regret estimate. Moreover, the average variance of the CB IS regret estimate can also be bounded by $2/\epsilon_T$. Seldin et al. [159] obtain a PAC-Bayes Bernstein bound on the difference between the expected regret and the IS regret estimate.

Theorem 3.16 (CB PAC-Bayes Bernstein bound for Δ^{IS} [159])**.** *For any $\delta \in (0, 1]$ and any $c > 1$, simultaneously for all $Q \in \mathcal{P}(\Pi)$ that satisfy*

$$\frac{NI_Q(S;A) + K(\ln(N) + \ln(K)) + \ln(m_T/\delta)}{2(e-2)T} \leq \frac{\epsilon_T}{c^2}, \tag{3.30}$$

with probability at least $1 - \delta$,

$$\Delta(Q) \leq \Delta^{\text{IS}}(Q, D_T) + (1 + c)\sqrt{\frac{2(e-2)\left(NI_Q(S;A) + K(\ln(N) + \ln(K)) + \ln(m_T/\delta)\right)}{T\epsilon_T}}, \tag{3.31}$$

where $m_T = \ln(\sqrt{(e-2)T/\ln(1/\delta)})/\ln(c)$, and for all Q that do not satisfy (3.30), with the same probability,

$$\Delta(Q) \leq \Delta^{\text{IS}}(Q, D_T) + \frac{2\left(NI_Q(S;A) + K(\ln(N) + \ln(K)) + \ln(m_T/\delta)\right)}{T\epsilon_T}. \tag{3.32}$$

In this PAC-Bayes Bernstein bound, the KL divergence penalty has been replaced with $I_Q(S;A)$, which is the mutual information between states and actions under the policy $Q(a|s)$. Let $\bar{Q}(a) = (1/N)\sum_s Q(a|s)$ denote the marginal distribution over $\mathcal{A}$ that corresponds to $Q(a|s)$ and a uniform distribution over $\mathcal{S}$. Then $I_Q(S;A)$ is defined as

$$I_Q(S;A) = \frac{1}{N}\sum_{s,a} Q(a|s)\ln\left(Q(a|s)/\bar{Q}(a)\right). \tag{3.33}$$

As shown by Seldin and Tishby [164], there exists a distribution P over Π such that for every Q over Π, we have

$$D_{\text{KL}}(Q||P) \leq NI_Q(S;A) + K\ln(N) + K\ln(K).$$

We could have also chosen P to be a uniform prior, in which case $D_{\text{KL}}(Q||P) \leq N\ln(K)$. However, $I_Q(S;A) \leq \ln(K)$, so when the number of states N is much larger than the number of actions K, we have $NI_Q(S;A) + K\ln(N) + K\ln(K) \leq N\ln(K)$. Seldin et al. [159] derive a cumulative regret bound for a family of contextual bandit algorithms. Let $Q_t(a)$ be an arbitrary distribution over $\mathcal{A}$ (for each t). We define

$$Q_t^{\exp}(a|s) \propto Q_t(a)e^{\gamma_t r^{\text{IS}}(s,a,D_t)}, \quad \tilde{Q}_t^{\exp}(a|s) = (1 - K\epsilon_{t+1})Q_t^{\exp}(a|s) + \epsilon_{t+1}. \tag{3.34}$$

Using the same regret decomposition as in Equation 3.24, one can obtain a per-round regret bound for playing $\tilde{Q}_t^{\exp}$. Seldin et al. [159] show that $\Delta^{\text{IS}}(Q_t^{\exp}, D_t) \leq \ln(1/\epsilon_{t+1})/\gamma_t$, and that $R(Q_t^{\exp}) - R(\tilde{Q}_t^{\exp}) \leq K\epsilon_{t+1}$ also holds in the CB setting. If the PAC-Bayes Bernstein bound from Theorem 3.16 is used to bound $\Delta(Q_t^{\exp}) - \Delta^{\text{IS}}(Q_t^{\exp}, D_t)$, then we obtain the following per-round regret bound.

Theorem 3.17 (CB PAC-Bayes Bernstein per-round regret bound [159])**.** *For any $\delta \in (0,1]$ and any $c > 1$, with probability at least $1-\delta$, for all policies Q^{exp} that satisfy Equation 3.30, the expected per-round regret $\Delta(\tilde{Q}_t^{\text{exp}})$ is bounded by*

$$\Delta(\tilde{Q}_t^{\text{exp}}) \leq (1+c)\sqrt{\frac{2(e-2)(NI_{Q_t^{\text{exp}}}(S;A) + K(\ln(N)+\ln(K)) + \ln(2m_t/\delta))}{t\epsilon_t}} + \frac{\ln(\epsilon_{t+1})}{\gamma_t} + K\epsilon_{t+1},$$

and for all Q^{exp} that do not satisfy Equation 3.30, with the same probability

$$\Delta(\tilde{Q}_t^{\text{exp}}) \leq \frac{2NI_{Q_t^{\text{exp}}}(S;A) + K(\ln(N)+\ln(K)) + \ln(2m_t/\delta)}{t\epsilon_t} + \frac{\ln(\epsilon_{t+1})}{\gamma_t} + K\epsilon_{t+1}.$$

If $\epsilon_t = \min(t^{-1/3}K^{-1/3}N^{1/3}, 1/K)$, then this gives a cumulative regret bound of order $\mathcal{O}(T^{2/3}K^{2/3}N^{1/3})$, ignoring log terms. If we were to upper bound $D_{\text{KL}}(Q||P)$ by $N\ln(K)$ instead of the mutual information, then choosing $\epsilon_t = \min(t^{-1/3}K^{-2/3}N^{1/3}, 1/K)$ would give a cumulative regret bound of order $\mathcal{O}(T^{2/3}K^{1/3}N^{1/3})$ ignoring log terms. Unfortunately, both of these bounds have sub-optimal scaling with T. For example, ignoring log terms, the EXP4.P algorithm of Beygelzimer et al. [28] has cumulative regret bounded by $\mathcal{O}(\sqrt{TKN})$ in this problem.

3.6 Optimising PAC-Bayes Bandit Bounds

3.6.1 The Choice of Prior

In this section, we give an overview of methods for choosing the prior. We first motivate the utility of "good" priors, using the PAC-Bayes Hoeffding-Azuma bound from Thm. 3.2 (shown below) as an example.

$$R(Q) \geq r^{\text{IS}}(Q, D_T) - \frac{\lambda}{8T\epsilon_T^2} - \frac{D_{\text{KL}}(Q||P) + \ln(1/\delta)}{\lambda}.$$

This lower bound is largest when $r^{\text{IS}}(Q, D_T)$ is large and $D_{\text{KL}}(Q||P)$ is close to 0. To achieve this, P must assign high probability to policies where $r^{\text{IS}}(Q, D_T)$ is large. This motivates priors that either depend on the data set D_T (data-dependent priors) or on the distribution of the data set (distribution-dependent priors). In fact, Dziugaite et al. [59] have shown that data-dependent priors are sometimes necessary for tight PAC-Bayes bounds. PAC-Bayes bounds with data/distribution-dependent priors are of practical interest because they can yield tighter performance guarantees. They are also of theoretical interest because they can yield bounds with improved rates.

We now detail various approaches for deriving PAC-Bayes bounds with data/distribution-dependent priors. Many of these techniques are compatible with essentially any PAC-Bayes bound. Where this is the case, we apply them to the PAC-Bayes kl bound for the IS estimate as an example, since we will later compare the PAC-Bayes kl^{-1} bound with various priors in our experiments.

Data-Dependent Priors via Sample Splitting

One way to use a data-dependent prior is to split the data set into disjoint subsets $D_T = D_{1:m} \cup D_{m+1:T}$, of size m and $T-m$, for some $m < T$. The first subset is used to learn a prior $P_{D_{1:m}}$.

A PAC-Bayes bound is then evaluated on the second subset with the learned prior. Since $P_{D_{1:m}}$ does not depend on $D_{m+1:T}$, this prior is a valid choice when the bound is evaluated on the second subset. The PAC-Bayes kl bound with a sample splitting prior is stated below.

Theorem 3.18 (PAC-Bayes kl Bound with Sample Splitting)**.** *For any $\delta \in (0,1)$ and any prior $P_{D_{1:m}} \in \mathcal{P}(\Pi)$ that may depend on the subset $D_{1:m}$, with probability at least $1-\delta$, for all $Q \in \mathcal{P}(\Pi)$ simultaneously*

$$kl\big(\epsilon_T r^{\mathrm{IS}}(Q, D_{m+1:T}) \,\big\|\, \epsilon_T R(Q)\big) \leq \frac{D_{\mathrm{KL}}(Q||P_{D_{1:m}})}{T-m} + \frac{\ln(2\sqrt{T-m}/\delta)}{T-m}.$$

This approach is very flexible, since the data-dependent prior can be learned in any way. We believe that Seeger [158] was the first to use this technique. Subsequently, it has been used by others, such as Catoni [36], Ambroladze et al. [14], and Germain et al. [67]. Recently, this approach has been used to obtain non-vacuous generalisation bounds for deep neural networks [146], [135], [133], [134].

Data-Dependent Priors Selected From a Restricted Set of Priors

Another way to use data-dependent priors is to define a set of priors in advance and then derive a modified PAC-Bayes bound that holds with high probability simultaneously for all priors in this set. One can then evaluate the modified PAC-Bayes bound with any prior from this set. The modified bound will contain an extra penalty that we must pay in order for the bound to hold for more than one prior.

Suppose we have a countable set of priors $\{P_i\}_{i=1}^{\infty}$ and we want the PAC-Bayes kl bound to hold with probability $1-\delta$ for all P_i simultaneously. We have that for each i, with probability at least $1-\delta_i$

$$kl\big(\epsilon_T r^{\mathrm{IS}}(Q, D_T) \,\big\|\, \epsilon_T R(Q)\big) \leq \frac{D_{\mathrm{KL}}(Q||P_i) + \ln(2\sqrt{T}/\delta_i)}{T}.$$

By the union bound, this bound holds with probability at least $1 - \sum_{i=1}^{\infty} \delta_i$ for all Q and all $i \in \mathbb{N}$ simultaneously. We can freely choose $\{\delta_i\}_{i=1}^{\infty}$ such that $\sum_{i=1}^{\infty} \delta_i = \delta$. Therefore, at the cost of replacing δ with δ_i, we can choose the prior in $\{P_i\}_{i=1}^{\infty}$ that results in the greatest lower bound.

This technique has previously been used to obtain parametric priors with data-dependent parameters, e.g. Gaussian priors with data-dependent variance [100], [60]. It can also been derived by using a prior that is a mixture of several priors $P = \sum_{i=1}^{\infty} p_i P_i$ [14], [131]. The weights p_i must satisfy $p_i > 0$ and $\sum_i^{\infty} p_i = 1$. This results in the same bound with $\delta_i = p_i \delta$.

A set of priors can be defined by fixing a learning algorithm and then restricting the choice of prior to be one that is learned from the data using this learning algorithm. If the learning algorithm is stable, meaning the prior it selects is almost unaffected by small changes to the data, then we call the learned prior a stable prior. Dziugaite and Roy [61] and Rivasplata et al. [144] proposed PAC-Bayes bounds with stable priors, where the stability of a prior is characterised by differential privacy.

Let $A : \mathcal{Z}^T \rightsquigarrow \mathcal{P}(\Pi)$ denote a randomised algorithm that maps a data set $D_T \in \mathcal{Z}^T$ to a prior $P \in \mathcal{P}(\Pi)$. Also, let the data set D_T consist of T i.i.d. samples.

Definition 3.19 (Differential privacy)**.** A randomised algorithm $A : \mathcal{Z}^T \rightsquigarrow \mathcal{P}(\Pi)$ is *η-differentially private* if for all pairs $D_T, D'_T \in \mathcal{Z}^T$ that differ at only one coordinate, and all measurable subsets $B \subseteq \mathcal{P}(\Pi)$, we have

$$\mathbb{P}(A(D_T) \in B) \leq e^{\eta}\mathbb{P}(A(D'_T) \in B).$$

Dziugaite and Roy [61] show that any PAC-Bayes bound that holds for any data-independent prior P with probability at least $1 - \delta'$ can be turned into a PAC-Bayes bound that holds for any η-differentially private prior P_{D_T} with probability at least $1 - \delta$ by replacing $D_{\mathrm{KL}}(Q||P)$ with $D_{\mathrm{KL}}(Q||P_{D_T})$ and replacing $\ln(1/\delta')$ with $T\eta^2/2 + \eta\sqrt{T\ln(4/\delta)/2} + \ln(2/\delta)$.

Theorem 3.20 (PAC-Bayes *kl* Bound with a Differentially Private Prior [61])**.** *If the data set D_T is drawn from a single, fixed behaviour policy, then for any $\delta \in (0, 1)$ and any η-differentially private prior $P_{D_T} \in \mathcal{P}(\Pi)$, with probability at least $1 - \delta$, for all $Q \in \mathcal{P}(\Pi)$ simultaneously:*

$$kl\big(\epsilon_T r^{\mathrm{IS}}(Q, D_T) \,\big\|\, \epsilon_T R(Q)\big) \leq \frac{D_{\mathrm{KL}}(Q||P_{D_T}) + \ln(4\sqrt{T}/\delta) + T\eta^2/2 + \eta\sqrt{T\ln(4/\delta)/2}}{T}.$$

Since differential privacy is defined only for data sets consisting of i.i.d. samples, this bound only holds when the data are all drawn from a single, fixed behaviour policy (to ensure that the data are i.i.d.).

Distribution-Dependent Priors

The motivation for using a data-dependent prior was that it would assign high probability to policies where $r^{\mathrm{IS}}(\pi, D_T)$ is large. Assuming $r^{\mathrm{IS}}(\pi, D_T)$ is close to $R(\pi)$, we could instead use a prior that assigns high probability to policies where $R(\pi)$ is large, such as $P(\pi) \propto \exp(R(\pi))$. This prior is data-independent, but we cannot calculate the KL divergence between Q and this prior, since $R(\pi)$ is unknown. Lever et al. [108], [109] provide a method for upper bounding the KL divergence between restricted sets of posteriors and distribution-dependent priors.

We restrict ourselves to the Gibbs distributions $P_{\beta r^{\mathrm{IS}}}$ and $P_{\beta R}$, which are defined as

$$P_{\beta r^{\mathrm{IS}}}(\pi) = \frac{P(\pi)e^{\beta r^{\mathrm{IS}}(\pi, D_T)}}{\mathbb{E}_{\pi\sim P}\left[e^{\beta r^{\mathrm{IS}}(\pi, D_T)}\right]}, \quad P_{\beta R}(\pi) = \frac{P(\pi)e^{\beta R(\pi)}}{\mathbb{E}_{\pi\sim P}\left[e^{\beta R(\pi)}\right]}. \tag{3.35}$$

where $\beta > 0$ and P is any data-independent reference distribution. Lever et al. [108] show that

$$D_{\mathrm{KL}}(P_{\beta r^{\mathrm{IS}}}||P_{\beta R}) \leq \beta\left(r^{\mathrm{IS}}(P_{\beta r^{\mathrm{IS}}}, D_T) - R(P_{\beta r^{\mathrm{IS}}}) + R(P_{\beta R}) - r^{\mathrm{IS}}(P_{\beta R}, D_T)\right). \tag{3.36}$$

Both expected values on the right-hand-side of Equation 3.36 can be upper bounded using any of the PAC-Bayes bounds for the IS reward estimate, with $P_{\beta R}$ as the prior. If the Pinsker bound is used, this results in a quadratic inequality for $D_{\mathrm{KL}}(P_{\beta r^{\mathrm{IS}}}||P_{\beta R})$, which holds with probability at least $1 - \delta$. The solution of this inequality tells us that with probability at least $1 - \delta$, we have

$$D_{\mathrm{KL}}(P_{\beta r^{\mathrm{IS}}}||P_{\beta R}) \leq \frac{2\beta}{\epsilon_T\sqrt{2T}}\sqrt{\ln(2\sqrt{T}/\delta)} + \frac{\beta^2}{2T\epsilon_n^2}.$$

See [108] or [162] for a detailed derivation. This upper bound can be substituted into the PAC-Bayes kl bound.

Theorem 3.21 (PAC-Bayes kl Lever Bound [108], [162])**.** *For any $\beta > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$*

$$kl\big(\epsilon_T r^{\mathrm{IS}}(P_{\beta r^{\mathrm{IS}}}, D_T) \,\big\|\, \epsilon_T R(P_{\beta r^{\mathrm{IS}}})\big) \leq \frac{\ln(4\sqrt{T}/\delta)}{T} + \frac{\beta^2}{2T^2\epsilon_T^2} + \frac{\beta\sqrt{2\ln(4\sqrt{T}/\delta)}}{T\sqrt{T}\epsilon_T}.$$

Since the PAC-Bayes Pinsker bound, which was used to upper bound the right-hand side of Equation (3.36), holds when the data are drawn from a sequence of dependent behaviour policies, so does the PAC-Bayes kl Lever bound. The best value of β will be large enough for $r^{\mathrm{IS}}(P_{\beta r^{\mathrm{IS}}}, D_T)$ to be large, but not so large that the bound is dominated by the β-dependent terms. When using the distribution-dependent prior $P_{\beta R}$, it is still helpful to have an informative reference distribution P, since then β can be close to 0 and $P_{\beta r^{\mathrm{IS}}}$ will still have high estimated reward. Lever et al.'s method of upper bounding the distribution-dependent KL divergence can be applied more generally to other kinds of Gibbs distributions. See [108] or [109] for more information.

In the case where the data D_T are i.i.d., Oneto et al. [130] proved a tighter upper bound on $D_{\mathrm{KL}}(P_{\beta r^{\mathrm{IS}}} || P_{\beta R})$. Oneto et al. [130] also proved another PAC-Bayes bound for empirical Gibbs posteriors. This bound only holds when the data D_T are i.i.d., so when there is a single, fixed behaviour policy. Let $P^{\setminus t}_{\beta r^{\mathrm{IS}}}$ denote the leave-one-out Gibbs distribution, which is defined as

$$P^{\setminus t}_{\beta r^{\mathrm{IS}}}(\pi) \propto P(\pi) \exp\left(\frac{\gamma}{T} \sum_{k=1, k\neq t}^{T} \frac{\pi(a_k)}{b(a_k)} r_k\right).$$

This is the Gibbs distribution $P_{\beta r^{\mathrm{IS}}}$ with the t^{th} datum removed. Following Oneto et al., one can show that any posterior that is symmetric (meaning it does not depend on the order of the training data), and has a certain distribution stability property, satisfies the following bound.

Theorem 3.22 (Distribution stability bound [130])**.** *If the data set D_T is drawn from a single, fixed behaviour policy and if the method for selecting the posterior Q and leave-one-out posteriors $Q^{\setminus t}$ from the data set D_T satisfies the distribution stability property*

$$\max_{a', r'} \left\{ \left| \mathop{\mathbb{E}}_{\pi \sim Q} \left[\frac{\pi(a')}{b(a')} r'\right] - \mathop{\mathbb{E}}_{\pi \sim Q^{\setminus t}} \left[\frac{\pi(a')}{b(a')} r'\right] \right| \right\} \leq \gamma,$$

then for all D_T, all $t \in \{1, \ldots, T\}$, and with γ that goes to 0 as $\mathcal{O}(1/T)$, then with probability at least $1 - \delta$, we have

$$\big|R(Q) - r^{\mathrm{IS}}(Q, D_T)\big| \leq 2\gamma + \left(4T\gamma + \frac{1}{\epsilon_T}\right)\sqrt{\frac{\ln(2/\delta)}{2T}}.$$

Oneto et al. show that $P_{\beta r^{\mathrm{IS}}}$ satisfies the distribution stability property with $\gamma \leq \frac{2\beta}{T\epsilon_T}$. Therefore, using Theorem 3.22, we have that, if the data D_T are drawn from a single behaviour policy, then with probability at least $1-\delta$, we have

$$\left|R(P_{\beta r^{\mathrm{IS}}}) - r^{\mathrm{IS}}(P_{\beta r^{\mathrm{IS}}}, D_T)\right| \leq \frac{4\beta}{T\epsilon_T} + \left(\frac{8\beta}{\epsilon_T} + \frac{1}{\epsilon_T}\right)\sqrt{\frac{\ln(2/\delta)}{2T}}. \tag{3.37}$$

Once again, there is a trade-off between setting β large enough for the empirical reward to be high, but not so large that the β-dependent penalty terms become too large.

Finally, we present another technique based on algorithmic stability for deriving PAC-Bayes bounds with certain distribution-dependent priors, which is due to Rivasplata et al. [145]. Let the data set $D_T \in \mathcal{Z}^T$ consist of T i.i.d. samples. Let $D_T^{(t)}$ be the data set D_T, except with it's t^{th} element z_t replaced with z_t'. The hypothesis sensitivity coefficients are defined as:

Definition 3.23 (Hypothesis sensitivity coefficients [145])**.** Consider a learning algorithm $A : \mathcal{Z}^T \to \mathcal{H}$ that maps a data set to a hypothesis in a separable Hilbert space $\mathcal{H}$. The hypothesis sensitivity coefficients of A are defined as

$$\gamma_T = \sup_{t\in[T]} \sup_{z_t, z_t'} \left\{ \left\| A(D_T) - A(D_T^{(t)}) \right\|_{\mathcal{H}} \right\}.$$

Rivasplata et al. use the posterior Q_A and the distribution-dependent prior P_A, which are defined as

$$Q_A = \mathcal{N}(A(D_T), \sigma^2 \boldsymbol{I}), \qquad P_A = \mathcal{N}(\mathbb{E}_{D_T}[A(D_T)], \sigma^2 \boldsymbol{I}).$$

The KL divergence between Q_A and P_A is equal to $\|A(D_T) - \mathbb{E}_{D_T}[A(D_T)]\|_{\mathcal{H}}^2 / (2\sigma^2)$. Rivasplata et al. show that if the algorithm A has hypothesis sensitivity coefficients γ_T, then the output of the algorithm $A(D_T)$ satisfies a concentration inequality, which implies an upper bound on $D_{\mathrm{KL}}(Q_A||P_A)$. With probability at least $1-\delta$

$$\|A(D_T) - \mathbb{E}_{D_T}[A(D_T)]\|_{\mathcal{H}} \leq \sqrt{T}\gamma_T \left(1 + \sqrt{\frac{1}{2}\ln\left(\frac{1}{\delta}\right)}\right).$$

One can then use the union bound to combine any PAC-Bayes bound using the posterior Q_A and prior P_A with the concentration inequality satisfied by the algorithm A.

Theorem 3.24 (PAC-Bayes *kl* Hypothesis Sensitivity Bound [145])**.** *If the data set D_T is drawn from a single, fixed behaviour policy, then for any $\delta \in (0,1)$ and any algorithm A with hypothesis sensitivity coefficients γ_T, with probability at least $1-\delta$*

$$kl\big(\epsilon_T r^{\mathrm{IS}}(Q_A, D_T) \,\big\|\, \epsilon_T R(Q_A)\big) \leq \frac{\ln(4\sqrt{n}/\delta)}{n} + \frac{n\gamma_n^2\left(1 + \sqrt{\ln(2/\delta)/2}\right)^2}{2\sigma^2}.$$

Unlike the previous techniques using distribution-dependent priors, this time there is no explicit dependence on a data-independent reference distribution.

Data-Dependent Approximations of Distribution-Dependent Priors

Distribution-dependent Gibbs priors were first used by Catoni [37]. Catoni proved that the KL divergence between an arbitrary posterior Q and a distribution-dependent Gibbs prior can be upper bounded by the KL divergence between Q and an empirical (data-dependent) Gibbs prior. We are not aware of any way to apply this technique to the PAC-Bayes *kl* bound. Therefore, we apply the technique, described in Section 1.3.4. of [38], to the PAC-Bayes Bernstein bound.

We use the distribution-dependent Gibbs distribution $P_{\beta R}$ and the data-dependent Gibbs distribution $P_{\beta r^{\text{IS}}}$ (see Equation (3.35)). Catoni showed that $D_{\text{KL}}(Q||P_{\beta R})$ is related to $D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}})$.

$$D_{\text{KL}}(Q||P_{\beta R}) = D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}}) + \beta \mathop{\mathbb{E}}_{\pi\sim Q}\left[r^{\text{IS}}(\pi, D_T) - R(\pi)\right] + \ln\left(\mathop{\mathbb{E}}_{\pi\sim P}\left[e^{\beta R(\pi)}\right]\right) - \ln\left(\mathop{\mathbb{E}}_{\pi\sim P}\left[e^{\beta r^{\text{IS}}(\pi, D_T)}\right]\right).$$

If we could upper bound $\ln\left(\mathbb{E}_{\pi\sim P}\left[\exp(\beta R(\pi))\right]\right) - \ln\left(\mathbb{E}_{\pi\sim P}\left[\exp(\beta r^{\text{IS}}(\pi, D_T))\right]\right)$, then we could upper bound the difference between $D_{\text{KL}}(Q||P_{\beta R})$ and $D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}})$. We could then combine the PAC-Bayes Bernstein bound, using the prior $P_{\beta R}$, with the upper bound on $D_{\text{KL}}(Q||P_{\beta R})$, to obtain a "localised" PAC-Bayes Bernstein bound for the IS estimate. In Appendix A.1.7, we show how this can be done, and that the result is the following bound.

Theorem 3.25 (Localised PAC-Bayes Bernstein Bound for r^{IS} [38], [160])**.** *For any $\lambda \in [0, T\epsilon_T]$, any β satisfying $0 \leq \beta < \lambda$, any $\delta \in (0,1)$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1-\delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously*

$$R(Q) \geq r^{\text{IS}}(Q, D_T) - \frac{(\lambda^2+\beta^2)(e-2)}{(\lambda-\beta)T\epsilon_T} - \frac{D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}}) + 2\ln(1/\delta)}{\lambda-\beta}.$$

For more information about Catoni's localisation technique and its consequences, see [38]. Finally, we describe two more techniques for obtaining PAC-Bayes bounds with data-dependent priors that are similar to the localisation technique. The first, by London and Sandler [115], also uses a data-dependent approximation of a distribution-dependent prior. We require i.i.d. data $D_T = \{z_t\}_{t=1}^T$ and we restrict the posterior and prior to be d-dimensional Gaussian distributions. Define

$$Q_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}), \qquad P_{\widehat{\boldsymbol{\theta}}} = \mathcal{N}(\mathbb{E}_{D_T}[\widehat{\boldsymbol{\theta}}], \sigma^2 \boldsymbol{I}).$$

$\boldsymbol{\theta}$ could be the parameter vector of a parametric policy and $\widehat{\boldsymbol{\theta}}$ could be an estimate of the parameters of the behaviour policy or the optimal policy. Define $\widehat{\boldsymbol{\theta}}$ as

$$\widehat{\boldsymbol{\theta}} = \mathop{\text{argmin}}_{\boldsymbol{\theta}} \left\{ \frac{1}{T}\sum_{t=1}^{T} L(\boldsymbol{\theta}, z_t) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2 \right\}.$$

London and Sandler [115] show that if $L(\cdot, z_t)$ is convex and β-Lipschitz for any z_t, then $\widehat{\boldsymbol{\theta}}$ satisfies a concentration inequality. With probability at least $1-\delta$, we have

$$||\widehat{\boldsymbol{\theta}} - \mathbb{E}_{D_T}[\widehat{\boldsymbol{\theta}}]||_2^2 \leq \frac{\beta}{\lambda}\sqrt{\frac{2\ln(2/\delta)}{T}}.$$

This can be used to upper bound the KL divergence between $Q_{\boldsymbol{\theta}}$ and $P_{\widehat{\boldsymbol{\theta}}}$ with high probability.

Theorem 3.26 (PAC-Bayes *kl* London and Sandler Bound [115])**.** *If the data set D_T is drawn from a single, fixed behaviour policy, then for any $\delta \in (0,1)$, with probability at least $1-\delta$, for all $\boldsymbol{\theta} \in \mathbb{R}^d$ simultaneously*

$$kl\big(\epsilon_T r^{\mathrm{IS}}(Q_{\boldsymbol{\theta}}, D_T) \,\big\|\, \epsilon_T R(Q_{\boldsymbol{\theta}})\big) \leq \frac{\ln(4\sqrt{T}/\delta)}{T} + \frac{\left(||\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}||_2 + (\beta/\lambda)\sqrt{2\ln(4/\delta)/T}\right)^2}{2T\sigma^2}.$$

This is similar to the PAC-Bayes *kl* Hypothesis Sensitivity bound in Theorem 3.24. Here though, the mean of the Gaussian posterior $Q_{\boldsymbol{\theta}}$ is unrestricted and the bound contains a data-dependent penalty term $||\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}||_2$. Rivasplata et al. [144] propose another method for deriving PAC-Bayes bounds with data-dependent Gibbs priors. Using standard PAC-Bayesian proof techniques, one can show that for any (possibly data-dependent) $P \in \mathcal{P}(\Pi)$, with probability at least $1-\delta$ and all $Q \in \mathcal{P}(\Pi)$ simultaneously

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda} - \frac{\ln\left(\mathbb{E}_{D_T}\,\mathbb{E}_{\pi\sim P}\left[e^{\lambda\left(r^{\mathrm{IS}}(\pi, D_T) - R(\pi)\right)}\right]\right)}{\lambda}. \tag{3.38}$$

At this point, one would typically use Tonelli's Theorem to swap the order of the expectations w.r.t. D_T and π and then upper bound $\mathbb{E}_{D_T}\left[\exp\left(\lambda\left(r^{\mathrm{IS}}(\pi, D_T) - R(\pi)\right)\right)\right]$. However, we can only swap the order of the expectations if P does not depend on the data D_T. Instead, if we could directly upper bound $\ln\left(\mathbb{E}_{D_T}\,\mathbb{E}_{\pi\sim P}\left[\exp\left(\lambda\left(r^{\mathrm{IS}}(\pi, D_T) - R(\pi)\right)\right)\right]\right)$ for a data-dependent P, then we would obtain a PAC-Bayes bound with a data-dependent prior. Following Rivasplata et al. [144], if the data D_T are i.i.d. and the prior is $P_{\beta r^{\mathrm{IS}}}$, then

$$\ln\left(\mathop{\mathbb{E}}_{D_T}\mathop{\mathbb{E}}_{\pi\sim P_{\beta r^{\mathrm{IS}}}}\left[e^{\lambda\left(r^{\mathrm{IS}}(\pi, D_T) - R(\pi)\right)}\right]\right) \leq \frac{2}{\epsilon_T^2}\left(1 + \frac{2\lambda\beta}{T}\right) + \ln\left(1 + e^{\frac{\lambda^2}{2T\epsilon_T^2}}\right).$$

Combining this with Equation 3.38, we obtain a PAC-Bayes bound with a data-dependent prior.

Theorem 3.27 (PAC-Bayes Hoeffding-Azuma Empirical Gibbs Bound[144])**.** *If the data set D_T is drawn from a single, fixed behaviour policy, then for any $\lambda > 0$, any $0 \leq \beta \leq \lambda$, any $\delta \in (0,1)$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1-\delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously*

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - \frac{2}{\lambda\epsilon_T^2} - \frac{4\beta}{T\epsilon_T^2} - \frac{D_{\mathrm{KL}}(Q||P_{\beta r^{\mathrm{IS}}}) + \ln((1 + e^{\frac{\lambda^2}{2T\epsilon_T^2}})/\delta)}{\lambda}.$$

This is similar to the localised PAC-Bayes Bernstein bound. However, this bound only holds for i.i.d. data and has worse dependence on ϵ_T.

Priors Learned From Other Data Sets

PAC-Bayesian meta learning [132], [15], [150], [151], [114], [88], [123], [62] is another line of work in which priors are learned from data. These methods use data sets from previous learning tasks to learn a distribution over priors. Flynn et al. [62] propose PAC-Bayes bounds for meta-learning priors over the policy class for multi-armed bandit problems.

3.6.2 Optimising Bound Parameters

Many PAC-Bayes bounds contain parameters that must be set before observing the data, such as λ in the PAC-Bayes Bernstein bound in Theorem 3.6. We would like to be able to choose optimal values of these parameters. However, the optimal values are usually data-dependent. For example, the optimal λ for the PAC-Bayes Bernstein bound from Theorem 3.6, using $V^{\mathrm{IS}}(Q, D_T) \leq 1/\epsilon_T$, is

$$\lambda^* = \operatorname*{argmin}_{\lambda} \left\{ \frac{\lambda(e-2)}{T\epsilon_T} + \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda} \right\} = \sqrt{\frac{T\epsilon_T \left(D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)\right)}{e-2}} \tag{3.39}$$

Since Q is (in general) data-dependent, λ^* is as well. With this choice of λ, we would obtain the (invalid) bound

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - 2\sqrt{\frac{(e-2)\left(D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)\right)}{T\epsilon_T}} \tag{3.40}$$

In this section we present methods for approximately optimising parameters of PAC-Bayes bounds, using the PAC-Bayes Bernstein bound as an example. We compare how close each of them is to the bound in Equation (3.40).

Sample Splitting

One approach is to split the data set into subsets of equal size $D_T = D_{1:T/2} \cup D_{T/2:T}$. The first subset is used to find a good value for λ. For example, we can approximate λ^* by

$$\hat{\lambda} = \operatorname*{argmax}_{\lambda} \left\{ \max_{\rho} \left\{ r^{\mathrm{IS}}(\rho, D_{1:n/2}) - \frac{\lambda 2(e-2)}{n\epsilon_n} - \frac{D_{\mathrm{KL}}(\rho||\mu) + \ln(1/\delta)}{\lambda} \right\} \right\}.$$

Since the Bernstein bound holds only for $\lambda \in (0, T\epsilon_T]$, we should take the minimum of $\hat{\lambda}$ and $(T/2)\epsilon_T$ ($\hat{\lambda}$ is always positive). The bound is then evaluated on the second subset with $\lambda = \tilde{\lambda}$. Since $\tilde{\lambda}$ does not depend on $D_{T/2:T}$, this yields a valid bound.

Theorem 3.28 (PAC-Bayes Bernstein Bound with a Subset λ)**.** *For any $\delta \in (0,1)$, any prior $P \in \mathcal{P}(\Pi)$ and $\tilde{\lambda} = \min(\hat{\lambda}, (T/2)\epsilon_T)$, with probability at least $1-\delta$, for all $Q \in \mathcal{P}(\Pi)$ simultaneously, we have*

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_{T/2:T}) - \frac{\tilde{\lambda} 2(e-2)}{T\epsilon_T} - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\tilde{\lambda}}.$$

If $\hat{\lambda}$ is an accurate approximation of λ^*, and $\hat{\lambda} \leq (T/2)\epsilon_T$, then the PAC-Bayes Bernstein bound evaluated on $D_{T/2:T}$ and with $\lambda = \hat{\lambda}$ is approximately

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_{T/2:T}) - 2\sqrt{2}\sqrt{\frac{(e-2)\left(D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)\right)}{T\epsilon_T}}.$$

Compared to the bound in Equation (3.40), this bound has a factor of $\sqrt{2}$ in front of the penalty term because it is evaluated using half as many samples.

Union Bounds and Grids

Another approach is to define a grid of parameter values, and then use the union bound to obtain a PAC-Bayes bound that holds simultaneously for all values in the grid with high probability. Suppose we choose the following grid $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ and that $\sum_{i=1}^{m} \delta_i = \delta$. We have that for each i, with probability at least $1 - \delta_i$

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - \frac{\lambda_i(e-2)}{T\epsilon_T} - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta_i)}{\lambda_i}.$$

By a union bound argument, this bound holds for all $\lambda_i \in \Lambda$ simultaneously with probability at least $1 - \delta$. This allows us to choose the best $\lambda \in \Lambda$ after observing the data.

We may also optimise λ over a continuous interval. For example, say we want the PAC-Bayes bound to hold with high probability for all λ in the interval $[a, b]$ simultaneously, where $0 < a \leq b \leq T\epsilon_T$. We can specify a geometric grid $\Lambda = \{c^k a | k \in \mathbb{N}\} \cap [a, b]$, where $c > 1$. The number of elements in Λ is no more than $\log_c(b/a) = \ln(b/a)/\ln(c)$. Using the union bound once more, and with $\delta_i = \frac{\ln(b/a)/\ln(c)}{\delta}$, the PAC-Bayes bound holds for all $\lambda \in \Lambda$ with probability at least $1 - \delta$. For any $\lambda \in [a, b]$, there exists a $\lambda' \in \Lambda$ with $\lambda' \leq \lambda \leq c\lambda'$. We can evaluate the bound at this λ' and then upper bound the terms containing λ' with terms containing λ. We then have that with probability at least $1 - \delta$

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - \min_{\lambda \in [a,b]} \left\{ \frac{\lambda(e-2)}{T\epsilon_T} + \frac{c\left(D_{\mathrm{KL}}(Q||P) + \ln\left(\frac{\ln(b/a)/\ln(c)}{\delta}\right)\right)}{\lambda} \right\}. \tag{3.41}$$

If the value of λ that optimises the bound in Equation (3.41) is in $[a, b]$, then the bound can be rewritten as

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - (1+c)\sqrt{\frac{(e-2)\left(D_{\mathrm{KL}}(Q||P) + \ln\left(\frac{\ln(b/a)/\ln(c)}{\delta}\right)\right)}{T\epsilon_T}}.$$

This bound is the same as the bound in Equation 3.40, except that there is a factor of $1 + c$ instead of 2 in front of the KL divergence penalty and $\ln(1/\delta)$ has been replaced with $\ln(\frac{\ln(b/a)/\ln(c)}{\delta})$). For best results, we need to choose a and b such that the optimal λ is in $[a, b]$, but $\ln(b/a)$ is not too large. We should choose c such that $1 + c$ is close to 2 and $1/\ln(c)$ is small. To choose a suitable a and b, we can lower and upper bound any data-dependent terms in the equation for the optimal λ^*, such as $D_{\mathrm{KL}}(Q||P)$ in Equation 3.39. With $a = \sqrt{T\epsilon_T \ln(1/\delta)/(e-2)}$ and $b = T\epsilon_T$, and following Seldin et al. [161], one can obtain the following theorem.

Theorem 3.29 (PAC-Bayes Bernstein Bound with a Geometric λ Grid [161])**.** *For any $\delta \in (0,1)$, any prior $P \in \mathcal{P}(\Pi)$ and any $c > 1$, with probability at least $1-\delta$, simultaneously for all $Q \in \mathcal{P}(\Pi)$ that satisfy*

$$\sqrt{\frac{D_{\mathrm{KL}}(Q||P) + \ln(\nu/\delta)}{T(e-2)V^{\mathrm{IS}}(Q, D_T)}} \leq \epsilon_T,$$

we have

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - (1+c)\sqrt{\frac{(e-2)\left(D_{\mathrm{KL}}(Q||P) + \ln(\nu/\delta)\right)}{T\epsilon_T}},$$

and for all other $Q \in \mathcal{P}(\Pi)$ with the same probability, we have

$$R(Q) \geq r^{\mathrm{IS}}(Q, D_T) - 2\frac{D_{\mathrm{KL}}(Q||P) + \ln(\nu/\delta)}{T\epsilon_T},$$

where $\nu = \ln(\sqrt{T\epsilon_T(e-2)/\ln(1/\delta)})/\ln(c)$.

We believe that Langford and Caruana [100] were the first to use a geometric grid. This approach can be extended to infinite (but countable) grids, which allows us to optimise λ over an interval $[a, \infty)$. For example, see [38] or [161]. One can use the same techniques to optimise the clipping parameter τ in any of the PAC-Bayes bounds for the CIS estimate from Section 3.4.2. London and Sandler [115] proved a variant of the PAC-Bayes bound *kl* bound for the CIS estimate where τ can be optimised over the interval $(0, 1)$.

3.7 Experimental Comparison

In this section, we compare the values and properties of the presented PAC-Bayes bandit bounds. In Section 3.7.1, we describe the benchmark tasks on which we evaluate the bounds. Then, we discuss insights gained from our experiments. In Section 3.7.2, we compare the cumulative regret bounds. In Section 3.7.3, we compare the reward bounds.

3.7.1 Benchmarks

We use three benchmark tasks: one MAB problem and two CB problems.

MAB Binary Benchmark

The first benchmark is a multi-armed bandit problem with a finite set of actions $\mathcal{A} = \{1, \ldots, K\}$. The rewards are always either 0 or 1, and the reward distribution for action a_i is a Bernoulli distribution with parameter p_i. The Bernoulli parameters p_i are drawn uniformly from the interval $[0, 0.8]$ and one action always has $p_i = 0.8$. For the policy class Π, we use the set of all deterministic policies, so $\Pi = \mathcal{A}$. We report results averaged over several instances of this problem, with different randomly generated Bernoulli parameters.

CB Binary Linear Benchmark

The next benchmark is a contextual bandit problem where the optimal policy is a linear function of the state. The set of states is $\mathcal{S} = \mathbb{R}^d$ and the set of actions is $\mathcal{A} = \{1, \ldots, K\}$. The state distribution $\mathbb{P}_S$ is a standard Gaussian distribution. The rewards are either 0 or 1. When creating an instance of this problem, we sample a linear classifier

$$f(s; \theta^*) = \underset{a \in \{1,\ldots,K\}}{\operatorname{argmax}} \left\{ \langle s, \theta^* \rangle_a \right\}.$$

The weight matrix $\theta^* \in \mathbb{R}^{d \times K}$ is drawn from a standard Gaussian distribution. $\langle s, \theta^* \rangle_a$ is the ath element of $\langle s, \theta^* \rangle$. For a given state s and action a, if $a = f(s; \theta^*)$, then the reward is drawn from a Bernoulli distribution with parameter 0.8. Otherwise, the reward is drawn from a Bernoulli distribution with parameter 0.2. For the policy class Π, we use

$$\Pi = \left\{ \pi_\theta(a|s) = \frac{\exp(\langle s, \theta \rangle_a)}{\sum_{a'} \exp(\langle s, \theta \rangle_{a'})} \middle| \theta \in \mathbb{R}^{d \times K} \right\}.$$

This policy class contains all linear softmax policies.

CB Classification Benchmark

For the final benchmark task, we turned four classification data sets found on OpenML [179] and the UCI Machine Learning Repository [54] into contextual bandit problems. The states are the covariates of the classification problem, the actions are predicted class labels and the rewards are 1 if the action matches the true class label and 0 otherwise. In the resulting contextual bandit problems, $\mathcal{S} \subseteq \mathbb{R}^d$, where d is between 7 and 64, and $\mathcal{A} = \{1, \ldots, K\}$, where K is 10 or 11. See Appendix A.2.1 for more information about the data sets used. For the policy class, we use multi-layer perceptrons with two hidden layers of 200 units each. The final layer has a softmax activation function and the remaining layers have the Elu activation function [45] with $\alpha = 1$.

3.7.2 Regret Bounds

In Section 3.5, we saw several PAC-Bayes cumulative regret bounds for certain (online) multi-armed bandit algorithms. We now evaluate these bounds and algorithms as well as the PAC-Bayes cumulative regret bound that would be possible if the improved bound on the variance of the IS estimate suggested by Seldin et al. [160] was proven for EXP3.

In the MAB Binary benchmark, with $K = 10$, we compared the multi-armed bandit algorithms described in Equation 3.23 with several settings of γ_t and ϵ_t. Motivated by the PAC-Bayes Hoeffding-Azuma cumulative regret bound, we tested ϵ-greedy with $\epsilon_t = \min(t^{-1/4}K^{-1/2}, 1/K)$ and an EXP3-like algorithm with $\gamma_t = t^{1/4}K^{-1/2}\sqrt{\ln(K)}$ and $\epsilon_t = \min(t^{-1/4}K^{-1/2}, 1/K)$. We call these algorithms HA ϵ-greedy and HA EXP3 respectively. Motivated by the PAC-Bayes Bernstein cumulative regret bound, we tested ϵ-greedy with $\epsilon_t = \min(t^{-1/3}K^{-2/3}, 1/K)$ and an EXP3-like algorithm with $\gamma_t = t^{1/3}K^{-1/3}\sqrt{\ln(K)}$ and $\epsilon_t = \min(t^{-1/3}K^{-2/3}, 1/K)$. We call these algorithms Bern ϵ-greedy and Bern EXP3 respectively. Finally, we run (standard) EXP3 and the UCB1 algorithm [21] for comparison.

We evaluate each cumulative regret bound with $\delta = 0.05$. For HA ϵ-greedy and Bern ϵ-greedy, the $\ln(K)/\gamma_t$ terms in their cumulative regret bounds (Theorem 3.14 and Theorem 3.15) can be removed, so we report different bound values for the ϵ-greedy and EXP3-like variants.

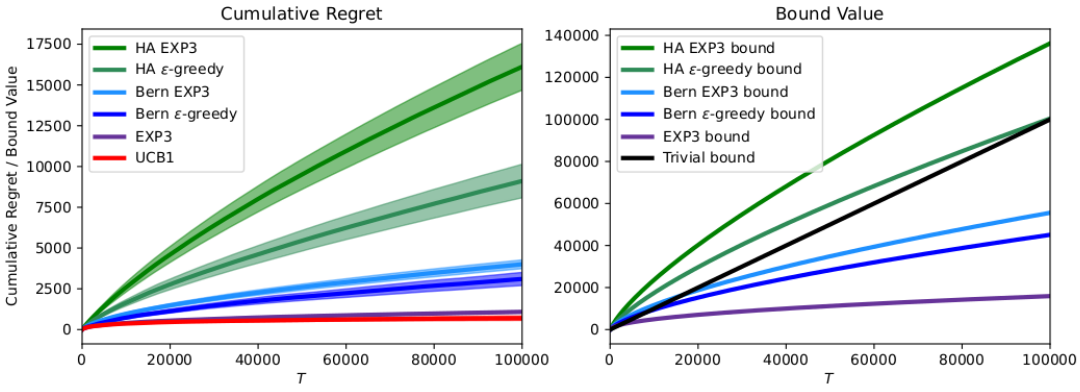


Figure 3.2: Comparison of the MAB algorithms and bounds in the MAB Binary benchmark with $K = 10$. The left plot shows the average cumulative regret plus/minus 1 standard deviation for each algorithm. The right plot shows the cumulative regret bounds, each with $\delta = 0.05$. The EXP3 bound is the cumulative regret bound that would be possible if the improved bound on the variance, suggested by Seldin et al. [160], was proven. The trivial bound assumes maximum regret (of 1) at each round.

Message 1. *The multi-armed bandit algorithms motivated by the PAC-Bayes Hoeffding-Azuma and Bernstein cumulative regret bounds all performed poorly compared to EXP3 and UCB1.*

Figure 3.2 shows both the actual cumulative regret (left) and the PAC-Bayes bounds on the cumulative regret (right) over 100000 steps. Surprisingly, EXP3 had very similar cumulative regret to UCB1. HA EXP3, HA ϵ-greedy, Bern EXP3 and Bern ϵ-greedy all had much higher cumulative regret. The PAC-Bayes cumulative regret bounds for each of these algorithms were loose, each being approximately a factor of 10 above the actual cumulative regret. On the bright side, the PAC-Bayes Bernstein regret bounds (blue) were far below the trivial regret bound at $T = 100000$. Note that the PAC-Bayes Hoeffding-Azuma regret bounds (green) would eventually drop below the trivial bound for large enough T. While the hypothetical PAC-Bayes bound for EXP3 is much lower than the other bounds, it is still quite far above the actual cumulative regret. The average cumulative regret for EXP3 at $T = 100000$ was roughly 1100, whereas the bound was roughly 15000.

Message 2. *The PAC-Bayes Bernstein cumulative regret bound was non-vacuous for $T > 20000$, but both the Hoeffding-Azuma and Bernstein cumulative regret bounds were quite loose. If the improved bound on the variance of the IS estimate suggested by Seldin et al. [160] was proven for EXP3, then a much better (though still not really tight) PAC-Bayes cumulative regret bound would be possible.*

3.7.3 Reward Bounds

In this section, we present our observations about the PAC-Bayes reward bounds for the IS and CIS estimates. Since we are not aware of a bound on the bias term in the Efron-Stein WIS bound, we only evaluate it in Appendix A.3.1, assuming the bias is 0. We compare the bounds in the (offline) MAB Binary and CB Binary Linear benchmarks. In each experiment we optimise each bound with respect to the posterior Q and then report the value of the bound and the expected reward for this Q. This allows us to compare the best possible value of each bound as well as which bound works the best as a learning objective. For details about how we optimise the various bounds with respect to Q and then evaluate them, see Appendix A.2.2.

We always use a data set of size $T = 1000$ in the MAB Binary benchmark and $T = 10000$ in the CB Binary Linear benchmark. Unless stated otherwise, we use $K = 10$ for the MAB benchmark, we use $d = 10$ and $K = 10$ for the CB benchmark, and the data set is generated using a uniform behaviour policy. At the end of Section 3.7.3, motivated by our observations, we evaluate a new offline PAC-Bayesian bandit algorithm in the CB Classification benchmark.

Insights About Different Bounds

We first investigate which of the PAC-Bayes bounds available for the IS and CIS estimates is best. We varied the number of actions K and the number of dimensions d of the state vector to investigate how each of the bounds scales with K and d. In the MAB benchmark, K varied from 2 to 50. In the CB benchmark, we ran the experiment twice. First, d was fixed at 10 and K varied from 2 to 50. Then, K was fixed at 10 and d varied from 2 to 50.

Message 3. *The PAC-Bayes kl^{-1} bound gives the greatest lower bound on the expected reward. The posterior learned by maximising the kl^{-1} bound achieves the highest expected reward.*

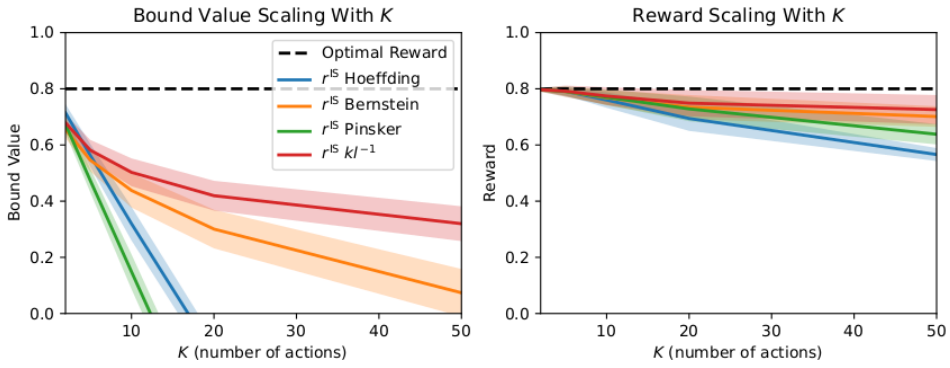


Figure 3.3: The bound value and expected reward for each bound in the MAB Binary benchmark. The number of actions K varies from 2 to 50 along the x axes.

In Figure 3.3 and Figure 3.4, we observe that increasing the number of actions causes the bound values to decay rapidly. As one would expect, due to its improved dependence on ϵ_T, the Bernstein bound decays at a much slower rate than the Hoeffding-Azuma and Pinsker bounds. The kl^{-1} bound scales up the best to large numbers of actions. As seen in Figure 3.5, increasing the number

of dimensions of the states appeared to have less effect on the bound values. The PAC-Bayes kl^{-1} bound consistently gave the greatest bound values and yielded posteriors with the greatest reward.

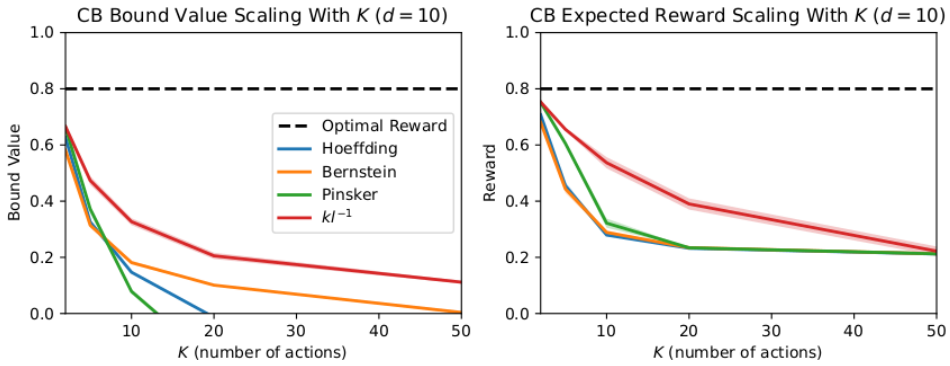


Figure 3.4: The bound value and expected reward for each bound in the CB Binary Linear benchmark. The number of dimensions of the states d is fixed at 10 and the number of actions K varies from 2 to 50 along the x axes.

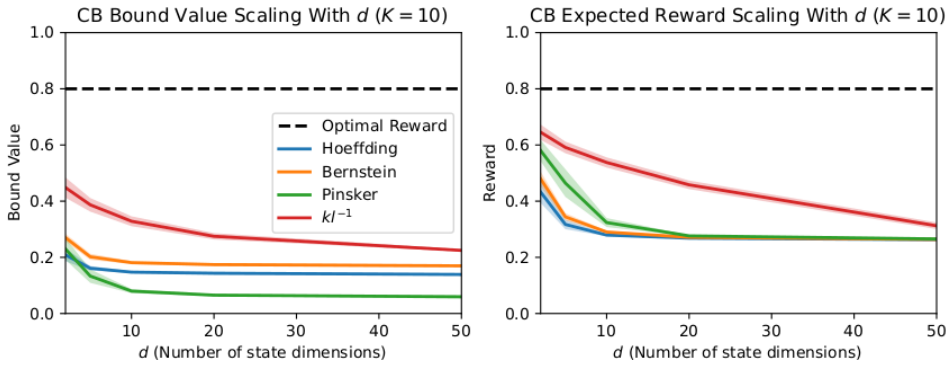


Figure 3.5: The bound value and expected reward for each bound in the CB Binary Linear benchmark. d varies from 2 to 50 along the x axes and K is fixed at 10.

Insights About Clipping

In this section, we compare the PAC-Bayes kl^{-1} bound for the IS and CIS estimates. Since clipping affects the importance weights, which are determined by the behaviour policy, we test the bounds with several behaviour policies to try and identify if and when clipping is helpful. First, we use a uniform behaviour policy. Next, we use an informative behaviour policy. In the MAB benchmark the informative policy was an ϵ-smoothed Gibbs policy.

$$b^{\text{inf}}(a) \propto e^{10R(a)}, \quad \tilde{b}^{\text{inf}}(a) = (1 - K\epsilon)b^{\text{inf}}(a) + \epsilon.$$

In the CB benchmark, the informative behaviour policy was another ϵ-smoothed policy.

$$b^{\text{inf}}(a|s) \propto e^{\langle s,\theta^* \rangle_a}, \quad \tilde{b}^{\text{inf}}(a|s) = (1 - K\epsilon)b^{\text{inf}}(a|s) + \epsilon.$$

θ^* is the weight matrix of the unknown linear classifier that generates the rewards. Finally, we use a randomly generated behaviour policy. In the MAB benchmark, the random behaviour policy was an ϵ-smoothed probability vector drawn randomly from a symmetric Dirichlet distribution with $\alpha = 1$. In the CB benchmark, the behaviour policy was an ϵ-smoothed linear softmax policy with a weight matrix θ drawn randomly from a standard Gaussian distribution. For both the informative and random behaviour policies, we used $\epsilon = 0.01$.

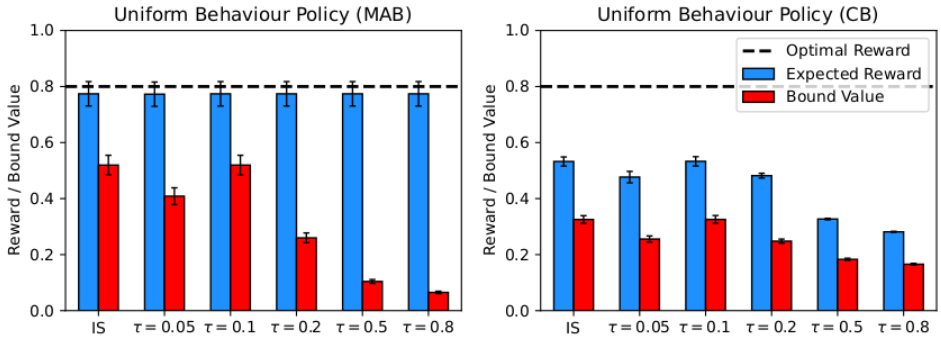


Figure 3.6: The expected reward (blue) and bound value (red) for each estimate in the MAB and CB benchmarks with a uniform behaviour policy.

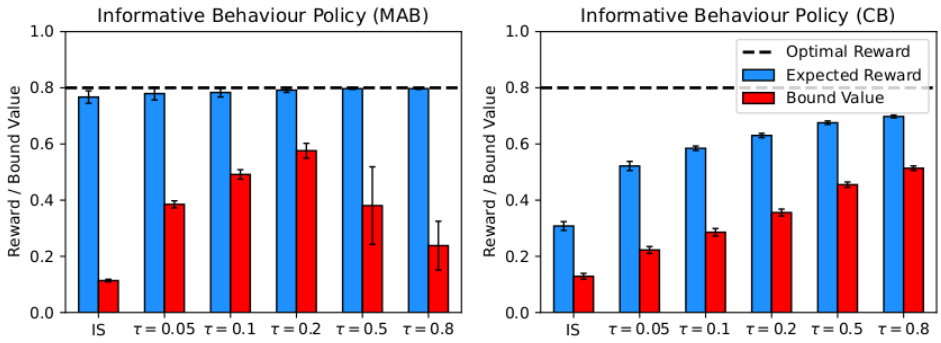


Figure 3.7: The expected reward and bound value for each estimate in the MAB and CB benchmarks with an informative behaviour policy.

Message 4. *Using the CIS estimate instead of the IS estimate can improve both the bound value and the expected reward of the learned posterior when the behaviour policy is non-uniform.*

Comparing Figure 3.6 and Figure 3.7, we see that the PAC-Bayes kl^{-1} bound for the IS estimate yields a lower bound value and lower expected reward with the informative behaviour policy than with the uniform behaviour policy. When the behaviour policy was uniform, the bound for the CIS estimate was no better than the bound for the IS estimate. However, when the behaviour policy was non-uniform, and particularly when it was informative, the kl^{-1} bound for the CIS estimate yielded greater bound values and greater expected reward.

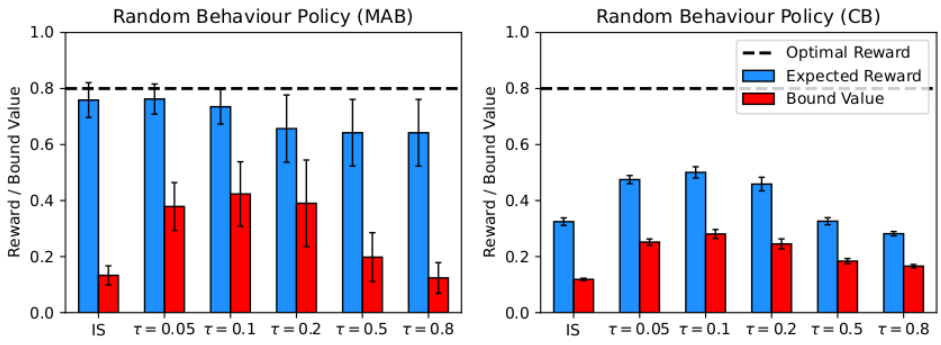


Figure 3.8: The expected reward and bound value for each estimate in the MAB and CB benchmarks with a random, non-uniform behaviour policy.

Insights About Choosing the Prior

In this section, we evaluate the presented methods for choosing the prior by using them to set the prior in a PAC-Bayes bound for the IS estimate. For the prior selection methods that work with any PAC-Bayes bound, we use them with the kl^{-1} bound, since this appeared to be the best in our earlier experiments.

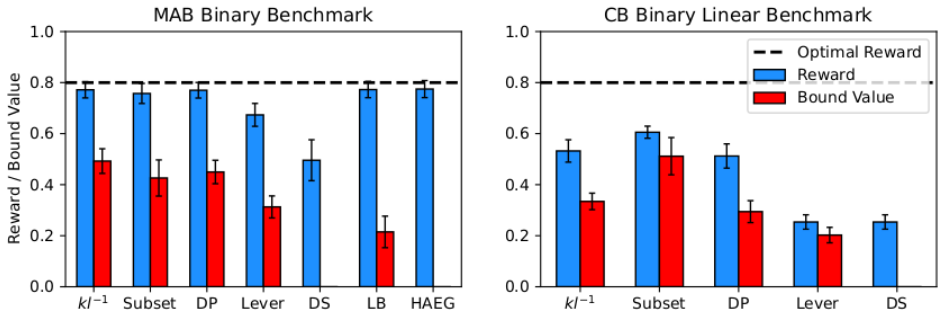


Figure 3.9: The expected reward (blue) and bound value (red) for each bound in our comparison of methods for choosing the prior. DP is the differentially private prior, DS is the distribution stability bound, LB is the localised Bernstein bound and HAEG is the Hoeffding-Azuma Empirical Gibbs bound.

In the MAB benchmark, the bounds we compared are: the kl^{-1} bound with a uniform prior (Theorem 3.5), the kl^{-1} bound with a prior learned using a subset of the data (Theorem 3.18), the kl^{-1} bound with a differentially-private prior (Theorem 3.20), the kl^{-1} Lever bound (Theorem 3.21), Oneto et al.'s distribution stability bound (Theorem 3.22), the localised PAC-Bayes Bernstein bound (Theorem 3.25) and the PAC-Bayes Hoeffding-Azuma Empirical Gibbs bound (Theorem 3.27). We do not evaluate the kl hypothesis sensitivity bound (Theorem 3.24) or the kl London and Sandler bound (Theorem 3.26) because we are not aware of a suitable learning algorithm with known hypothesis sensitivity coefficients for the first or a suitable convex and β-Lipschitz function L for the second. We compare the same bounds in the CB benchmark, but without the localised

PAC-Bayes Bernstein bound or the PAC-Bayes Hoeffding-Azuma Empirical Gibbs bound. This is because we cannot calculate $D_{\mathrm{KL}}(Q||P_{\beta r^{\mathrm{IS}}})$ for the linear softmax policy class. In Appendix A.2.3, we describe how each of these bounds was implemented.

Message 5. *A data-dependent prior learned using a subset of the data appears to be the best way to set the prior.*

Figure 3.9 shows the expected reward and bound values for the bounds we compared. In the MAB benchmark, none of the bounds with data-dependent or distribution-dependent priors achieved higher reward or higher bound values than the kl^{-1} bound with a uniform prior. In this problem, and with a uniform prior, $D_{\mathrm{KL}}(Q||P) \leq \ln(K)$. Since this is already small (relative to $\ln(1/\delta)$), it not so surprising that the more sophisticated priors did not help. The localised Bernstein bound and the Hoeffding-Azuma Empirical Gibbs bound were both greatest when $\beta = 0$. With this choice of β, the empirical Gibbs prior $P_{\beta r^{\mathrm{IS}}}$ is a uniform prior. The distribution stability bound and the Hoeffding-Azuma Empirical Gibbs bound were both vacuous, with average values of -3.613 and -5.608 respectively.

In the CB benchmark, the kl^{-1} bound with a prior learned from a subset of the data had a greater expected reward and bound value compared to the kl^{-1} bound with a standard Gaussian prior. With the η-differentially private prior, we found that as soon as η is large enough that the prior is informative, the η-dependent penalty terms become large enough to offset this benefit. The bound value was greatest when η was very close to 0, and we observe that the expected reward and bound value for the η-DP prior and the uninformative prior are almost the same. With the Lever and distribution stability bounds for the Gibbs posterior $P_{\beta r^{\mathrm{IS}}}$, we found that when β is large enough for $P_{\beta r^{\mathrm{IS}}}$ to have large empirical reward, the bounds on $D_{\mathrm{KL}}(P_{\beta r^{\mathrm{IS}}}||P_{\beta R})$ are large enough to offset this. Consequently, these two bounds were greatest when β was small, resulting in underfitting, low expected reward and low bound values. The average bound value for the distribution stability bound was -3.807. Our results suggest that using a subset of the data to learn a prior appears to be the best way to set the prior, at least for large enough policy classes.

Insights About Choosing Bound Parameters

We compare the methods presented in Section 3.6.2 for approximately optimising PAC-Bayes bounds with respect to their parameters. We use each method to set the λ parameter of the r^{IS} PAC-Bayes Bernstein bound. In both the MAB and CB benchmarks, we compare the Bernstein bound with λ learned using a subset of the data (Theorem 3.28) and the Bernstein bound with λ optimised over a geometric grid (Theorem 3.29).

We compare the grid bound with several choices of the grid parameter $c \in \{1.1, 1.2, 1.5\}$. We also compare against some baselines: the kl^{-1} bound (Theorem 3.5), the Bernstein bound with a fixed value of λ (Theorem 3.6) and the idealised Bernstein bound with the optimal choice of λ (Equation (3.40)). For the fixed value of λ, we use $\lambda = \sqrt{T\epsilon_T \ln(1/\delta)/(e-2)}$, which is equal to the optimal value when $D_{\mathrm{KL}}(Q||P) = 0$. The kl^{-1} bound represents the best parameter-free bound, the Bernstein bound with a fixed λ represents a naive choice of λ, and the idealised Bernstein bound is the best bound we could hope to achieve by optimising λ.

In Figure 3.10, we can see that in both the MAB and CB benchmarks, the sample splitting λ and the grid λ's all yield almost identical expected reward and bound values. We find that both

methods of approximately optimising the Bernstein bound w.r.t. λ give worse bound values than the fixed data-independent choice of λ. Surprisingly, the fixed λ is almost as good as the optimal λ.

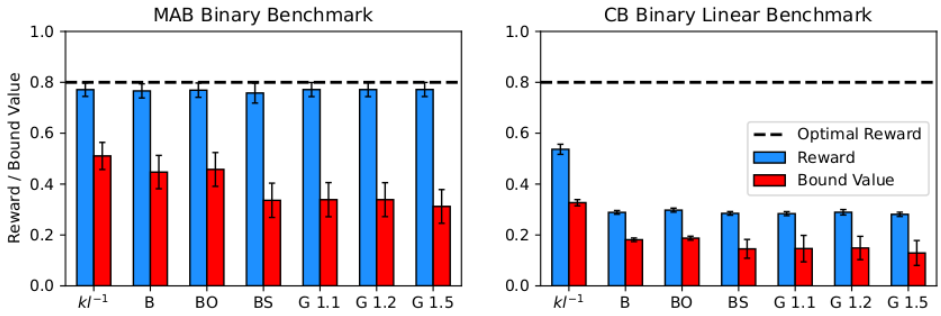


Figure 3.10: The expected reward (blue) and bound value (red) in our comparison of the methods for choosing the λ parameter of the PAC-Bayes Bernstein bound. B is the Bernstein bound with a fixed choice of λ, BO is the (invalid) Bernstein bound with the optimal λ, BS is the Bernstein bound with λ learned using a subset of the data and G 1.1, G 1.2 and G 1.5 are the Bernstein bound with λ optimised over a geometric grid with $c = 1.1$, $c = 1.2$ and $c = 1.5$.

In Appendix A.3.2, we briefly explore why the fixed value of λ was almost as good as the optimal value. It turns out that whenever T is large enough for the Bernstein bound to be non-vacuous for some value of λ, the minimum of the Bernstein bound with respect to λ is flat, which means that any reasonable data-independent guess for λ is almost as good as the optimum value.

A Method For Offline Bandits

Using the insights gained from the previous experiments, we propose a method for offline contextual bandit problems and we test it in the Contextual Bandit Classification problem where the policy class is a set of neural networks.

For the first step of our method, we use the first half of the training data $D_{1:T/2}$ to learn a diagonal Gaussian prior over the neural network weights $\boldsymbol{\theta}$ by maximising

$$\mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim P_D}\left[r^{\mathrm{IS}}(\pi_{\boldsymbol{\theta}}, D_{1:T/2})\right] - \beta D_{\mathrm{KL}}(P_D||P) \tag{3.42}$$

with respect to P_D. $\pi_{\boldsymbol{\theta}}$ is a neural network with weights $\boldsymbol{\theta}$. P is a standard Gaussian distribution. To choose β, we split $D_{1:T/2}$ into a training set D_{tr} and a validation set D_{val}. We learn diagonal Gaussian priors by maximising Equation 3.42 for $\beta \in \{10^{-k}|k \in \{1,\ldots,6\}\}$. We choose the value of β where the resulting prior P_D maximises $\mathbb{E}_{\boldsymbol{\theta}\sim P_D}[r^{\mathrm{IS}}(\pi_{\boldsymbol{\theta}}, D_{\mathrm{val}})]$. Next, we learn the clipping parameter τ. With P_D fixed, and using the first half of the training data, we optimise the following objective with respect to τ:

$$\frac{1}{\tau} kl^{-1}\left(\tau r^{\mathrm{CIS}}(P_D, D_{1:T/2}), \frac{\ln(\sqrt{2T}/\delta)}{T/2}\right).$$

This approximates the value of τ that would be optimal if we were to use the posterior $Q = P_D$. Now that we have our data-dependent prior P_D and data-dependent τ, we learn the posterior by maximising the kl^{-1} bound with respect to Q using the second half of the training data.

$$\frac{1}{\tau} kl^{-1} \left(\tau r^{\text{CIS}}(Q, D_{T/2+1:T}), \frac{D_{\text{KL}}(Q||P_D)) + \ln(\sqrt{2T}/\delta)}{T/2} \right). \tag{3.43}$$

Finally, we evaluate the bound (Equation 3.43) at the learned posterior, using the second half of the training data and the data-dependent P_D and τ.

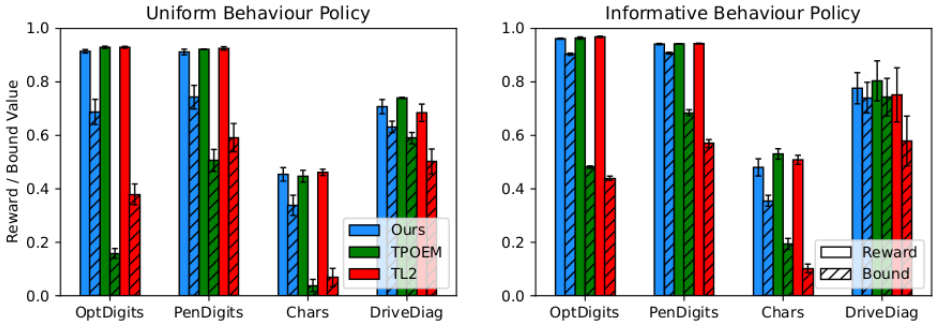


Figure 3.11: The expected reward (solid bars) and bound value (striped bars) for our proposed offline bandit algorithm (blue), the TPOEM baseline (green) and the TL2 baseline (red) in the CB classification benchmark.

We compare the expected reward and bound values of our method against two baselines. The first baseline is inspired by the POEM algorithm and PAC bound by Swaminathan and Joachims [169]. The POEM PAC bound uses covering numbers to measure the complexity of the policy class. Based on the covering number bounds for neural networks by Anthony and Bartlett [17], we expect that the original POEM PAC bound is vacuous for the CB Classification benchmark with our neural network policy class. Therefore, for a tougher comparison, we compare our proposed method to a test set bound inspired by the original POEM bound. We call this TestPOEM (TPOEM). Like the original POEM algorithm, it uses the sample variance of the CIS estimate to regularise the policy selection. We also compare against a second baseline that is similar to TPOEM, except it uses the ℓ_2 norm of the neural network weights to regularise the policy selection. For TPOEM and TL2, we use $\tau = 1/K$ since in Section 3.7.3 we saw that this was the best choice for uniform behaviour policies and a good choice for the non-uniform behaviour policies.

Message 6. *Our proposed PAC-Bayes offline contextual bandit algorithm can learn neural network policies that achieve competitive expected reward and can provide tighter reward bounds than TPOEM and TL2.*

We test our proposed method, TPOEM and TL2 in the CB Classification benchmark, first with a data set drawn using a uniform behaviour policy and then with a data set drawn using a more informative behaviour policy. For each CB Classification problem, we train a neural network classifier using 10% of the original classification data set. The ϵ-smoothed class probabilities of these

classifiers, with $\epsilon = 0.01$, are the action probabilities of the informative behaviour policies. Figure 3.11 shows the expected reward and bound values for the three methods. When the behaviour policy was uniform, our method (blue) learned policies with competitive expected reward while providing greater bound values than TPOEM (green) and TL2 (red). When the behaviour policy was informative, our method once again learned policies with competitive expected reward while providing greater bound values on all except the drive diagnosis problem, where the bound value for our method and TPOEM were comparable. The bound value for our method on the PenDigits problem was remarkably tight: the expected reward was 0.94 and the bound value was 0.91.

3.8 Conclusion

We have surveyed and empirically evaluated the available PAC-Bayes reward and regret bounds for bandit problems. In this section, we discuss our findings and highlight some open problems.

3.8.1 Findings

The results of our offline bandit experiments suggest that PAC-Bayes bounds are a useful tool for designing offline bandit algorithms with performance guarantees. In Section 3.7, we saw that the choice of bound, the choice of estimator, and the choice of the prior can each have a large impact on both the performance of the learned policy and the tightness of the performance guarantee. In Figure 3.11, we saw that a well-chosen bound, estimator and prior yields an offline bandit algorithm with competitive performance and very tight performance guarantees - even when the policy class is a set of neural networks. Similarly good performance guarantees with neural network-based policies would certainly not be possible with algorithms such as POEM [169], which measure the complexity of the policy class with covering numbers.

Our survey yields a less positive picture for existing online bandit algorithms. The cumulative regret bounds presented in Section 3.5 had sub-optimal growth rates in T and the algorithms motivated by these bounds performed poorly compared to EXP3 and UCB1. However, we believe that it would be premature to dismiss PAC-Bayes as a tool for designing online bandit algorithms with cumulative regret bounds. Rather, we believe that these less encouraging findings are indicative of PAC-Bayesian bandit algorithms being a topic that deserves further exploration. In Section 3.8.2, we describe several topics for future work that may lead to PAC-Bayesian online bandit algorithms with improved cumulative regret bounds and improved empirical performance.

3.8.2 Outlook

Tighter PAC-Bayes Bounds For "Better" Estimators

It is known that the WIS estimate often achieves lower mean squared error than the IS estimate [79]. However, the Efron-Stein PAC-Bayes reward bound for the WIS estimate was looser than some of the reward bounds that used the IS estimate (see Figure A.1). Whether improved PAC-Bayes bounds can be derived for the WIS estimate may be a key question to answer. In addition, it may be worthwhile to investigate PAC-Bayes bounds for other improved reward estimates, such as the doubly robust estimate [55].

Improved cumulative regret bounds

The PAC-Bayes Bernstein cumulative regret bound from Thm. 3.15 has a sub-optimal growth rate of $\mathcal{O}(T^{2/3}K^{1/3}\ln(K)^{1/2})$ because it uses a loose upper bound on the variance of the IS estimate. In a follow-up paper, Seldin et al. [163] used a more sophisticated bound on the variance of the IS estimate to prove a high probability regret bound of order $\mathcal{O}(\sqrt{TK})$ (ignoring log terms) for EXP3. Investigating whether this more-sophisticated variance bound is compatible with PAC-Bayes analysis is one path towards PAC-Bayesian bandit algorithms with improved cumulative regret bounds.

Beyond policy search

Following the literature on PAC-Bayesian bandits, we have focused exclusively on policy search methods, which directly learn a policy from data. However, PAC-Bayes bounds are compatible with other approaches to bandits. We briefly describe two different kinds of bandit algorithms and how PAC-Bayes bounds might be incorporated.

Broadly speaking, oracle-based bandit algorithms, such as Epoch-Greedy [102], ILOVETOCONBANDIT [6] and SquareCB [64], reduce bandit problems to supervised learning problems, such as predicting the expected reward of each action. For example, SquareCB is a meta-algorithm that turns any online regression algorithm into an online contextual bandit algorithm. In addition, if the online regression algorithm has a regret bound for online regression with an optimal growth rate, then the resulting online contextual bandit algorithm enjoys a cumulative regret bound with an optimal growth rate. This is an appealing approach for designing PAC-Bayesian bandit algorithms because it allows us to utilise PAC-Bayesian supervised learning algorithms, which are plentiful. For instance, there are PAC-Bayesian algorithms for online regression problems (e.g. [66, 74]) that are compatible with SquareCB.

Confidence bounds are a key ingredient of online bandit algorithms that follow the optimism in the face of uncertainty principle (e.g.[21], [48]) and offline bandit algorithms that follow the pessimism in the face of uncertainty principle (e.g. [142]). Upper/lower confidence bounds are estimates of the expected reward for each action that, with high probability, are guaranteed to be above/below the expected reward. In principle, PAC-Bayes bounds could be used to construct confidence bounds suitable for bandits, though we are not aware of any in the literature. We believe that investigation of PAC-Bayesian confidence bounds, as well as bandit algorithms that use them, is a fruitful direction for future work.

Chapter 4

PAC-Bayes-Style Algorithms for Linear Bandits

In this chapter, we present PAC-Bayes-style algorithms with worst-case regret guarantees for the stochastic linear bandit problem. The widely used "optimism in the face of uncertainty" principle reduces a stochastic bandit problem to the construction of a confidence sequence for the unknown reward function. The performance of the resulting bandit algorithm depends on the size of the confidence sequence, with smaller confidence sets yielding better empirical performance and stronger regret guarantees. In this chapter, we use a novel PAC-Bayes-style tail bound for adaptive martingale mixtures to construct confidence sequences which are suitable for stochastic bandits. These confidence sequences allow for efficient action selection via convex programming. We prove that a linear bandit algorithm based on our confidence sequences is guaranteed to achieve competitive worst-case regret. We also show that our confidence bounds are tighter than competitors, both empirically and theoretically. Finally, we demonstrate that our tighter confidence bounds give improved performance in several hyperparameter tuning tasks.

4.1 Introduction

The stochastic linear bandit problem is a generalisation of the classical multi-armed bandit problem [147], which was seen in the previous chapter. In each round t of a stochastic linear bandit problem, a learner chooses an action a_t and then receives a stochastic reward r_t for its choice of action. The expected value of each reward is a linear function $\phi(a_t)^\top \boldsymbol{\theta}^*$ of a known feature vector $\phi(a_t)$ associated with the corresponding action, while $\boldsymbol{\theta}^*$ is unknown. The linear bandit problem has attracted a great deal of attention because it is expressive enough to be a faithful model of many real-world decision-making problems, such as news recommendation [111] and dynamic pricing [46], yet it is simple enough to make theoretical analysis tractable.

A popular way to design algorithms for (sparse) linear bandits is to follow the principle of optimism in the face of uncertainty. This principle states that we should choose actions as if the expected reward function is as nice as plausibly possible. For linear bandits, the principle can be instantiated with a confidence sequence $\Theta_0, \Theta_1, \ldots$ for the parameter vector $\boldsymbol{\theta}^* \in \Theta$ of the expected reward function. A confidence sequence is a sequence of subsets of the full parameter space Θ, which is

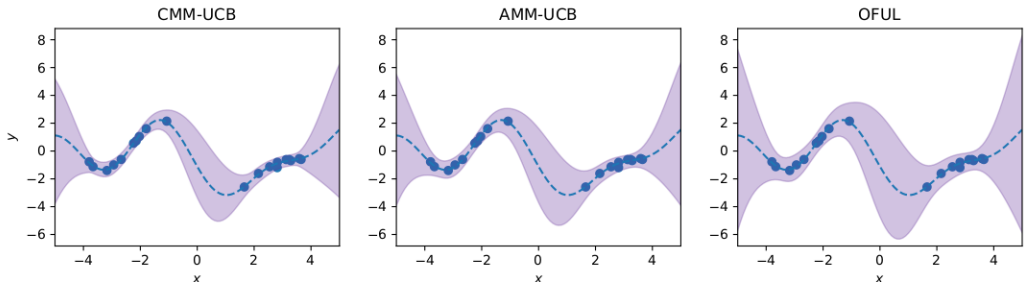


Figure 4.1: The upper and lower confidence bounds of CMM-UCB (left), AMM-UCB (middle), and OFUL [3] (right) for a test function linear in random Fourier features. The bounds from CMM-UCB and AMM-UCB are visibly closer to the true function (dashed line) than those of OFUL. The CMM-UCB confidence bounds are slightly tighter than the ones of AMM-UCB.

built iteratively as data becomes available and is constructed such that with high probability over the observed data, $\boldsymbol{\theta}^*$ is contained in each confidence set Θ_t. One can then run an upper confidence bound (UCB) algorithm, which at round t chooses an action $a_{t+1} = a$ by maximising the UCB $\max_{\boldsymbol{\theta} \in \Theta_t} \{\phi(a)^\top \boldsymbol{\theta}\}$.

The popularity of UCB algorithms stems from the fact that they come with worst-case regret guarantees and often perform well in practice. However, the performance of a UCB algorithm is intimately tied to the size of the confidence sets it uses. The smaller the subsets in the confidence sequence, the better the regret bound and, perhaps more importantly, the better the algorithm performs in practice.

Contributions. In this chapter, we develop a novel general-purpose tail bound for martingale mixtures, which can be used to construct new confidence sequences. When we specialise our general results to the linear bandit problem, the maximisation problem to compute the UCB is a convex program. We maximize the UCB over actions via gradient-based methods, and investigate two procedures for computing the UCB along with its gradient: (a) *Convex Martingale Mixture UCB (CMM-UCB):* We employ a convex solver for the UCB maximisation and calculate its gradients via differentiable convex optimisation [7]; (b) *Analytic Martingale Mixture UCB (AMM-UCB):* We exploit weak Lagrangian duality to obtain an analytic upper bound on the UCB whose gradient can be computed in closed-form or via standard automatic differentiation procedures.

Figure 4.1 highlights a key observation: both of our UCBs are tighter than those used in the state-of-the-art OFUL algorithm [3] for stochastic linear bandits. We prove this claim in Section 4.7.1 and verify it empirically in Section 4.8.1. In Section 4.8.2, we evaluate CMM-UCB, AMM-UCB, OFUL and several other linear bandit algorithms in several hyperparameter tuning problems. We find that our tighter UCBs result linear bandit algorithms with better performance.

4.2 Related Work

Algorithms with regret guarantees have been developed for several variants of the stochastic linear bandit problem. [48], [2] and [152] proposed algorithms for a linear bandit problem where the action set is a fixed, possibly infinite subset of a finite-dimensional vector space. [20] and [42]

proposed algorithms for linear bandit problems where the action set has finite cardinality, but may change over time. [3] proposed the OFUL algorithm for linear bandit problems with a changing and possibly infinite action set, which is essentially the same as the non-sparse linear bandit problem that we investigate. We consider stochastic linear bandit problems where the reward function is a composition of a possibly non-linear feature map and a linear function. This can be seen as a restricted version of the stochastic kernelised bandit problem, where the kernel feature map is finite dimensional. [168, 177, 41, 35, 155, 112] proposed algorithms with regret guarantees for various kernelised bandit problems.

In the bandit literature, confidence sets and confidence bounds constructed from online (e.g. non-i.i.d.) observation points for unknown linear functions have been proposed by [48, 152] and [3]. Online confidence sets/bounds for unknown functions in separable Hilbert spaces and reproducing kernel Hilbert spaces (RKHSs) have been proposed by [168, 1, 95, 57, 128]. [153] derived online confidence sets for unknown functions belonging to arbitrary function classes.

We use the term "mixture of martingales", or martingale mixture, to refer to a martingale of the form $\mathbb{E}_{v\sim P}[M_t(v)]$, where $(M_t(v)|t \in \mathbb{N})$ is a collection of martingales indexed by the variable $v \in \mathcal{V}$. Martingale mixtures can be traced back to [49, 148], and have been used to construct confidence sequences since at least the work of [99]. Proofs of tail bounds for martingale mixtures typically use the method of mixtures, which was first used by [149] and was later popularised by [50, 51]. Methods for martingale mixtures have seen renewed interest in the sequential testing literature [81, 91, 141]. Examples of confidence sequences for bandits that use martingale mixtures include the works of [3, 1, 95, 57, 128]. Unlike in these examples, we construct confidence sequences based on *adaptive* martingale mixtures $(\mathbb{E}_{v\sim P_t}[M_t(v)]|t \in \mathbb{N})$, where the mixture distribution P_t is refined as more data is acquired with time t.

4.3 Problem Statement and Background

We consider a problem in which a learner plays a game over a sequence of T rounds, where T may not be known in advance. In each round t, the learner observes an action set $\mathcal{A}_t$ and must choose an action $a_t \in \mathcal{A}_t$. The learner then receives a reward $r_t = \phi(a_t)^\top \boldsymbol{\theta}^* + \epsilon_t$. The feature map $\phi : \mathcal{A} \to \mathbb{R}^d$ is a known function that maps actions to d-dimensional feature vectors, where $\mathcal{A} = \bigcup_t \mathcal{A}_t$. $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is an unknown parameter with Euclidean norm bounded by some known $B_2 > 0$, i.e. $\|\boldsymbol{\theta}^*\|_2 \leq B_2$. $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$ are conditionally zero-mean σ-sub-Gaussian noise variables. These assumptions on $\boldsymbol{\theta}^*$ and $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$ are standard in the linear bandit literature, see e.g. [3].

The goal of the learner is to choose a sequence of actions that maximises the total expected reward, which is equal to $\sum_{t=1}^T \phi(a_t)^\top \boldsymbol{\theta}^*$ after T rounds. We use cumulative regret, which is the difference between the total expected reward of the learner and the optimal strategy, to evaluate the learner. For a single round, we define the regret as $\Delta(a_t) = \phi(a_t^*)^\top \boldsymbol{\theta}^* - \phi(a_t)^\top \boldsymbol{\theta}^*$, where $a_t^* = \text{argmax}_{a\in\mathcal{A}_t}\{\phi(a)^\top \boldsymbol{\theta}^*\}$. After T rounds, the cumulative regret is $\Delta_{1:T} = \sum_{t=1}^T \Delta(a_t)$.

In the special case where $\mathcal{A}_t = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d\}$ is the standard orthonormal basis of $\mathbb{R}^d$ and $\phi(a) = a$, we recover the standard multi-armed bandit problem. We focus mainly on linear bandit problems where the action sets $\mathcal{A}_t$ are continuous subsets of $\mathbb{R}^{d_\mathcal{A}}$, although our regret analysis applies to any action sets.

Confidence Sequences. For any level $\delta \in (0, 1]$, a $(1-\delta)$-confidence sequence for the parameter vector $\boldsymbol{\theta}^*$ is a sequence $\Theta_1, \Theta_2, \ldots$ of subsets of $\mathbb{R}^d$, such that each Θ_t can be calculated using the data available just after reward r_t is revealed (i.e., $a_1, r_1, \ldots, a_t, r_t$) and the sequence satisfies

$$\mathbb{P}_{\substack{a_1,a_2,\ldots,\\ r_1,r_2,\ldots,}} [\forall t \geq 1 : \boldsymbol{\theta}^* \in \Theta_t] \geq 1 - \delta.$$

A confidence sequence $\Theta_1, \Theta_2, \ldots$ is thus a sequence of data-dependent confidence sets such that with high probability over the the random actions and rewards, $\boldsymbol{\theta}^* \in \Theta_t$ holds for all $t \geq 1$ simultaneously. We remark that the confidence sets in this chapter (and the subsequent chapter) are random closed sets in the sense of Definition 1.1.1 of [126], which implies that the event $\boldsymbol{\theta} \in \Theta_t$ is measurable for any $\boldsymbol{\theta} \in \mathbb{R}^d$. In other words, given the data $a_1, r_1, \ldots, a_t, r_t$, we can always say whether any given $\boldsymbol{\theta}$ is in Θ_t.

4.4 UCB Algorithms for Linear Bandits

We describe here how to transform confidence sets for $\boldsymbol{\theta}^*$ into a UCB algorithm for the linear bandit problem. Such algorithms have appeared under various names, such as LinRel [20], LinUCB [111] and OFUL [3]. We refer to this meta algorithm as LinUCB, and give its pseudo-code in Algorithm 1. When we run LinUCB with our confidence sets, we call this algorithm CMM-UCB or AMM-UCB (see Section 4.6).

Algorithm 1: LinUCB

for $t = 0, 1, 2, \ldots$ **do**
- Construct a confidence set Θ_t from $\{(a_k, r_k)\}_{k=1}^t$
- Observe next action set $\mathcal{A}_{t+1}$
- Play next action $a_{t+1} = \operatorname{argmax}_{a \in \mathcal{A}_{t+1}} \{\mathrm{UCB}_{\Theta_t}(a)\}$
- Observe next reward r_{t+1}

end

In each round t, the first step is to construct a confidence set Θ_t from the previous observations $\{(a_k, r_k)\}_{k=1}^t$. If $\boldsymbol{\theta}^* \in \Theta_t$ with high probability, then for any action a,

$$\mathrm{UCB}_{\Theta_t}(a) := \max_{\theta \in \Theta_t} \{\phi(a)^\top \boldsymbol{\theta}\} \tag{4.1}$$

is an upper confidence bound (UCB) on $\phi(a)^\top \boldsymbol{\theta}^*$. Taking $\min_{\theta \in \Theta_t}$ in (4.1) yields the lower confidence bound $\mathrm{LCB}_{\Theta_t}(a)$. Once a confidence set Θ_t has been constructed and the next action set $\mathcal{A}_{t+1}$ has been observed, the LinUCB algorithm chooses the action

$$a_{t+1} = \operatorname*{argmax}_{a \in \mathcal{A}_{t+1}} \{\mathrm{UCB}_{\Theta_t}(a)\}, \tag{4.2}$$

which maximises the UCB. The remaining challenge lies in the construction of the confidence sets.

4.5 Confidence Sequences from Martingale Mixtures

In this Section 4.5.1, we develop a general-purpose tail bound for adaptive martingale mixtures. Then, in Section 4.5.2, we specialise our general result to the stochastic linear bandit setting, described in Section 4.3, and construct confidence sequences for the parameter $\boldsymbol{\theta}^*$.

4.5.1 General-Purpose Tail Bound for Adaptive Martingale Mixtures

We consider a general setting where we are given a filtration $(\mathcal{D}_t | t \in \mathbb{N})$, a sequence of adapted random functions $(Z_t : \mathbb{R} \to \mathbb{R} | t \in \mathbb{N})$, and a sequence of predictable random variables $(\lambda_t | t \in \mathbb{N})$. For a sequence of function values $(g_t | t \in \mathbb{N})$ (each g_t is in $\mathbb{R}$), we define the conditional cumulant generating function $\psi_t(g_t, \lambda_t)$ as

$$\psi_t(g_t, \lambda_t) := \ln\left(\mathbb{E}\left[\exp(\lambda_t Z_t(g_t)) | \mathcal{D}_{t-1}\right]\right),$$

where the expectation $\mathbb{E}$ is over $Z_t(g_t)$ and λ_t (although λ_t is non-random when conditioned on $\mathcal{D}_{t-1}$). We use the shorthand $\boldsymbol{g}_t := (g_1, g_2, \ldots, g_t)$ and $\boldsymbol{\lambda}_t := (\lambda_1, \lambda_2, \ldots, \lambda_t)$. Let

$$M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t) = \exp\left(\sum_{k=1}^{t} \lambda_k Z_k(g_k) - \psi_k(g_k, \lambda_k)\right). \tag{4.3}$$

$M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)$ is a slight generalisation of the martingale used in Appendix B.1 of [153]. One can show that for any sequence of function values $(g_t | t \in \mathbb{N})$, $(M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t) | t \in \mathbb{N})$ is a martingale and $\mathbb{E}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)] = 1$ (see Appendix B.1.1). We will now construct an *adaptive* martingale mixture.

We call a data-dependent sequence of probability distributions $(P_t | t \in \mathbb{N})$ an *adaptive sequence of mixture distributions* if: (a) P_t is a distribution over $\boldsymbol{g}_t \in \mathbb{R}^t$; (b) P_t is $\mathcal{D}_{t-1}$-measurable; (c) the distributions are consistent in the sense that their marginals coincide, i.e. $\int P_t(\boldsymbol{g}_t) dg_t = P_{t-1}(\boldsymbol{g}_{t-1})$ for all t. These conditions on the sequence of distributions ensure that the martingale mixture $(\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)] | t \in \mathbb{N})$ is in fact a martingale. In Appendix B.1.1, we verify this and show that $\mathbb{E}[\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]] = 1$. From here, we can use Ville's inequality for non-negative supermartingales [181] to obtain our general purpose tail bound.

Theorem 4.1 (Tail Bound for Adaptive Martingale Mixtures)**.** *For any $\delta \in (0, 1)$, any sequence of predictable random variables $(\lambda_t | t \in \mathbb{N})$, and any adaptive sequence of mixture distributions $(P_t | t \in \mathbb{N})$, with probability at least $1 - \delta$,*

$$\ln\left(\mathop{\mathbb{E}}_{\boldsymbol{g}_t \sim P_t}\left[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)\right]\right) \leq \ln(1/\delta) \quad \textit{for all } t \geq 1. \tag{4.4}$$

We provide a proof of this result in App. B.1.2. Note that if $\psi_k(g_k, \lambda_k)$ in (4.3) is replaced by an upper bound on $\psi_k(g_k, \lambda_k)$, the statement of the theorem still holds. We could use the Donsker-Varadhan change of measure inequality to re-write (4.4) as

$$\sup_{Q \in \mathcal{P}(\mathbb{R}^t)} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{g}_t \sim Q}\left[\sum_{k=1}^{t} \lambda_k Z_k(g_k) - \psi_k(g_k, \lambda_k)\right] - D_{\mathrm{KL}}(Q || P_t) \right\} \leq \ln(1/\delta). \tag{4.5}$$

This is a PAC-Bayes-style inequality which is time-uniform and uses a data-dependent mixture distribution (or prior) P_t. Since the supremum in (4.5) is reached when $Q(\boldsymbol{g}_t) \propto P_t(\boldsymbol{g}_t)M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)$, we can view (4.4) as a PAC-Bayes-style inequality evaluated at this posterior Q. When we construct our confidence sequence for linear bandits (see Section 4.5.2), we don't need to introduce a posterior Q via the Donsker-Varadhan change of measure inequality. However, we do use (4.5) in the proofs of some of our regret bounds (see Lemma B.16 in Appendix B.4.2).

Theorem 4.1 is closely related to the general-purpose anytime-valid PAC-Bayes bound in Theorem 3.1 of [43]. Like the general-purpose bound in [43], (4.4) would also hold for any collection of non-negative (super)martingales $(M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)|t \in \mathbb{N})$, indexed by $\boldsymbol{g}_t$ and $\boldsymbol{\lambda}_t$, which satisfies $M_0 \leq 1$. The main difference is that we consider mixture distributions/priors over function values, which allows us to derive PAC-Bayes-style inequalities with *adaptive* sequences of mixture distributions/priors. PAC-Bayes bounds with somewhat similar adaptive sequences of priors have recently been proposed by [74, 75].

4.5.2 Confidence sequences for stochastic linear bandits

We now specialise Theorem 4.1 to the stochastic linear bandit setting. For the filtration $(\mathcal{D}_t|t \in \mathbb{N})$, we set $\mathcal{D}_t$ to be the σ-algebra generated by $(a_1, r_1, \ldots, a_t, r_t, a_{t+1})$. For reasons that will become clear, we choose $Z_t(g_t) = (g_t - \phi(a_t)^\top \boldsymbol{\theta}^*)\epsilon_t$. Since $Z(g_t)$ is linear in the noise variable ϵ_t, $\psi_t(g_t, \lambda_t)$ can be upper bounded using the sub-Gaussian property of ϵ_t. We have

$$\psi_t(g_t, \lambda_t) = \ln\left(\mathbb{E}\left[\exp\left(\lambda_t(g_t - \phi(a_t)^\top \boldsymbol{\theta}^*)\epsilon_t\right)|\mathcal{D}_{t-1}\right]\right) \leq \lambda_t^2\sigma^2(g_t - \phi(a_t)^\top \boldsymbol{\theta}^*)^2/2. \tag{4.6}$$

With this upper bound on $\psi_t(g_t, \lambda_t)$, Theorem 4.1 implies that, with probability at least $1 - \delta$

$$\mathbb{E}_{\boldsymbol{g}_t \sim P_t}\left[\exp\left\{\sum_{k=1}^{t} \lambda_k(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)(r_k - \phi(a_k)^\top \boldsymbol{\theta}^*) - \frac{\sigma^2}{2}\lambda_k^2(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)^2\right\}\right] \leq \frac{1}{\delta}. \tag{4.7}$$

Since $Z_f(g_t)$ is linear in g_t, this integral has a closed-form solution whenever the mixture distribution is a Gaussian $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$. Although there is a closed-form solution for any predictable sequence $(\lambda_t|t \in \mathbb{N})$ (see Appendix B.2.1), we choose $\lambda_t \equiv 1/\sigma^2$, which yields a relatively simple, convex quadratic constraint for $\boldsymbol{\theta}^*$. Collecting the feature vectors in $\Phi_t := \big(\phi(a_1), \ldots, \phi(a_t)\big)^\top \in \mathbb{R}^{t \times d}$ and the rewards in the vector $\boldsymbol{r}_t := (r_1, \ldots, r_t)^\top$, we arrive at (see Appendix B.2.2)

$$\|\Phi_t\boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 \leq (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2}\right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{r}_t) + \sigma^2 \ln\left(\det\left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2}\right)\right) + 2\sigma^2 \ln\frac{1}{\delta} =: R_{\mathrm{MM},t}^2. \tag{4.8}$$

This inequality has an attractive interpretation. At each step t of the bandit process, the (unknown) ground-truth reward vector $\Phi_t^\top \boldsymbol{\theta}^*$ lies in a sphere around the observed reward vector $\boldsymbol{r}_t$, with radius $R_{\mathrm{MM},t}$. We can think of the mean vector $\boldsymbol{\mu}_t$ as a prediction of the reward vector $\boldsymbol{r}_t$, given the previous data $a_1, r_1, \ldots, a_{t-1}, r_{t-1}, a_t$. The covariance matrix $\boldsymbol{T}_t$ can be thought of as the uncertainty associated with the prediction $\boldsymbol{\mu}_t$. If the distance between $\boldsymbol{\mu}_t$ and $\boldsymbol{r}_t$ is close to 0 (i.e., $\boldsymbol{\mu}_t$ is a good predictor of $\boldsymbol{r}_t$), then the quadratic "prediction error" term in (4.8) will be close to 0,

and we can afford to choose $\boldsymbol{T}_t$ to be close to zero to minimise the log determinant penalty. In this situation, (4.8) can give a much tighter constraint than the naive bound $\sim t\sigma^2$, especially when σ is a pessimistic upper bound on the true sub-Gaussian parameter. The naive bound follows from the fact that $\|\Phi_t\boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 = \|\boldsymbol{\epsilon}_t\|_2^2$, where $\boldsymbol{\epsilon}_t = (\epsilon_1, \ldots, \epsilon_t)$.

Combining the constraint in (4.8) with our assumption $\|\boldsymbol{\theta}^*\|_2 \leq B_2$ yields our confidence sequence.

Corollary 4.2 (Martingale Mixture Confidence Sequence)**.** *For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, it holds with probability at least $1-\delta$ that for all $t \geq 1$ simultaneously $\boldsymbol{\theta}^*$ lies in the set*

$$\Theta_t^{\ell_2} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \|\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t} \quad \text{and} \quad \|\boldsymbol{\theta}\|_2 \leq B_2 \right\}. \tag{4.9}$$

The boundaries of the constraints in (4.9) are d-dimensional ellipses, which means the set $\Theta_t^{\ell_2}$ is the intersection of two d-dimensional ellipses. The leftmost plot on the title page depicts $\Theta_t^{\ell_2}$ for the case when $d = 2$.

4.6 Martingale Mixture UCB Algorithms

In this section, we describe our CMM-UCB and AMM-UCB algorithms, which are two different implementations of LinUCB (Algorithm 1) with our confidence sequence $\Theta_t^{\ell_2}$ from Corollary 4.2. Both of our algorithms require us to specify the mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$ and compute the radius $R_{\mathrm{MM},t}$, which appears in (4.9). We highlight some sensible choices for the mixture distributions in Section 4.6.4 and describe how $R_{\mathrm{MM},t}$ can be computed efficiently in Section 4.6.5.

4.6.1 UCB Computation and Optimisation

To run the LinUCB action selection rule with our confidence sequence, we need to be able to maximise $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ with respect to a. When we use our confidence sequence from Corollary 4.2, the value of the UCB at the action a is the solution of the following convex optimisation problem.

$$\mathrm{UCB}_{\Theta_t^{\ell_2}}(a) \;=\; \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \;\; \phi(a)^\top \boldsymbol{\theta} \quad \text{s.t.} \;\; \|\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t} \quad \text{and} \quad \|\boldsymbol{\theta}\|_2 \leq B_2. \tag{4.10}$$

One can also obtain lower confidence bounds (LCBs) by replacing $\max_{\boldsymbol{\theta}\in\mathbb{R}^d}$ with $\min_{\boldsymbol{\theta}\in\mathbb{R}^d}$. If the action sets have finite cardinality, $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ can be maximised by solving (4.10) for each $a \in \mathcal{A}_t$ and then comparing the solutions.

If the action sets are continuous subsets of $\mathbb{R}^{d_{\mathcal{A}}}$, then exact maximisation of $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ is (in general) infeasible. For example, consider the case when the feature map ϕ is linear in a. If this is the case, $\mathrm{UCB}_{\Theta_t^{\ell_2}}(\cdot)$ is the maximum over a set of linear functions, which is a convex function of a (see Equation (3.7) in Section 3.2.3 of [32]). Since maximisation of a convex function is in general NP-hard, exact maximisation of $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ is NP-hard when ϕ is linear. If ϕ is non-linear, maximising $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ is still difficult in general.

For this reason, when the action sets are continuous subsets of $\mathbb{R}^{d_{\mathcal{A}}}$, we approximately maximise $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ via gradient-based local search (possibly augmented with several restarts). Note that this requires that the feature map is differentiable. In summary, to (approximately) run the LinUCB algorithm with discrete or continuous action sets, we must be able to: (a) compute $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$; (b) compute the gradient of $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ with respect to a.

4.6.2 Convex Martingale Mixture UCB Algorithm

Our Convex Martingale Mixture UCB (CMM-UCB) algorithm is based on computing (4.10) using numerical convex (conic) solvers from the CVXPY library [52, 8]. Note that (4.10) is already stated in a conic form, which is favourable for conic solvers [32]. Solving (4.10) numerically gives the tightest UCBs (and LCBs) that can be achieved using our confidence sequence.

To compute the gradient of $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ with respect to the action a, we use recently developed methods for differentiating conic programs at their optimum [7], which are implemented in the cvxpylayers library.

When strong duality holds for (4.10), one could consider using numerical convex solvers to optimise the Lagrangian dual function with respect to the Lagrange multipliers, which would also yield the solution of (4.10). An advantage of this approach, which one could call Dual CMM-UCB, it that the dual problem is a convex optimisation problem with only 2 variables, whereas the primal in (4.10) has d variables. Unfortunately, the expression for the dual function (see Appendix B.3.1) contains the inverse of a $d \times d$ matrix, so it may in fact be more efficient to solve the primal problem.

4.6.3 Analytic Martingale Mixture UCB Algorithm

Our Analytic Martingale Mixture UCB (AMM-UCB) algorithm uses an analytic upper bound on the solution of (4.10). The resulting analytic confidence bounds are looser than the numerical confidence bounds used by CMM-UCB, but are cheaper to evaluate and maximise. Theorem 4.3 states our upper bound on the solution of (4.10).

Theorem 4.3 (Analytic UCB)**.** *For all $\alpha > 0$, we have*

$$\mathrm{UCB}_{\Theta_t^{\ell_2}}(a) = \max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\} \;\leq\; \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{AMM},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)}, \tag{4.11}$$

$$\textit{where} \qquad \widehat{\boldsymbol{\theta}}_{\alpha,t} = \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \Phi_t^\top \boldsymbol{r}_t,$$

$$R_{\mathrm{AMM},t}^2 = R_{\mathrm{MM},t}^2 + \alpha B_2^2 - \boldsymbol{r}_t^\top \boldsymbol{r}_t + \boldsymbol{r}_t^\top \Phi_t \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \Phi_t^\top \boldsymbol{r}_t.$$

In addition, if $\Theta_t^{\ell_2}$ has an interior point, then for all $\alpha > 0$, we have

$$\mathrm{UCB}_{\Theta_t^{\ell_2}}(a) = \max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\} = \min_{\alpha > 0} \left\{ \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{AMM},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)} \right\}. \tag{4.12}$$

In Appendix B.3.1, we derive this analytic UCB by partial optimisation of the Lagrangian dual function. Using strong duality, one can show that the analytic UCB minimised with respect to α is

equal to $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$. Due to the closed-form expression of the analytic UCB in (4.11), its gradient with respect to a can be computed with standard automatic differentiation packages. Computing the gradient (with respect to a) of the optimised analytic UCB in (4.12) is less straightforward, because the optimal α will in general depend on a. Note that optimising α as in (4.12) is equivalent to the Dual-CMM approach described in the previous section (although in (4.12) we have already set one of the optimisation variables/Lagrange multipliers to its optimum value).

4.6.4 Choosing the Mixture Distributions

The mixture distributions in our confidence sequences play a role similar to the priors used in the PAC-Bayes or luckiness [68, 69] frameworks. Our confidence sequences and regret bounds are valid for any (Gaussian) choice of the mixture distributions, but if better mixture distributions are used, then our confidence sequences will get smaller, the performance of our CMM-UCB and AMM-UCB algorithms with get better and their regret bounds (see Section 4.7) will get tighter.

Choosing the mixture distributions used in CMM-UCB and AMM-UCB is therefore an important design decision. In the remainder of this section, we first describe some standard choices for the mixture distributions. The standard mixture distributions allow for more efficient computation of the radius $R^2_{\mathrm{MM},t}$ (defined in (4.8). Then, we describe a general method for updating the mixture distributions in a more data-dependent fashion, where the mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{T}_t$ are refined using previously observed rewards.

Standard Mixture Distributions. In order for a sequence of Gaussian mixture distributions $(\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)|t \in \mathbb{N})$ to be a *sequence of adaptive mixture distributions* (as defined in Section 4.5.1), we require: (a) $\boldsymbol{\mu}_t$ and $\boldsymbol{T}_t$ can only depend on $a_1, \ldots, a_t$ and $r_1, \ldots, r_{t-1}$; (b) the first $t-1$ elements of $\boldsymbol{\mu}_t$ must be equal to $\boldsymbol{\mu}_{t-1}$; (c) the upper left $t-1 \times t-1$ block of $\boldsymbol{T}_t$ must be $\boldsymbol{T}_{t-1}$; (d) $\boldsymbol{T}_t$ must be positive (semi-)definite. These conditions are all satisfied if we use a mean vector $\boldsymbol{\mu}_t$ and covariance matrix $\boldsymbol{T}_t$ of the form

$$\boldsymbol{\mu}_t = [m(a_1), m(a_2), \ldots, m(a_t)]^\top, \qquad \boldsymbol{T}_t = \begin{bmatrix} k(a_1,a_1) & k(a_1,a_2) & \cdots & k(a_1,a_t) \\ k(a_2,a_1) & k(a_2,a_2) & \cdots & k(a_2,a_t) \\ \vdots & \vdots & \ddots & \vdots \\ k(a_t,a_1) & k(a_t,a_2) & \cdots & k(a_t,a_t) \end{bmatrix}, \tag{4.13}$$

where $m : \mathcal{A} \to \mathbb{R}$ is a mean function and $k : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is a positive definite kernel function. Note that the distribution $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$ is the marginal distribution of a Gaussian process (GP) [186] with mean function m and kernel function k, at the actions $a_1, \ldots, a_t$.

For the linear bandit problem, in which we know that the ground-truth reward function values $\phi(a_1)^\top \boldsymbol{\theta}^*, \phi(a_2)^\top \boldsymbol{\theta}^*, \ldots$ come from a linear reward function $\phi(\cdot)^\top \boldsymbol{\theta}^*$, it is natural to use a linear mean function $m(a) = \phi(a)^\top \boldsymbol{\theta}_0$ and a linear kernel function $k(a, a') = \phi(a)^\top \boldsymbol{\Sigma}_0 \phi(a')$ (where $\boldsymbol{\Sigma}_0$ is symmetric and positive-definite), since this corresponds to a Gaussian process on linear functions. By direct computation, the Gaussian mixture distribution with this m and k, and with $\boldsymbol{\mu}_t$ and $\boldsymbol{T}_t$ as in (4.13), is $P_t = \mathcal{N}(\Phi_t \boldsymbol{\theta}_0, \Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top)$. When $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$, we recover what we call the *standard mixture distributions* $P_t = \mathcal{N}(\mathbf{0}, \Phi_t \Phi_t^\top)$.

The mixture distribution $P_t = \mathcal{N}(\Phi_t \boldsymbol{\theta}_0, \Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top)$ can also be motivated using Bayesian principles. Suppose that we believe, in a Bayesian sense, that $\boldsymbol{\theta}^* \sim \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$. This distribution over $\boldsymbol{\theta}^*$ induces a prior distribution over the vector of function values $\Phi_t \boldsymbol{\theta}^*$, which is $\mathcal{N}(\Phi_t \boldsymbol{\theta}_0, \Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top)$. This viewpoint merits our decision to call $P_t = \mathcal{N}(\mathbf{0}, \Phi_t \Phi_t^\top)$ the *standard* mixture distributions, since this corresponds to a standard Gaussian prior over $\boldsymbol{\theta}^*$.

Adaptive Mixture Distributions. The requirement that the sequence $(\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t) | t \in \mathbb{N})$ is an adaptive sequence of mixture distributions allows for more general choices of $\boldsymbol{\mu}_t$ and $\boldsymbol{T}_t$ than those seen above. Here, we investigate a method for refining $\boldsymbol{\mu}_t$ and $\boldsymbol{T}_t$ based on the previously observed actions *and* rewards. As before, m and k are any fixed mean and (positive definite) kernel function. Each new row and column of $\boldsymbol{T}_t$ is set using an adaptive kernel function k_{t-1}. For $\beta > 0$, define

$$k_t(a, a') := k(a, a') - \boldsymbol{k}_t(a)^\top \left(\boldsymbol{K}_t + \beta \boldsymbol{I}\right)^{-1} \boldsymbol{k}_t(a'),$$

where $\boldsymbol{k}_t(a) = [k(a, a_1), \ldots, k(a, a_t)]^\top$ and $\boldsymbol{K}_t$ is the kernel matrix whose i, j-th element is $k(a_i, a_j)$. The covariance matrix $\boldsymbol{T}_t$ of the mixture distribution at time t becomes

$$\boldsymbol{T}_t = \begin{bmatrix} k_0(a_1, a_1) & k_1(a_1, a_2) & \cdots & k_{t-1}(a_1, a_t) \\ k_1(a_2, a_1) & k_1(a_2, a_2) & \cdots & k_{t-1}(a_2, a_t) \\ \vdots & \vdots & \ddots & \vdots \\ k_{t-1}(a_t, a_1) & k_{t-1}(a_t, a_2) & \cdots & k_{t-1}(a_t, a_t) \end{bmatrix}. \tag{4.14}$$

The t^{th} column and t^{th} row of this matrix depend only on only $a_1, r_1, \ldots, a_t$. Our motivation for this choice of the kernel function is: (a) generalising the usual Bayesian GP posterior covariance, one can show that if the kernel function k is positive definite, then $\boldsymbol{T}_t$ is positive semi-definite; (b) k_t is the Bayesian GP posterior covariance function (with a Gaussian likelihood with variance β). Each new element of $\boldsymbol{\mu}_t$ is set by evaluating an adaptive mean function m_{t-1} at the latest action a_t. Define

$$m_t(a) := m(a) - \boldsymbol{k}_t(a)^\top \left(\boldsymbol{K}_t + \beta \boldsymbol{I}\right)^{-1} \left(\boldsymbol{m}_t - \boldsymbol{r}_t\right).$$

The mean $\boldsymbol{\mu}_t$ of the mixture distribution at time t becomes

$$\boldsymbol{\mu}_t = [m_0(a_1), m_1(a_2), \ldots, m_{t-1}(a_t)]^\top. \tag{4.15}$$

The t^{th} element, $m_{t-1}(a_t)$, depends on only $a_1, r_1, \ldots, a_t$, so this is a valid choice for $\boldsymbol{\mu}_t$. Note that m_t is the Bayesian GP posterior mean function (again with a Gaussian likelihood with variance β).

4.6.5 Efficient Radius Computation

If we compute the squared radius $R^2_{\text{MM},t}$ using the expression in (4.8), then we have to compute the inverse and determinant of the $t \times t$ matrix $\boldsymbol{I} + \boldsymbol{T}_t / \sigma^2$. We will now show that for any mixture distribution of the form $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top)$, where $\boldsymbol{\Sigma}_0$ is symmetric and positive-definite, we can re-write the expression for $R^2_{\text{MM},t}$ such that we instead need to compute the inverse and determinant of a $d \times d$ matrix.

When $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top)$, the squared radius $R^2_{\mathrm{MM},t}$ is equal to

$$R^2_{\mathrm{MM},t} = (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \left(\boldsymbol{I} + \frac{\Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top}{\sigma^2} \right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{r}_t) + \sigma^2 \ln \left(\det \left(\boldsymbol{I} + \frac{\Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top}{\sigma^2} \right) \right) + 2\sigma^2 \ln(1/\delta).$$

By using the Weinstein–Aronszajn identity, and then doing some algebra, we have

$$\det \left(\boldsymbol{I} + \frac{\Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top}{\sigma^2} \right) = \det \left(\boldsymbol{I} + \frac{\boldsymbol{\Sigma}_0^{1/2} \Phi_t^\top \Phi_t \boldsymbol{\Sigma}_0^{1/2}}{\sigma^2} \right) = \det(\boldsymbol{\Sigma}_0/\sigma^2) \det \left(\Phi_t^\top \Phi_t + \sigma^2 \boldsymbol{\Sigma}_0^{-1} \right).$$

Using Lemma B.6 with $\gamma = \sigma^2$, $\boldsymbol{v} = \boldsymbol{\mu}_t - \boldsymbol{r}_t$ and $\boldsymbol{M} = \Phi_t \boldsymbol{\Sigma}_0^{1/2}$, we have

$$\begin{aligned}
&(\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \left(\boldsymbol{I} + \frac{1}{\sigma^2} \Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top \right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{r}_t) = (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top (\boldsymbol{\mu}_t - \boldsymbol{r}_t) \\
&- (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \Phi_t \boldsymbol{\Sigma}_0^{1/2} \left(\boldsymbol{\Sigma}_0^{1/2} \Phi_t^\top \Phi_t \boldsymbol{\Sigma}_0^{1/2} + \sigma^2 \boldsymbol{I} \right)^{-1} \boldsymbol{\Sigma}_0^{1/2} \Phi_t^\top (\boldsymbol{\mu}_t - \boldsymbol{r}_t) \\
&= (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top (\boldsymbol{\mu}_t - \boldsymbol{r}_t) - (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \Phi_t \boldsymbol{\Sigma}_0^{1/2} \left(\boldsymbol{\Sigma}_0^{1/2} \left(\Phi_t^\top \Phi_t + \sigma^2 \boldsymbol{\Sigma}_0^{-1} \right) \boldsymbol{\Sigma}_0^{1/2} \right)^{-1} \boldsymbol{\Sigma}_0^{1/2} \Phi_t^\top (\boldsymbol{\mu}_t - \boldsymbol{r}_t) \\
&= (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top (\boldsymbol{\mu}_t - \boldsymbol{r}_t) - (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \Phi_t \left(\Phi_t^\top \Phi_t + \sigma^2 \boldsymbol{\Sigma}_0^{-1} \right)^{-1} \Phi_t^\top (\boldsymbol{\mu}_t - \boldsymbol{r}_t).
\end{aligned}$$

The resulting expression for $R^2_{\mathrm{MM},t}$ is rather cumbersome, but the upshot is that we now (only) need to compute the inverse and determinant of the $d \times d$ matrix $\Phi_t^\top \Phi_t + \sigma^2 \boldsymbol{\Sigma}_0^{-1}$. Since, we compute $R^2_{\mathrm{MM},t}$ at each round t, we can update the inverse of $\Phi_t^\top \Phi_t + \sigma^2 \boldsymbol{\Sigma}_0^{-1}$ incrementally using the Sherman-Morrison formula [166]. The determinant of $\Phi_t^\top \Phi_t + \sigma^2 \boldsymbol{\Sigma}_0^{-1}$ can be updated incrementally using the relation

$$\det \left(\Phi_t^\top \Phi_t + \sigma^2 \boldsymbol{\Sigma}_0^{-1} \right) = \det(\Phi_{t-1}^\top \Phi_{t-1} + \sigma^2 \boldsymbol{\Sigma}_0^{-1})(1 + \phi(a_t)^\top (\Phi_{t-1}^\top \Phi_{t-1} + \sigma^2 \boldsymbol{\Sigma}_0^{-1})^{-1} \phi(a_t)),$$

which can be found in Equation (6) in Lemma 11 of [3].

4.7 Theoretical Analysis

In this section, we analyse the tightness of our CMM-UCB and AMM-UCB confidence bounds relative to the OFUL confidence bounds [3]. We also establish cumulative regret bounds for our CMM-UCB and AMM-UCB algorithms. First, we state a data-dependent regret bound which illustrates how the radius of the analytic UCB from Sec. 4.6.3 influences the regret of both algorithms. Then, we prove a data-independent regret bound which illustrates the worst-case growth rate of the cumulative regret, with explicit dependence on the feature vector dimension d and the number of rounds T. We begin by stating the assumptions (which are standard) under which our analysis holds.

Assumption 4.4 (Sub-Gaussian noise)**.** Let $\mathcal{D}_k$ be the σ-algebra generated by $(a_1, r_1, \ldots, a_k, r_k, a_{k+1})$. Each noise variable ϵ_k is conditionally zero-mean and σ-sub-Gaussian, which means

$$\mathbb{E}\left[\epsilon_k | \mathcal{D}_{k-1}\right] = 0, \quad \text{and} \quad \forall \lambda \in \mathbb{R}, \;\; \mathbb{E}\left[\exp(\lambda \epsilon_k) | \mathcal{D}_{k-1}\right] \leq \exp(\lambda^2 \sigma^2 / 2).$$

Assumption 4.5 (Bounded parameter vector)**.** For some $B_2 > 0$, $\|\boldsymbol{\theta}^*\|_2 \leq B_2$.

Assumption 4.6 (Bounded feature vectors)**.** For some $L_2 > 0$, $\|\phi(a)\|_2 \leq L_2$ for all $a \in \mathcal{A}$.

Assumption 4.7 (Bounded expected reward)**.** For some $C > 0$, $\phi(a)^\top \boldsymbol{\theta}^* \in [-C, C]$ for all $a \in \mathcal{A}$.

We remark that to run our algorithms and evaluate the data-dependent regret bound in Theorem 4.8, we only need to know (upper bounds on) the sub-Gaussian parameter σ and the norm bound B_2. Note that Assumption 4.5 and Assumption 4.6 together imply that Assumption 4.7 must hold with $C \leq L_2 B_2$. We nevertheless state it as a separate assumption because: (a) this is in line with the conventions of other linear bandit analyses (e.g. Section 19.3 of [104]); (b) this leaves open the possibility that a better (than $L_2 B_2$) value for C is known.

4.7.1 OFUL vs AMM-UCB (and CMM-UCB)

For any $\alpha > 0$ (in [3], what we call α is called λ), the OFUL UCB [3] states that

$$\phi(a)^\top \boldsymbol{\theta}^* \leq \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{OFUL},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I}\right)^{-1} \phi(a)},$$
$$\text{where} \quad R_{\mathrm{OFUL},t} = \sigma \sqrt{\ln\left(\det\left(\frac{1}{\alpha}\Phi_t^\top \Phi_t + \boldsymbol{I}\right)\right) + 2\ln(1/\delta)} + \sqrt{\alpha} B_2.$$

$\widehat{\boldsymbol{\theta}}_{\alpha,t}$ is defined in (4.11). For any $\alpha > 0$ and any $\delta \in (0, 1]$, this statement holds with probability at least $1 - \delta$ for all $t \geq 1$ and all $a \in \mathcal{A}_t$. By comparison, our AMM-UCB holds uniformly over all $t \geq 0$, all $a \in \mathcal{A}_t$ *and* all $\alpha > 0$ (i.e., we could optimise the AMM-UCB with respect to α in a data-dependent manner, which would yield our CMM-UCB).

The OFUL UCB is the same as our AMM-UCB, except that the radius quantity $R_{\mathrm{AMM},t}$ is replaced with $R_{\mathrm{OFUL},t}$. The same is true for the LCBs of OFUL and AMM-UCB (with the same radius $R_{\mathrm{OFUL},t}$), so we only focus on the UCBs. In Appendix B.3.2, we show that for any history $a_1, r_1, a_2, r_2, \ldots$ and any $\alpha > 0$, we can chose a sequence of Gaussian mixture distributions (of the form $P_t = \mathcal{N}(\boldsymbol{0}, c\Phi_t \Phi_t^\top)$ for some $c > 0$) such that $R_{\mathrm{AMM},t} < R_{\mathrm{OFUL},t}$. This means that the UCBs of our CMM-UCB and AMM-UCB algorithms are always better than the OFUL UCB.

Note that the mixture distributions which we use to prove this inequality are not necessarily the the mixture distributions that minimise $R_{\mathrm{AMM},t}$. With a better choice of the mixture distributions, $R_{\mathrm{AMM},t}$ will be smaller and the gap between AMM-UCB and OFUL will be greater.

4.7.2 Data-Dependent Regret Bounds

Several authors [48, 3, 153] have shown that the cumulative regret of a UCB algorithm can be upper bounded by the sum of the widths of the confidence sets or confidence bounds that it uses. The width of a confidence set Θ_t at the action a is the difference between the UCB and the LCB at a (i.e., $\max_{\theta \in \Theta_t}\{\phi(a)^\top \boldsymbol{\theta}\} - \min_{\theta \in \Theta_t}\{\phi(a)^\top \boldsymbol{\theta}\}$). In App. B.4.1, we show that if $a_1, a_2, \ldots, a_T$ are the actions selected by our CMM-UCB algorithm, then

$$\sum_{t=1}^{T} \Delta(a_t) \leq \sum_{t=1}^{T} \max_{\theta \in \Theta_{t-1}^{\ell_2}} \{\phi(a_t)^\top \boldsymbol{\theta}\} - \min_{\theta \in \Theta_{t-1}^{\ell_2}} \{\phi(a_t)^\top \boldsymbol{\theta}\}. \tag{4.16}$$

This gives a data-dependent cumulative regret bound for CMM-UCB. AMM-UCB has a similar data-dependent cumulative regret bound. In App. B.4.1, we show that if $a_1, a_2, \ldots, a_T$ are the actions selected by our CMM-UCB algorithm, then

$$\sum_{t=1}^{T} \Delta(a_t) \leq \sum_{t=1}^{T} \mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a_t) - \mathrm{ALCB}_{\Theta_{t-1}^{\ell_2}}(a_t), \tag{4.17}$$

where $\mathrm{AUCB}_{\Theta_t^{\ell_2}}(a)$ is the right-hand-side of (4.11) and $\mathrm{ALCB}_{\Theta_t^{\ell_2}}(a)$ is the equivalent analytic LCB. Since, the analytic UCB/LCB is an upper/lower bound on the numerical UCB/LCB, the bound in Equation (4.17) also holds for the actions selected by CMM-UCB. By substituting in the expressions for the analytic UCB/LCBs, we obtain the following data-dependent cumulative regret bound for CMM-UCB and AMM-UCB.

Theorem 4.8. *Suppose that assumptions 4.4-4.5 hold. For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, any $\delta \in (0,1)$, any $\alpha > 0$ and all $T \geq 1$, with probability at least $1-\delta$, the cumulative regret of both CMM-UCB and AMM-UCB is bounded by*

$$\Delta_{1:T} \leq \sum_{t=1}^{T} 2R_{\mathrm{AMM},t-1} \sqrt{\phi(a_t)^\top \left(\Phi_{t-1}^\top \Phi_{t-1} + \alpha \boldsymbol{I}\right)^{-1} \phi(a_t)}.$$

A proof is given in App. B.4.1. This regret bound tells us that if we choose an adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, such that the radii $R_{\mathrm{AMM},t}$ are small, then we can expect to have small cumulative regret.

4.7.3 Data-Independent Regret Bounds

We now state a data-independent cumulative regret bound for the special case when the sequence of mixture distributions is $P_t = \mathcal{N}(\mathbf{0}, c\Phi_t\Phi_t^\top)$, and $\alpha = \sigma^2/c$, for any $c > 0$.

Theorem 4.9. *Suppose that assumptions 4.4-4.7 hold. If for any $c > 0$, the sequence of mixture distributions is $P_t = \mathcal{N}(\mathbf{0}, c\Phi_t\Phi_t^\top)$, then for all $T \geq 1$, with probability at least $1-\delta$, the cumulative regret of both CMM-UCB and AMM-UCB (with $\alpha = \sigma^2/c$) is bounded by*

$$\Delta_{1:T} \leq \frac{2}{\sqrt{\ln 2}} \max\left\{C, \sigma\sqrt{d\ln\left(1+\frac{cL_2^2T}{\sigma^2 d}\right) + \frac{B_2^2}{c} + 2\ln\frac{1}{\delta}}\right\} \sqrt{dT\ln\left(1+\frac{cL_2^2T}{\sigma^2 d}\right)} = \mathcal{O}(d\sqrt{T}\ln(T)).$$

Proof sketch. Choosing $P_t = \mathcal{N}(\mathbf{0}, c\Phi_t\Phi_t^\top)$ and $\alpha = \sigma^2/c$ means that the two quadratic terms in $R_{\mathrm{AMM},t}^2$ cancel out. We then find a data-independent upper bound for the log det term in $R_{\mathrm{AMM},t}^2$. Following [3], the sum of norms $\sqrt{\phi(a_t)^\top \left(\Phi_{t-1}^\top \Phi_{t-1} + \alpha \boldsymbol{I}\right)^{-1} \phi(a_t)}$ can upper bounded using an elliptical potential lemma. The result is the data-independent bound in the statement of the theorem. □

Note that if we choose $c \propto B_2$, the dependence of the regret bound on B_2 is improved from $\mathcal{O}(B_2)$ to $\mathcal{O}(\sqrt{B_2})$. In App. B.4.2, we give a proof of this special case. In addition we also treat a more

general case when the sequence of mixture distributions is $P_t = \mathcal{N}(\Phi_t \boldsymbol{\theta}_0, \sigma_0^2 \Phi_t \Phi_t^\top)$ and α is any positive number. Focusing on the dependence on d and T, this regret bound (and the more general one in App. B.4.2) is at most $\mathcal{O}(d\sqrt{T}\ln(T))$, which matches OFUL and is minimax optimal up to the $\ln(T)$ factor. If (upper bounds on) σ^2, B_2, L_2 and C are known, then we can evaluate this cumulative regret bound before running the algorithm.

4.8 Experiments

We evaluate our CMM and AMM confidence bounds and our CMM-UCB and AMM-UCB linear bandit algorithms. In our experimental evaluation, we want to compare the tightness of our UCBs against UCBs for linear (reward) functions. We also want to compare the empirical performance of our CMM-UCB and AMM-UCB algorithms against other linear bandit algorithms with comparable worst-case regret guarantees.

4.8.1 Upper and Lower Confidence Bounds

First, we evaluate our CMM and AMM confidence bounds. We evaluate their tightness and observe how the choice of the mixture distribution affects the resulting confidence bounds.

Compared Methods. We evaluate the following upper/lower confidence bounds: (a) *CMM-UCB:* our numerical UCBs/LCBs from Section 4.6.2; (b) *AMM-UCB:* our analytic UCBs/LCBs from Theorem 4.3; (c) *OFUL:* the UCBs/LCBs used by the OFUL algorithm [3]; (d) *Bayes:* a Bayesian credible interval constructed from the Bayesian posterior for linear regression with a Gaussian prior and likelihood (see Appendix B.5.1 for details).

Experimental Setup. We conduct experiments on randomly generated linear functions of the form $f(\boldsymbol{x}) = \phi(\boldsymbol{x})^\top \boldsymbol{\theta}^*$, with inputs $\boldsymbol{x} \in \mathbb{R}^{d_\mathcal{X}}$ and $\boldsymbol{\theta}^* \in \mathbb{R}^d$. In each experiment, $\boldsymbol{\theta}^*$ is drawn from a standard Gaussian distribution and if necessary scaled down to $\|\boldsymbol{\theta}^*\|_2 \leq 10 =: B_2$. For the feature map $\phi : \mathbb{R}^{d_\mathcal{X}} \to \mathbb{R}^d$, we use Random Fourier Features (cf. Algorithm 1 of [140]). We investigate the properties of upper and lower confidence bounds constructed from random data sets $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^T$, where $y_t = \phi(\boldsymbol{x}_t)^\top \boldsymbol{\theta}^* + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 0.1$. Unless stated otherwise, we use the standard mixture distributions $P_t = \mathcal{N}(\mathbf{0}, \Phi_t \Phi_t^\top)$ for our CMM and AMM confidence bounds. For our AMM confidence bounds and OFUL, we always choose $\alpha = \sigma^2$. For each confidence bound, we set $\delta = 0.01$.

UCB/LCB Tightness. Figure 4.1 shows the data $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^T$ and the CMM, AMM and OFUL UCBs/LCBs for a randomly generated linear function with $d_\mathcal{X} = 1$ and $d = 20$. In this example, the confidence bounds of CMM-UCB are slightly tighter than those of AMM-UCB, which are considerably tighter than those of OFUL. Next, we investigate the tightness of the confidence bounds for functions with higher dimensional inputs ($d_\mathcal{X} = 10$), a range of data set sizes ($T \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$) and a range feature vector dimensions ($d \in \{1, 2, 5, 10, 20, 50, 100\}$). For each T and d, we sample a random feature map ϕ and weight vector $\boldsymbol{\theta}^*$ of appropriate size. Then, we sample random training data $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^T$ and random test points $\{\boldsymbol{x}'_t\}_{t=1}^{100}$, where $\boldsymbol{x}_t$ and $\boldsymbol{x}'_t$ are drawn uniformly from the $d_\mathcal{X}$-dimensional unit hypercube. Finally, we use the training data to construct confidence bounds with each method and calculate

the average width at the test points. Figure 4.2 shows the average width of the CMM, AMM and OFUL confidence bounds with: $d = 10$ and varying T (left); $T = 100$ and varying d (right). We observe the same pattern at every d and T: our CMM confidence bounds are the tightest, followed by AMM and then OFUL. These results agree with our theoretical comparison of the CMM, AMM and OFUL confidence bounds (see Section 4.7.1).

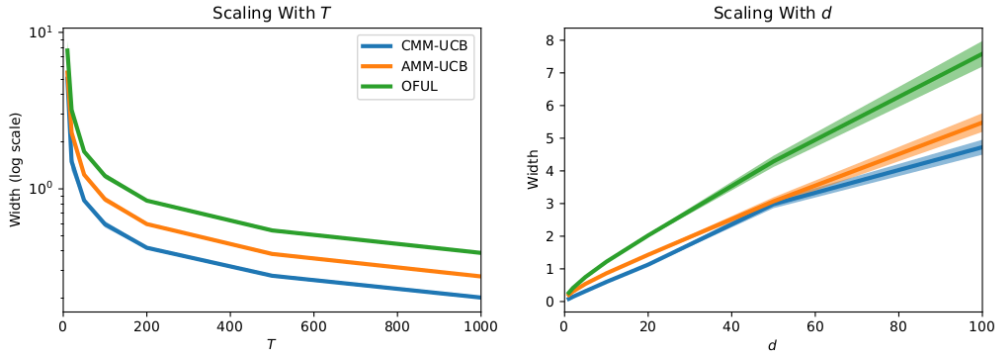


Figure 4.2: The confidence bound width for different data set sizes T and feature dimensions d. We show the mean and standard devation of the widths over 10 runs.

Effect of the Mixture Distributions. Figure 4.3 shows our CMM confidence bounds (left) and a the Bayesian credible interval (right) for a randomly generated linear function with $d_{\mathcal{X}} = 1$ and $d = 20$. In each row of Figure 4.3, we use different choices for different for the mixture distributions/priors. We use equivalent mixture distributions/priors for both methods. If the Bayesian credible interval uses the prior $\boldsymbol{\theta}^* \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, then the mixture distribution for the CMM confidence bounds is the induced distribution $\Phi_t \boldsymbol{\theta}^*$, which is $P_t = \mathcal{N}(\Phi_t \boldsymbol{\mu}_0, \Phi_t \boldsymbol{\Sigma}_0 \Phi_t^\top)$. Note that, since we are using Gaussian noise to generate the data, the likelihood function for the Bayesian credible interval is well-specified.

In Figure 4.3 (top row), we see that when the mixture distribution/prior ($P_t = \mathcal{N}(\mathbf{0}, B_2 \Phi_t \Phi_t^\top)$) is uninformative, the CMM confidence bounds and the Bayesian credible interval are almost the same in regions near the data, but the CMM confidence bounds are noticeably better in regions far away from the data. In the middle row of Figure 4.3, where the mixture distribution/prior ($P_t = \mathcal{N}(\Phi_t \boldsymbol{\theta}^*, 0.1 \Phi_t \Phi_t^\top)$) is centred at the ground-truth function values/parameter vector, we observe that the CMM confidence bounds are slightly looser than the Bayesian credible interval. In the bottom row of Figure 4.3, where the mixture distribution/prior ($P_t = \mathcal{N}(-\Phi_t \boldsymbol{\theta}^*, 0.1 \Phi_t \Phi_t^\top)$) is misspecified (in a Bayesian sense), the CMM confidence bounds become looser, but still contain the ground-truth function. In contrast, the Bayesian credible interval no longer contains the ground-truth function.

In summary, our CMM confidence bounds give valid uncertainty estimates for any mixture distribution, whereas the Bayesian credible interval may not if the prior is chosen badly. While the Bayesian credible interval was slightly tighter when the prior was chosen very well, our CMM confidence bounds provided as good as or tighter intervals otherwise.

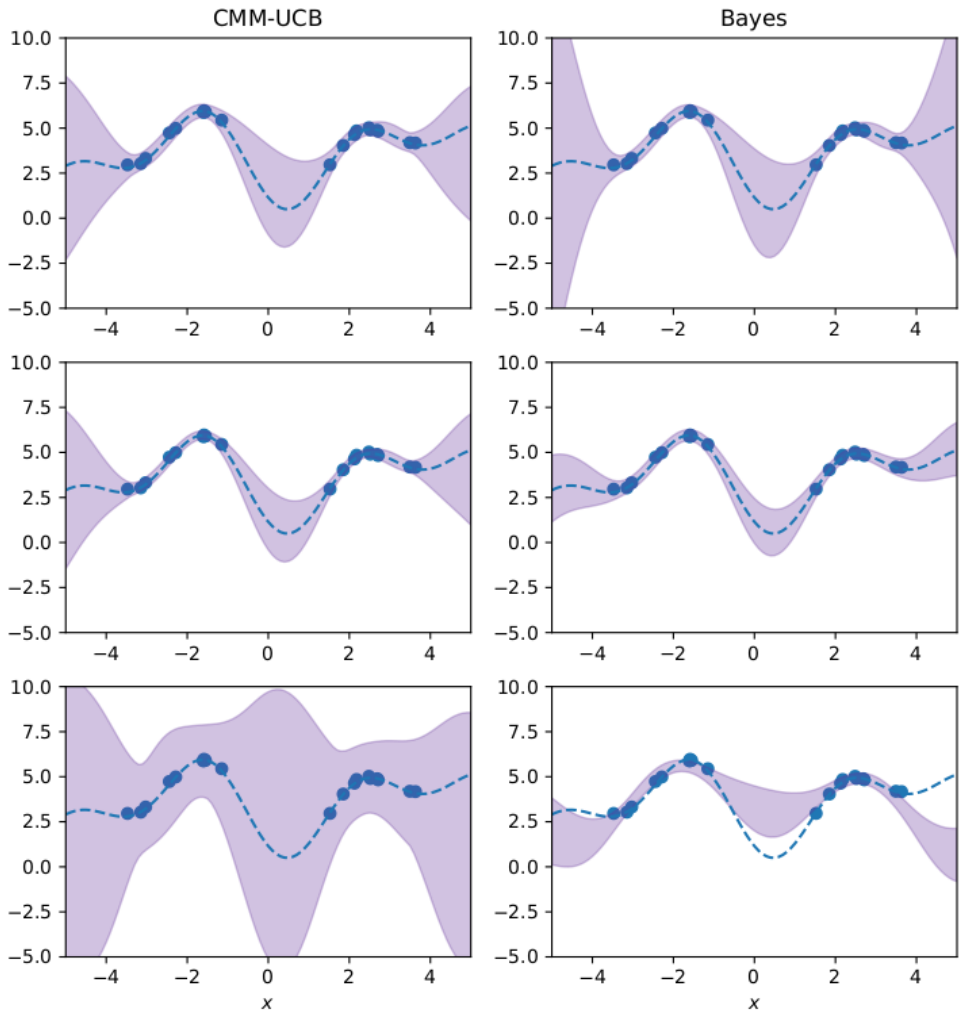


Figure 4.3: The upper and lower confidence bounds of our CMM-UCB method (left) and Bayesian posterior credible intervals (right) with different choices of the prior. The top row uses the prior $\boldsymbol{f}_t \sim \mathcal{N}(\mathbf{0}, B_2\Phi_t\Phi_t^\top)$ for CMM-UCB and $\boldsymbol{\theta}^* \sim \mathcal{N}(\mathbf{0}, B_2\boldsymbol{I})$) for Bayes. The middle row uses an informative prior: $\boldsymbol{f}_t \sim \mathcal{N}(\Phi_t\boldsymbol{\theta}^*, 0.1\Phi_t\Phi_t^\top)$ for CMM-UCB and $\boldsymbol{\theta}^* \sim \mathcal{N}(\boldsymbol{\theta}^*, 0.1\boldsymbol{I})$) for Bayes. The bottom row uses a misspecified prior: $\boldsymbol{f}_t \sim \mathcal{N}(-\Phi_t\boldsymbol{\theta}^*, 0.1\Phi_t\Phi_t^\top)$ for CMM-UCB and $\boldsymbol{\theta}^* \sim \mathcal{N}(-\boldsymbol{\theta}^*, 0.1\boldsymbol{I})$) for Bayes.

4.8.2 Linear Bandits

We now investigate whether our tighter (than OFUL) confidence bounds translate to better UCB algorithms for linear bandits.

Compared Methods. (a) *CMM-UCB:* cf. Section 4.6.2; (b) *AMM-UCB:* cf. Section 4.6.3; (c) *OFUL:* the OFUL algorithm [3]; (d) *IDS:* the frequentist Information Directed Sampling (IDS) algorithm [95], specifically the deterministic DIDS-F version; (e) *Freq-TS:* Thompson Sampling with posterior covariance inflation [9], which we call Frequentist Thompson Sampling.

Experimental Setup. We use our linear bandit algorithms to optimise the hyperparameters of a kernel Support Vector Machine (SVM) for three classification data sets from the UCI Machine Learning Repository [54]: Raisin [44], Maternal [10], and Banknotes. The expected reward function $f^*(a)$ is the average test set accuracy of a kernel SVM trained using an ARD RBF kernel with hyperparameters $a = (C, \boldsymbol{\gamma})$. $C > 0$ is the regularisation hyperparameter and $\boldsymbol{\gamma} \in \mathbb{R}^{d_{\mathcal{X}}}$ is a vector of lengthscales, where $d_{\mathcal{X}}$ is the number of covariates in the classification problem (which is between 4 and 7). The observed reward r_t is the validation set accuracy at a_t. We record the mean test accuracy (expected reward), which is roughly equivalent to the cumulative regret, and the maximum test accuracy achieved over $T = 500$ steps.

For the feature map ϕ, we use a neural network layer with 20 outputs and random weights. We choose $\sigma = 0.05$ for the sub-Gaussian parameter and $B_2 = 10$, which means we are assuming that $f^*(a) \approx \phi(a)^\top \boldsymbol{\theta}^*$ for some $\|\boldsymbol{\theta}^*\|_2 \leq 10$. We use the standard mixture distributions $P_t = \mathcal{N}(\mathbf{0}, \Phi_t \Phi_t^\top)$ for our CMM-UCB and AMM-UCB algorithms. For AMM-UCB, OFUL, and IDS we set $\alpha = \sigma^2$. For algorithm, we set $\delta = 0.01$.

Table 4.1: Average test accuracy and maximum test accuracy of our CMM-UCB, AMM-UCB, OFUL, IDS and Freq-TS in the SVM hyperparameter tuning problems after $T = 500$ rounds. We report the mean and standard deviation over 100 repetitions.

	Raisin		Maternal		Banknotes	
	Mean Acc	Max Acc	Mean Acc	Max Acc	Mean Acc	Max Acc
CMM-UCB (Ours)	**0.818** ± 0.018	**0.893** ± 0.019	**0.744** ± 0.020	**0.829** ± 0.023	**0.954** ± 0.005	**1.000** ± 0.000
AMM-UCB (Ours)	0.800 ± 0.017	0.892 ± 0.020	0.736 ± 0.020	**0.829** ± 0.023	0.948 ± 0.005	**1.000** ± 0.000
OFUL	0.764 ± 0.019	0.891 ± 0.019	0.722 ± 0.019	0.827 ± 0.022	0.929 ± 0.006	**1.000** ± 0.000
IDS	0.706 ± 0.048	0.891 ± 0.020	0.714 ± 0.019	0.827 ± 0.024	0.926 ± 0.007	**1.000** ± 0.000
Freq-TS	0.527 ± 0.022	0.884 ± 0.019	0.616 ± 0.018	0.823 ± 0.022	0.808 ± 0.012	**1.000** ± 0.000

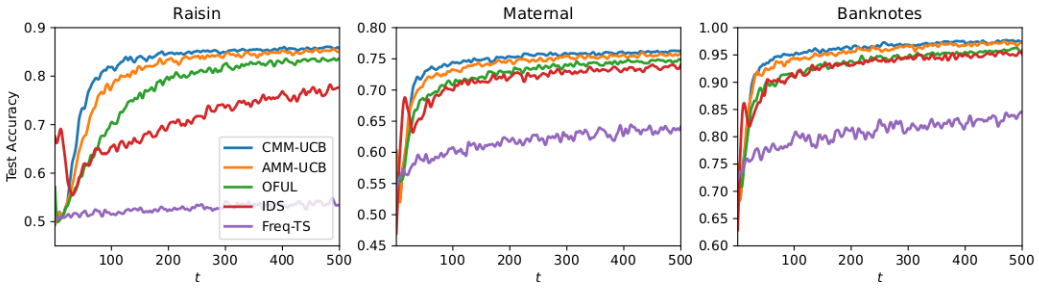


Figure 4.4: The smoothed per-round test accuracy (expected reward) of our CMM-UCB, AMM-UCB, OFUL, IDS and Freq-TS in the SVM hyperparameter tuning experiments. We show the mean reward over 100 runs of each experiment, after Gaussian kernel smoothing.

Results. Figure 4.4 shows the average test accuracy (expected reward) obtained by each bandit algorithm for each data set and at each round $t = 1, \ldots, 500$. Our CMM-UCB and AMM-UCB algorihtms outperform all the other methods. From the reward curves of CMM-UCB (blue), AMM-UCB (orange) and OFUL (green), we observe that CMM-UCB outperforms AMM-UCB, which outperforms OFUL. Therefore, we conclude that our tighter confidence bounds do indeed lead to UCB algorithms with improved performance.

4.8.3 Adaptive Mixture Distributions

Finally, we investigate the behaviour of our CMM-UCB and AMM-UCB algorithms when using the adaptive mixture distributions described in Section 4.6.4.

Experimental Setup. First, we compare our CMM confidence bounds with the standard mixture distributions $P_t = \mathcal{N}(\mathbf{0}, \Phi_t \Phi_t^\top)$ and the adaptive mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, with $\boldsymbol{\mu}_t$ as in (4.15) and $\boldsymbol{T}_t$ as in (4.14)). For the adaptive mixture distributions, we set $m(a) = 0$, $k(a, a') = \phi(a)^\top \phi(a')$ and $\beta = 4\sigma^2$. With each choice of the mixture distributions, we plot the CMM confidence bounds for a randomly generated linear (in Fourier features) function with $d_{\mathcal{X}} = 1$ and $d = 20$.

Next, we compare the AMM-UCB algorithm with the same standard and adaptive mixture distributions in the SVM hyperparameter tuning problem for the Raisin data set (as described in Sec. 4.8.2). We record the test accuracy and the value of the radius $R_{\text{AMM},t}$ in each round.

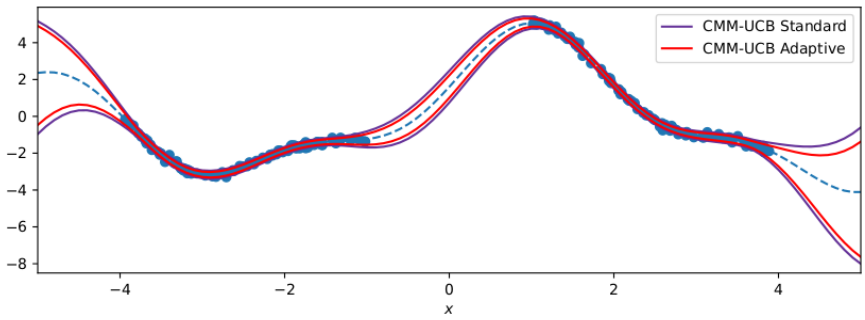


Figure 4.5: Our CMM confidence bounds with the standard mixture distributions (purple) and the adaptive mixture distributions (red).

Results. In Figure 4.5, we observe that our CMM confidence bounds are slightly tighter when we use the adaptive mixture distributions. In Figure 4.6 (right), we observe that AMM-UCB achieves slightly higher test accuracy with the adaptive mixture distributions. In Figure 4.6 (left), we see that the radius $R_{\text{AMM},t}$ is smaller, and therefore the confidence bounds are tighter, with the adaptive mixture distributions. Interestingly, $R_{\text{AMM},t}$ grows with T when we use the standard mixture distributions, but appears to be uniformly (over T) bounded by a constant when we use the adaptive mixture distributions. Overall, it appears that using the adaptive mixture distributions leads to slightly tighter confidence bounds and improved bandit algorithms.

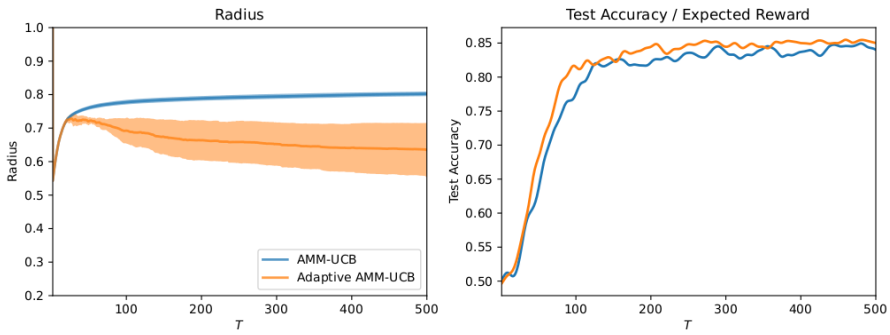


Figure 4.6: The radius $R_{\text{AMM},t}$ (left) and test accuracy (right) for our AMM-UCB algorithm with the standard mixture distributions (blue) and the adaptive mixture distributions (orange). On the left, we show the mean and standard deviation of the radius $R_{\text{AMM},t}$ over 10 runs. On the right, we plot the smoothed mean reward over 10 runs, with Gaussian kernel smoothing.

4.9 Conclusion

In this chapter, we developed a novel PAC-Bayes-style tail bound for adaptive martingale mixtures and showed that it can be used to construct confidence sequences for linear bandits that allow for efficient action selection via convex programming. We proved that our CMM-UCB and AMM-UCB algorithms match the worst-case regret of OFUL. We proved, and verified empirically, that our confidence bounds are tighter than those used by OFUL. In our expriments, we found that this allowed our CMM-UCB and AMM-UCB algorithms to achieve greater average and maximum reward in several hyperparameter tuning problems.

We believe that further investigation into tail bounds (or PAC-Bayes-style bounds as in (4.5)) with adaptive mixture distributions/priors (as described in Section 4.6.4) is an exciting direction to pursue in future research. Focusing on the application of adaptive mixture distributions to the linear bandit problem, we suspect that, by using adaptive mixture distributions, it is possible to obtain data-independent regret bounds for CMM-UCB or AMM-UCB with improved growth rates.

Our data-independent regret bound in Theorem 4.9 used the fact that the radius $R_{\text{AMM},T}$ can be upper bounded by a data-independent quantity of order $\mathcal{O}(\sqrt{d\ln(T)})$ (see Appendix B.4.2), when we use the standard mixture distributions. In Figure 4.6, we saw that $R_{\text{AMM},t}$ does appear to grow roughly logarithmically with T when we use the standard mixture distributions. However, we also saw in Figure 4.6 that $R_{\text{AMM},t}$ appears to be bounded by a constant when we use the adaptive mixture distributions. If, when using adaptive mixture distributions, we could prove a data-independent bound on $R_{\text{AMM},T}$ of order $\mathcal{O}(\sqrt{d})$, then we would be able to improve our data-independent cumulative regret bounds to $\mathcal{O}(d\sqrt{T\ln(T)})$ (rather than $\mathcal{O}(d\sqrt{T}\ln(T))$). This would be within a $\sqrt{\ln(T)}$ factor of the minimax lower bound $\Omega(d\sqrt{T})$ (see e.g. Theorem 24.1 or Theorem 24.2 of [104] for a lower bound).

A limitation of both the CMM-UCB and AMM-UCB algorithms is that their regret bound grows linearly in d, so they may not perform well when the feature vectors are high-dimensional. Note however, that the minimax lower bound $\Omega(d\sqrt{T})$ implies that no algorithm can be guaranteed to have good worst-case performance in high-dimensional linear bandit problems.

Chapter 5

PAC-Bayes-Style Algorithms for Sparse Linear Bandits

5.1 Introduction

In the previous chapter, we developed the Convex Martingale Mixture Upper Confidence Bound (CMM-UCB) algorithm for stochastic linear bandits. We showed that this algorithm has cumulative regret no worse than $\mathcal{O}(d\sqrt{T}\ln(T))$, which means it is guaranteed to perform well when the dimension d of the feature vectors is small relative to the number of rounds T.

However, the linear growth rate of the regret in the feature vector dimension, combined with the assumption that the reward function is linear in the feature vectors, brings about an unfortunate trade-off in the selection of the feature map. On the one hand, we would like to choose a low-dimensional feature map so that the cumulative regret is not too high. On the other hand, we need to choose a sufficiently high-dimensional feature map so that the expected reward function can be expressed as a linear function of the feature vectors. Moreover, it is often the case that there are many candidate features to choose from and it not obvious which should be included in the feature map. For example, suppose we use a random feature map (e.g. Random Fourier features [140]). It is likely that many of the features are not useful, but it is not obvious in advance which features are useful. One would like to be able to allow a linear bandit algorithm to use many features, but still suffer low regret if it turns out that only a small number of features were useful. This can achieved by linear bandit algorithms that exploit sparsity.

In sparse stochastic linear bandit problems, the unknown parameter vector $\boldsymbol{\theta}^*$ of the expected reward function $\phi(a)^\top \boldsymbol{\theta}^*$ is assumed to contain only a small number (say s) of non-zero elements. A sparse parameter vector can be much easier to estimate than a non-sparse (or dense) parameter vector. As a result, one can design algorithms for sparse linear bandit problems that have improved dependence on the dimension d. At the same time, sparsity assumptions are often not too restrictive because natural data can typically be well-approximated by sparse linear combinations of appropriate basis functions. All of this means that sparse linear bandits are a suitable model for real-world bandit problems.

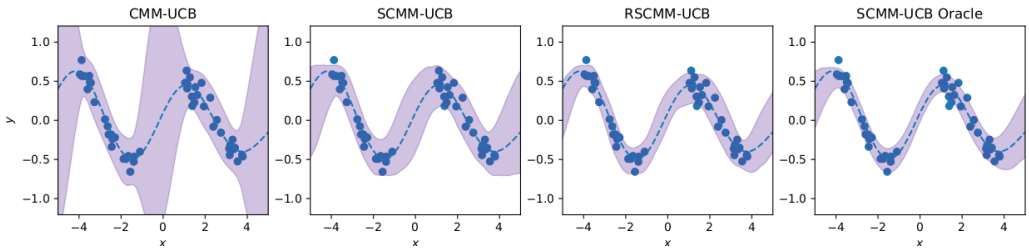


Figure 5.1: The upper and lower confidence bounds of CMM-UCB (left), SCMM-UCB (left-middle), RSCMM-UCB (right-middle), and SCMM-UCB Oracle (right) for a test function which is linear in a 100-dimensional feature map, but whose parameter vector has only 5 non-zero elements.

Contributions. In this chapter, we design and analyse PAC-Bayes-style algorithms for sparse linear bandits. We re-use our general-purpose tail bound for martingale mixtures from Theorem 4.1 to construct confidence sequences for sparse linear bandits. The maximisation problem to compute the corresponding upper confidence bounds (UCBs) is still convex, which allows us to derive modified versions of the CMM-UCB algorithm that exploit sparsity. We propose two sparse versions of the CMM-UCB algorithm: (a) *Sparse Convex Martingale Mixture UCB (SCMM-UCB):* We replace the $\|\boldsymbol{\theta}^*\|_2 \leq B_2$ constraint in the confidence set in Corollary 4.2 with the constraint $\|\boldsymbol{\theta}^*\|_1 \leq B_1$; (b) *Restricted Sparse Convex Martingale Mixture UCB (RSCMM-UCB):* We add an initial feature selection phase to the SCMM-UCB algorithm, in which we use the thresholded Lasso estimate [172, 122, 189] to select a subset of available features.

Figure 5.1 highlights a key finding. When the unknown ground-truth function is sparse and linear, the upper and lower confidence bounds used by our proposed SCMM-UCB and RSCMM-UCB algortihms can be much tighter than the confidence bounds used by CMM-UCB. Moreover, they can be almost as tight as upper and lower confidence bounds that can only be constructed when the locations of the non-zero elements of $\boldsymbol{\theta}^*$ are known in advance (SCMM-UCB Oracle).

5.2 Related Work

Several variants of the sparse stochastic linear bandit problem have been studied. Abbasi-Yadkori et al. [5] proposed an online-to-confidence-set conversion and used it design algorithms for a sparse linear bandit problem where the action set is an arbitrary fixed subset of a finite-dimensional vector space. Ignoring logarithmic terms, the online-to-confidence-set approach achieves an $\mathcal{O}(\sqrt{sdT})$ regret bound, where s is the number of non-zero elements in $\boldsymbol{\theta}^*$ and T is the number of rounds. A matching lower bound for this problem was later established by Lattimore and Szepesvári [104].

To achieve regret bounds with better than $\mathcal{O}(\sqrt{d})$ dependence on the feature vector dimension, several authors have studied sparse linear bandits with additional restrictions on the action set, the feature map ϕ, or the parameter vector $\boldsymbol{\theta}^*$. For the case where the action set is the binary hypercube and the feature map is the identity function, Lattimore et al. [103] developed the selective explore-then-commit algorithm, and showed that it has an $\mathcal{O}(s\sqrt{T})$ regret bound. Hao et al. [76] and Jang et al. [85] proposed explore-the-sparsity-then-commit algorithms that have $\mathcal{O}(s^{2/3}T^{2/3})$ regret bounds when the action set spans $\mathbb{R}^{d_{\mathcal{A}}}$ and the feature map is the identity function. Under an additional minimum signal condition, meaning each non-zero element of $\boldsymbol{\theta}^*$ satisfies $|\theta_i^*| \geq m$ for

$m > 0$, Hao et al. [76] and Jang et al. [85] also proposed restricted phase elimination algorithms, which have $\mathcal{O}(\sqrt{sT\ln(K)})$ regret bounds when the action set is finite with K elements.

It has recently become popular to study the stochastic contextual linear bandit problem with sparsity assumptions. In this setting, an action set containing a fixed finite number of feature vectors is randomly generated at the start of each round. Under suitable conditions on the distribution from which the actions sets are drawn, Bastani and Bayati [26], Wang et al. [184], Kim and Paik [92], Oh et al. [129], Ariu et al. [18] and Chakraborty et al. [40] have all proposed algorithms that achieve regret bounds with $\mathcal{O}(\ln(d))$ or $\mathcal{O}(\text{poly}(\ln(d)))$ dependence on the feature vector dimension.

5.3 Problem Statement and Background

We consider the stochastic linear bandit problem with a sparsity assumption. A learner plays a game over a sequence of T rounds, where T may not be known in advance. In each round t, the learner must choose an action $a_t \in \mathcal{A}$ from a fixed action set $\mathcal{A}$. The learner then receives a reward $r_t = \phi(a_t)^\top \boldsymbol{\theta}^* + \epsilon_t$. The feature map $\phi : \mathcal{A} \to \mathbb{R}^d$ is a known function that maps actions to d-dimensional feature vectors. $\boldsymbol{\theta}^* \in \mathbb{R}^d$ is an unknown parameter with ℓ_1 norm bounded by some known $B_1 > 0$, i.e. $\|\boldsymbol{\theta}^*\|_1 \leq B_1$. We assume that $\boldsymbol{\theta}^*$ satisfies a hard sparsity property:

$$\|\boldsymbol{\theta}^*\|_0 := \sum_{i=1}^{d} \mathbb{I}\{\theta_i^* \neq 0\} = s. \tag{5.1}$$

We consider both simple regret and cumulative regret as objectives of the learner. For a single round, we define the regret as

$$\Delta(a_t) = \phi(a^*)^\top \boldsymbol{\theta}^* - \phi(a_t)^\top \boldsymbol{\theta}^*, \tag{5.2}$$

where $a^* = \text{argmax}_{a \in \mathcal{A}} \{\phi(a)^\top \boldsymbol{\theta}^*\}$. After T rounds, the *simple regret* is

$$\Delta_T := \min_{t \in \{1,\dots,T\}} \{\Delta(a_t)\} = \phi(a^*)^\top \boldsymbol{\theta}^* - \max_{t \in \{1,\dots,T\}} \left\{\phi(a_t)^\top \boldsymbol{\theta}^*\right\}, \tag{5.3}$$

which is the difference between the maximum of the reward function and the maximum restricted to the selected actions $a_1, \dots, a_T$. After T rounds, the *cumulative regret* is

$$\Delta_{1:T} := \sum_{t=1}^{T} \Delta(a_t) = \sum_{t=1}^{T} \phi(a^*)^\top \boldsymbol{\theta}^* - \phi(a_t)^\top \boldsymbol{\theta}^*, \tag{5.4}$$

which is the difference between the total expected reward of the optimal strategy and the learner. Since a minimum is upper bounded by an average, we have

$$\Delta_T = \min_{t \in \{1,\dots,T\}} \{\Delta(a_t)\} \leq \frac{1}{T} \sum_{t=1}^{T} \Delta(a_t) = \frac{\Delta_{1:T}}{T}. \tag{5.5}$$

We design algorithms primarily with cumulative regret bounds in mind, but due to (5.5), this approach yields simple regret bounds as well.

Notation. For any integer $n \geq 1$, let $[n] = \{1, \ldots, n\}$. For a set of indices $S \subseteq [d]$, let S^{c} denote its complement. For a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and a set of indices S, let $\boldsymbol{\theta}_S \in \mathbb{R}^{|S|}$ denote the subvector with elements indexed by S. Let $\mathrm{supp}(\boldsymbol{\theta})$ denote the support of the vector $\boldsymbol{\theta}$, which is the set of indices where $\theta_i \neq 0$. For a matrix $\boldsymbol{A}$, let $\nu_{\min}(\boldsymbol{A})$ and $\nu_{\max}(\boldsymbol{A})$ denote its minimum and maximum eigenvalues.

Lasso Estimation. Let $\Phi_t = [\phi(a_1), \ldots, \phi(a_t)]^\top$ denote the $t \times d$ matrix of feature vectors, which we will call the design matrix. Let $\boldsymbol{r}_t = [r_1, \ldots, r_t]^\top$ denote the vector of rewards. For a penalty strength $\eta > 0$, we define the regularised Lasso estimate [172] as

$$\widehat{\boldsymbol{\theta}} := \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \frac{1}{2t} \left\| \Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t \right\|_2^2 + \eta \left\| \boldsymbol{\theta} \right\|_1 \right\}. \tag{5.6}$$

The Lasso estimate is widely used for both prediction and feature selection in high-dimensional sparse linear regression problems because: (a) the ℓ_1 norm penalty encourages sparse solutions; (b) (5.6) is a convex program that can be solved efficiently for reasonably large d. We use $\widehat{S} = \mathrm{supp}(\widehat{\boldsymbol{\theta}})$ to denote the support of the Lasso estimate.

In practice, it is not uncommon for the Lasso estimate to have several components that are very close to 0, but not exactly zero. For this reason, it is a common practice to apply additional thresholding to remove these small components. For a threshold level τ, which satisfies $0 \leq \tau < m$, the thresholded Lasso estimate (see Eq. (2.7) of [27]) is $\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}$, where

$$\widehat{S}_\tau := \left\{ i \in \{1, \ldots, d\} \middle| |\widehat{\theta}_i| \geq \tau \right\}. \tag{5.7}$$

We use the thresholded Lasso estimate to estimate the support of $\boldsymbol{\theta}^*$ in the feature selection phase of our RSCMM-UCB algorithm.

Compatibility. In the analysis of the feature selection phase of our RSCMM-UCB algorithm, we use a compatibility condition [34]. Let S be a set of indices and κ a positive constant. Define

$$C(S, \kappa) := \left\{ \boldsymbol{A} \in \mathbb{R}^{d \times d} \middle| \forall \boldsymbol{v} \in \mathbb{R}^d \text{ such that } \left\| \boldsymbol{v}_{S^{\mathsf{c}}} \right\|_1 \leq 3 \left\| \boldsymbol{v}_S \right\|_1, \left\| \boldsymbol{v}_S \right\|_1^2 \leq \frac{|S| \boldsymbol{v}^\top \boldsymbol{A} \boldsymbol{v}}{\kappa^2} \right\}. \tag{5.8}$$

We say that a matrix $\boldsymbol{\Sigma}$ satisfies the *compatibility condition* if $\boldsymbol{\Sigma} \in C(S, \kappa)$. To understand what this condition means, suppose that we have a positive-definite matrix $\boldsymbol{\Sigma}$ with minimum eigenvalue $\nu > 0$. Since $\boldsymbol{\Sigma}$ is positive-definite, for all $S \subseteq [d]$ and all vectors $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$|S| \boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v} \geq |S| \nu \left\| \boldsymbol{v} \right\|_2^2 \geq |S| \nu \left\| \boldsymbol{v}_S \right\|_2^2 \geq \nu \left\| \boldsymbol{v}_S \right\|_1^2.$$

Therefore, the compatibility condition is a weaker version of positive-definiteness. Roughly speaking, if a matrix $\boldsymbol{\Sigma}$ satisfies $\boldsymbol{\Sigma} \in C(S, \kappa)$, then for all for all vectors $\boldsymbol{v}$ in the cone $\{\boldsymbol{v} \in \mathbb{R}^d | \left\| \boldsymbol{v}_{S^{\mathsf{c}}} \right\|_1 \leq 3 \left\| \boldsymbol{v}_S \right\|_1\}$, $\boldsymbol{\Sigma}$ behaves as if it is positive-definite with minimum eigenvalue κ^2.

5.4 Confidence Sequences for Sparse Linear Bandits

The key component of our SCMM-UCB and RSCMM-UCB algorithms are confidence sequences. We use the general-purpose tail bound for adaptive martingale mixtures from Theorem 4.1 of Chapter 4 to construct confidence sequences for sparse linear bandits. First, we recall the tail bound in Theorem 4.1 and the general setting in which it holds.

We are given a filtration $(\mathcal{D}_t | t \in \mathbb{N})$, a sequence of adapted random functions $(Z_t : \mathbb{R} \to \mathbb{R} | t \in \mathbb{N})$, a sequence of predictable random variables $(\lambda_t | t \in \mathbb{N})$ and a fixed sequence of function values $(g_t | t \in \mathbb{N})$. We use the shorthand $\boldsymbol{g}_t = (g_1, \ldots, g_t)$ and $\boldsymbol{\lambda}_t = (\lambda_1, \ldots, \lambda_t)$. We call a data-dependent sequence of probability distributions $(P_t | t \in \mathbb{N})$ an *adaptive sequence of mixture distributions* if: (a) P_t is a distribution over $\boldsymbol{g}_t \in \mathbb{R}^t$; (b) P_t is $\mathcal{D}_{t-1}$-measurable; (c) the distributions are consistent in the sense that their marginals coincide, i.e. $\int P_t(\boldsymbol{g}_t) dg_t = P_{t-1}(\boldsymbol{g}_{t-1})$ for all t. Theorem 4.1 provides a time-uniform tail bound for the martingale mixture $(\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)] | t \in \mathbb{N})$, where

$$M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t) = \exp\left(\sum_{k=1}^{t} \lambda_k Z_k(g_k) - \psi_k(g_k, \lambda_k)\right), \quad \psi_t(g_t, \lambda_t) := \ln\left(\mathbb{E}\left[\exp(\lambda_t Z_t(g_t)) | \mathcal{D}_{t-1}\right]\right). \quad (5.9)$$

Theorem 4.1 states that, for any level $\delta \in (0, 1]$, any sequence of predictable random variables $(\lambda_t | t \in \mathbb{N})$ and any adaptive sequence of mixture distributions $(P_t | t \in \mathbb{N})$, with probability at least $1 - \delta$, we have

$$\forall t \geq 1, \quad \ln\left(\mathop{\mathbb{E}}_{\boldsymbol{g}_t \sim P_t}\left[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)\right]\right) \leq \ln(1/\delta). \quad (5.10)$$

This statement is proved in Appendix B.1. If $\psi_k(g_k, \lambda_k)$ is replaced by an upper bound on $\psi_k(g_k, \lambda_k)$, the inequality in Equation 5.10 still holds.

5.4.1 SCMM Confidence Sequence

For our first confidence sequence for sparse linear bandits, we specialise the inequality in Equation 5.10 in the same way as in Chapter 4. For the filtration $(\mathcal{D}_t | t \in \mathbb{N})$, we set $\mathcal{D}_t$ to be the σ-algebra generated by $(a_1, r_1, \ldots, a_t, r_t, a_{t+1})$. We choose $Z_t(g_t) = (g_t - \phi(a_t)^\top \boldsymbol{\theta}^*)\epsilon_t$, $\lambda_t \equiv 1/\sigma^2$ and we use Gaussian mixture distributions, i.e. $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$. This choice of $Z_t(g_t)$ allows us to upper bound $\psi_t(g_t, \lambda_t)$ by exploiting the sub-Gaussian property of ϵ_t. These choices of $Z_t(g_t)$, λ_t and P_t allow us to derive a relatively simple convex quadratic constraint for $\boldsymbol{\theta}^*$ from the inequality in Equation (5.10). As in Section 4.5.2, we have

$$\|\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 \leq (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2}\right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{r}_t) + \sigma^2 \ln\det\left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2}\right) + 2\sigma^2 \ln\frac{1}{\delta} =: R_{\mathrm{MM},t}^2. \quad (5.11)$$

See Section 4.5 for the derivation of this inequality. In the previous chapter, we combined the constraint Equation (5.11) with the constraint $\|\boldsymbol{\theta}^*\|_2 \leq B_2$. This time, we combine the constraint in Equation (5.11) with the constraint $\|\boldsymbol{\theta}^*\|_1 \leq B_1$.

Corollary 5.1 (SCMM Confidence Sequence)**.** *For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, it holds with probability at least $1 - \delta$ that for all $t \geq 1$ simultaneously $\boldsymbol{\theta}^*$ lies in*

the set

$$\Theta_t^{\ell 1} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \,\middle|\, \|\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t\|_2 \le R_{\mathrm{MM},t} \quad \text{and} \quad \|\boldsymbol{\theta}\|_1 \le B_1 \right\}. \tag{5.12}$$

The set $\Theta_t^{\ell 1}$ is the intersection of a d-dimensional ellipse and a d-dimensional ℓ_1 ball of radius B_1. The rightmost plot on the title page depicts $\Theta_t^{\ell 1}$ for the case when $d = 2$. This confidence sequence is related to the constrained form of the Lasso estimate (see e.g. Equation (11.2) of [77]). In particular, whenever $\Theta_t^{\ell 1}$ is non-empty, it contains the constrained form of the Lasso estimate.

5.4.2 Restricted SCMM Confidence Sequence

For our second confidence sequence for sparse linear bandits, we incorporate an initial feature selection phase, which is described in more detail in Section 5.5.2. In the feature selection phase, we sample T_1 actions from a fixed *exploration distribution* ρ. We then use the *exploration data* $\{(a'_t, r'_t)\}_{t=1}^{T_1}$ to compute the support $\widehat{S}_\tau$ of the thresholded Lasso estimate $\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}$. The *restricted feature map* is defined as $\widetilde{\phi}(a) = \phi(a)_{\widehat{S}_\tau}$, which is the vector of only the features corresponding to the indices in $\widehat{S}_\tau$. For validity of this confidence sequence, we require that $\widehat{S}_\tau \supseteq S^*$, where $S^* = \mathrm{supp}(\boldsymbol{\theta}^*)$. This ensures that $\widetilde{\phi}(a)^\top \boldsymbol{\theta}^*_{\widehat{S}_\tau} = \phi(a)^\top \boldsymbol{\theta}^*$ for all $a \in \mathcal{A}$, which means we can construct a confidence sequence for $\boldsymbol{\theta}^*_{\widehat{S}_\tau}$ instead of $\boldsymbol{\theta}^*$. We analyse the conditions under which this requirement can be guaranteed in Section 5.6.1.

For the filtration $(\mathcal{D}_t | t \in \mathbb{N})$, we set $\mathcal{D}_t$ to be the σ-algebra generated by $a'_1, r'_1, \cdots, a'_{T_1}, r'_{T_1}, a_1, r_1, \ldots, a_t, r_t, a_{t+1}$. This choice of the filtration ensures that the restricted feature map $\widetilde{\phi}$ is $\mathcal{D}_0$-measurable. We choose $Z_t(g_t) = (g_t - \widetilde{\phi}(a_t)^\top \boldsymbol{\theta}^*_{\widehat{S}_\tau})\epsilon_t$, $\lambda_t \equiv 1/\sigma^2$ and we use Gaussian mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$. Note that, for every $t \ge 1$, $\boldsymbol{\mu}_t$ and $\boldsymbol{T}_t$ can depend on all of the exploration data (and also on $\widetilde{\phi}$). Following the derivation of Equation (5.11) in Section 4.5, one can obtain a similar convex quadratic constraint for $\boldsymbol{\theta}^*_{\widehat{S}_\tau}$. Collecting the restricted feature vectors in $\widetilde{\Phi}_t := (\widetilde{\phi}(a_1), \ldots, \widetilde{\phi}(a_1))^\top$, we have

$$\left\| \widetilde{\Phi}_t \boldsymbol{\theta}^*_{\widehat{S}_\tau} - \boldsymbol{r}_t \right\|_2^2 \le (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2} \right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{r}_t) + \sigma^2 \ln \det \left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2} \right) + 2\sigma^2 \ln \frac{1}{\delta} = \widetilde{R}^2_{\mathrm{MM},t}. \tag{5.13}$$

Using the fact that $\boldsymbol{\theta}^*_{\widehat{S}_\tau}$ satisfies the same ℓ_1 norm constraint as $\boldsymbol{\theta}^*$, we can obtain a confidence sequence for $\boldsymbol{\theta}^*_{\widehat{S}_\tau}$ by combining (5.13) and $\|\boldsymbol{\theta}^*_{\widehat{S}_\tau}\|_1 \le B_1$.

Corollary 5.2 (Restricted SCMM Confidence Sequence)**.** *For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, it holds with probability at least $1 - \delta$ that for all $t \ge 1$ simultaneously $\boldsymbol{\theta}^*_{\widehat{S}_\tau}$ lies in the set*

$$\widetilde{\Theta}_t^{\ell 1} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{|\widehat{S}_\tau|} \,\middle|\, \left\| \widetilde{\Phi}_t \boldsymbol{\theta} - \boldsymbol{r}_t \right\|_2 \le \widetilde{R}_{\mathrm{MM},t} \quad \text{and} \quad \|\boldsymbol{\theta}\|_1 \le B_1 \right\}. \tag{5.14}$$

$\widetilde{\Theta}_t^{\ell 1}$ is the intersection of a $|\widehat{S}_\tau|$-dimensional ellipse and a $|\widehat{S}_\tau|$-dimensional ℓ_1 ball of radius B_1.

5.5 Algorithms for Sparse Linear Bandits

In this section, we describe our proposed sparse linear bandit algorithms. At a high level, each algorithm runs the LinUCB algorithm (described in Algorithm 1 in Chapter 4) with either the SCMM or RSCMM confidence sequence.

5.5.1 SCMM-UCB Algorithm

Our first algorithm is called Sparse Convex Martingale Mixture UCB (SCMM-UCB). We run the LinUCB algorithm with the SCMM confidence sequence from Corollary 5.1. To run the LinUCB action selection rule the SCMM confidence sequence, we need to be able to maximise $\mathrm{UCB}_{\Theta_t^{\ell 1}}(a)$, which is the solution of the convex program

$$\max_{\boldsymbol{\theta} \in \mathbb{R}^d} \ \phi(a)^\top \boldsymbol{\theta} \quad \text{s.t.} \ \left\| \Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t \right\|_2 \leq R_{\mathrm{MM},t} \quad \text{and} \quad \left\| \boldsymbol{\theta} \right\|_1 \leq B_1. \tag{5.15}$$

Since $\mathrm{UCB}_{\Theta_t^{\ell 1}}(a)$ is a maximum over a set of linear functions, our discussion on the feasibility of exact maximisation of $\mathrm{UCB}_{\Theta_t^{\ell 2}}(a)$ from Section 4.6 also applies to $\mathrm{UCB}_{\Theta_t^{\ell 1}}(a)$. If the action set has finite cardinality, $\mathrm{UCB}_{\Theta_t^{\ell 1}}(a)$ can be maximised exactly. For each $a \in \mathcal{A}$, we compute the numerical solution of (5.15) using convex solvers from the CVXPY library [52, 8], and then select the action a where the solution is greatest. If the action set is a continuous subset of $\mathbb{R}^{d_{\mathcal{A}}}$, then exact maximisation of $\mathrm{UCB}_{\Theta_t^{\ell 1}}(a)$ is infeasible. We select actions by approximately maximising $\mathrm{UCB}_{\Theta_t^{\ell 1}}(a)$ via gradient-based local search. To compute the gradient of $\mathrm{UCB}_{\Theta_t^{\ell 1}}(a)$ with respect to a, we use differentiable convex optimisation methods from the cvxpylayers library [7].

In Chapter 4, we proposed the AMM-UCB algorithm, which used an analytic upper bound on $\mathrm{UCB}_{\Theta_t^{\ell 2}}(a)$. The analytic upper bound was derived by partial optimisation of the Lagrangian dual function associated with the convex program in (4.10). In Appendix C.1, we show that one can use the same approach to obtain an analytic upper bound on the solution of (5.15). However, it is difficult to find satisfactory values of the Lagrange multipliers in this analytic upper bound, so we do not pursue this approach any further.

5.5.2 Restricted SCMM-UCB Algorithm

Our second algorithm begins with an initial feature selection phase, in which a restricted feature map $\widetilde{\phi}(a) = \phi(a)_{\widehat{S}}$ is learned, where $\widehat{S}$ is an estimate of the support of $\boldsymbol{\theta}^*$. We then run the SCMM-UCB algorithm using the restricted feature map.

This algorithm is similar to the Restricted Phase Elimination (RPE) algorithm [76] and the RPE with Warm-PopArt algorithm [85], which both perform an initial feature selection phase and then run the Phased Elimination algorithm [105] (or OFUL [3] if the action set is a continuous subset of $\mathbb{R}^{d_{\mathcal{A}}}$) using the restricted feature map. While these algorithms use the Lasso or PopArt [85] estimators to select features, we use the thresholded Lasso estimate. We call our second algorithm Restricted SCMM-UCB.

The feature selection phase is described in Algorithm 2. We sample T_1 actions (and rewards) from an exploration distribution ρ and then use the collected data $\{(a'_t, r'_t)\}_{t=1}^{T_1}$ to compute the support $\widehat{S}_\tau$ of the thresholded Lasso estimate (see Equation (5.7)) with regularisation parameter η and

threshold level τ. In Section 5.6.1, we show that for suitable choices of ρ, T_1, η and τ, $\widehat{S}_\tau$ contains the support of $\boldsymbol{\theta}^*$ and a total of $\mathcal{O}(s)$ features with high probability. This means that the original d-dimensional linear bandit problem is reduced to an $\mathcal{O}(s)$-dimensional linear bandit problem.

Algorithm 2: Thresholded Lasso feature selection

Input: exploration distribution ρ, exploration length T_1, regularisation parameter η, threshold level τ

for $t = 1, 2, \ldots, T_1$ **do**

  Play an action $a'_t \sim \rho$

  Observe a reward $r'_t = \phi(a'_t)^\top \boldsymbol{\theta}^* + \epsilon_t$

end

Compute the Lasso estimate $\widehat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \frac{1}{2T_1} \left\| \Phi_{T_1} \boldsymbol{\theta} - \boldsymbol{r}_{T_1} \right\|_2^2 + \eta \left\| \boldsymbol{\theta} \right\|_1 \right\}$

Return: $\widehat{S}_\tau = \left\{ i \in \{1, \ldots, d\} | |\widehat{\theta}_i| \geq \tau \right\}$

Choosing the Exploration Distribution. For any exploration distribution ρ (on $\mathcal{A}$), let $\boldsymbol{\Sigma}_\rho := \mathbb{E}_{a \sim \rho}[\phi(a)\phi(a)^\top]$ denote the population covariance matrix under the distribution ρ and let $\widehat{\boldsymbol{\Sigma}}_{\rho,T_1} := \frac{1}{T_1} \sum_{t=1}^{T_1} \phi(a'_t)\phi(a'_t)^\top$ denote the empirical covariance matrix. To guarantee successful feature selection (see Section 5.6.1), we require that $\widehat{\boldsymbol{\Sigma}}_{\rho,T_1}$ satisfies the compatibility condition $\widehat{\boldsymbol{\Sigma}}_{\rho,T_1} \in C(S^*, \kappa)$ (see Equation 5.8), where $S^* = \operatorname{supp}(\boldsymbol{\theta}^*)$ and $\kappa > 0$. Using Lemma EC.6 of [26], this can be guaranteed with high probability if $\boldsymbol{\Sigma}_\rho \in C(S^*, \sqrt{2}\kappa)$ and T_1 is sufficiently large. Since $\boldsymbol{\Sigma}_\rho$ is symmetric, its eigenvalues are all real. If we choose ρ such that $\nu_{\min}(\boldsymbol{\Sigma}_\rho) > 0$ (which is possible if and only if the feature vectors $(\phi(a))_{a \in \mathcal{A}}$ span $\mathbb{R}^d$ (see Remark 3.2. of [76])), then for any set of indices $S \subset [d]$, we have $\boldsymbol{\Sigma}_\rho \in C(S, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)})$. This means that, for sufficiently large T_1, $\widehat{\boldsymbol{\Sigma}}_{\rho,T_1} \in C(S^*, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$. The required sample size T_1 decreases as $\nu_{\min}(\boldsymbol{\Sigma}_\rho)$ increases (see Section 5.6.1), so the best choice of ρ is

$$\rho^* := \operatorname*{argmax}_\rho \left\{ \nu_{\min}(\boldsymbol{\Sigma}_\rho) \right\}. \tag{5.16}$$

When $\mathcal{A}$ is finite, any distribution $\rho \in \mathcal{P}(\mathcal{A})$ can be expressed as a probability vector $\boldsymbol{p}$. As discussed in Remark 4.1. of [76], the minimum eigenvalue of $\boldsymbol{\Sigma}_\rho = \sum_{a \in \mathcal{A}} p_a \phi(a)\phi(a)^\top$ is concave in $\boldsymbol{p}$ (see Example 4.43 of [32]), which means maximising the minimum eigenvalue of $\boldsymbol{\Sigma}_\rho$ with respect to $\boldsymbol{p}$ is a convex optimisation problem. In fact, the maximum minimum eigenvalue ν and the optimal distribution $\boldsymbol{p}$ can be found by solving the semidefinite program

$$\max_{\nu \in \mathbb{R}, \boldsymbol{p} \in \mathbb{R}^{|\mathcal{A}|}} \quad \nu \quad \text{such that} \quad \sum_{a \in \mathcal{A}} p_a \phi(a)\phi(a)^\top - \nu \boldsymbol{I} \succ 0, \quad \sum_{a \in \mathcal{A}} p_a = 1, \quad p_a \geq 0. \tag{5.17}$$

The matrix inequality means that $\sum_{a \in \mathcal{A}} p_a \phi(a)\phi(a)^\top - \nu \boldsymbol{I}$ must be positive definite. This means that, for finite $\mathcal{A}$, ρ^* can be computed efficiently using (for example) the CVXPY library [52, 8]. If $\mathcal{A}$ is not finite, one can approximate ρ^* by generating a large (finite) set of random actions $\mathcal{A}_{\text{rand}} \subset \mathcal{A}$, and then solving (5.17) for $\boldsymbol{p} \in \mathbb{R}^{|\mathcal{A}_{\text{rand}}|}$. Alternatively, one can choose ρ to be a simple (e.g. uniform) distribution on $\mathcal{A}$ that is easy to sample from.

5.5.3 Choosing the Mixture Distributions

The SCMM-UCB and RSCMM-UCB algorithms both require the user to select the mixture distributions $(P_t | t \in \mathbb{N})$. In the remainder of this section, we describe some choices for the mixture distributions in the SCMM-UCB and RSCMM-UCB algorithms.

Standard Mixture Distributions for SCMM-UCB. For the SCMM-UCB algorithm, we use Gaussian mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, which allows us to use the analytic expression for the radius $R_{\mathrm{MM},t}$ in Equation (5.11). One can use the Gaussian process-like mixture distributions described in Section 4.6.4 of Chapter 4. We choose a mean function $m : \mathcal{A} \to \mathbb{R}$ and a positive-definite kernel function $k : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ and then set $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, where

$$\boldsymbol{\mu}_t = [m(a_1), m(a_2), \ldots, m(a_t)]^\top, \qquad \boldsymbol{T}_t = \begin{bmatrix} k(a_1, a_1) & k(a_1, a_2) & \cdots & k(a_1, a_t) \\ k(a_2, a_1) & k(a_2, a_2) & \cdots & k(a_2, a_t) \\ \vdots & \vdots & \ddots & \vdots \\ k(a_t, a_1) & k(a_t, a_2) & \cdots & k(a_t, a_t) \end{bmatrix}. \tag{5.18}$$

As discussed in Section 4.6.4, it is natural to choose a linear mean function $m(a) = \phi(a)^\top \boldsymbol{\theta}_0$ and a linear kernel $k(a, a') = \phi(a)^\top \boldsymbol{\Sigma}_0 \phi(a)$, where $\boldsymbol{\Sigma}_0$ is symmetric and positive-definite. If we set $m(a) = 0$ and $k(a, a') = \phi(a)^\top \phi(a)$, we recover the standard mixture distributions $P_t = \mathcal{N}(\mathbf{0}, \Phi_t \Phi_t^\top)$.

Sparsity Inducing Mixture Distributions for SCMM-UCB. Despite their convenience, Gaussian mixture distributions may not be the most suitable choice for sparse linear bandits because they do not take the sparsity of $\boldsymbol{\theta}^*$ into account. As an alternative, one could start with a sparsity inducing prior for $\boldsymbol{\theta}^*$ and then set each P_t to be the induced distribution over the function values $\Phi_t \boldsymbol{\theta}^*$. Options for the sparsity inducing prior over $\boldsymbol{\theta}^*$ include the spike and slab [125, 65, 84], and the priors used in the relevance vector machine [173] and in sparse PAC-Bayesian regression [47, 13, 12, 71].

Such a mixture distribution has attractive properties: (a) it only assigns non-zero probability to vectors of function values that could be generated by linear functions; (b) it assigns high probability to function values that could be generated by sparse linear functions. The main issue with using these priors is that one can no longer use the analytic form for the radius $R_{\mathrm{MM},t}$ in Equation (5.11). We leave martingale mixture confidence sequences with sparsity inducing mixture distributions as a challenge to address in future research.

Standard Mixture Distributions for RSCMM-UCB. For the RSCMM-UCB algorithm, we use Gaussian process-like mixture distributions as in (5.18), which allows us to use the analytic expression for the radius $\widetilde{R}_{\mathrm{MM},t}$ in Equation (5.13). We could recover standard mixture distributions for RSCMM-UCB by choosing

$$m(a) = 0, \qquad k(a, a') = \widetilde{\phi}(a)^\top \widetilde{\phi}(a'), \tag{5.19}$$

where $\widetilde{\phi}$ is the restricted feature map. However, our choice of m and k can depend on all the exploration data $\{(a'_t, r'_t)\}_{t=1}^{T_1}$. We therefore propose to set the mixture distribution to the Bayesian

Gaussian process posterior mean and kernel/covariance, given the exploration data and the Gaussian process prior specified by (5.19), which is

$$m(a) = \widetilde{\phi}(a)^\top (\widetilde{\Phi}_{T_1}^\top \widetilde{\Phi}_{T_1} + \sigma^2 \boldsymbol{I})^{-1} \widetilde{\Phi}_{T_1}^\top \boldsymbol{r}'_{T_1},$$
$$k(a, a') = \widetilde{\phi}(a)^\top \widetilde{\phi}(a') - \widetilde{\phi}(a)^\top (\widetilde{\Phi}_{T_1}^\top \widetilde{\Phi}_{T_1} + \sigma^2 \boldsymbol{I})^{-1} \widetilde{\Phi}_{T_1}^\top \widetilde{\Phi}_{T_1} \widetilde{\phi}(a').$$

The corresponding Gaussian process-like mixture distribution is $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, where

$$\boldsymbol{\mu}_t = \widetilde{\Phi}_t (\widetilde{\Phi}_{T_1}^\top \widetilde{\Phi}_{T_1} + \sigma^2 \boldsymbol{I})^{-1} \widetilde{\Phi}_{T_1}^\top \boldsymbol{r}'_{T_1}, \quad \boldsymbol{T}_t = \widetilde{\Phi}_t (\boldsymbol{I} - (\widetilde{\Phi}_{T_1}^\top \widetilde{\Phi}_{T_1} + \sigma^2 \boldsymbol{I})^{-1} \widetilde{\Phi}_{T_1}^\top \widetilde{\Phi}_{T_1}) \widetilde{\Phi}_t^\top. \tag{5.20}$$

5.6 Theoretical Analysis

In this section, we analyse the behaviour of the SCMM-UCB and RSCMM-UCB algorithms. In Section 5.6.1 we provide feature selection guarantees for the Thresholded Lasso Feature Selection procedure described in Algorithm 2. We show that, for sufficiently large T and appropriate values of η and τ, Algorithm 2 is guaranteed to return the exact support of $\boldsymbol{\theta}^*$.

In Section 5.6.2, we provide data-dependent upper bounds on the simple regret for SCMM-UCB and RSCMM-UCB. After a number of actions have been selected, these bounds tell us the gap between the reward of the best selected action and the optimal action. One could our simple regret bounds to design design stopping criteria, as in e.g. [116, 83].

Finally, in Section 5.6.3, we provide data-independent (cumulative and simple) regret bounds for SCMM-UCB and RSCMM-UCB, which have explicit dependence on the number of round T, the feature vector dimension d and the sparsity level s. We begin by stating the assumptions under which our analysis holds.

Assumption 5.3 (Sub-Gaussian noise)**.** Let $\mathcal{D}_k$ be the σ-algebra generated by $(a_1, r_1, \ldots, a_k, r_k, a_{k+1})$. Each noise variable ϵ_k is conditionally zero-mean and σ-sub-Gaussian, which means

$$\mathbb{E}\left[\epsilon_k | \mathcal{D}_{k-1}\right] = 0, \quad \text{and} \quad \forall \lambda \in \mathbb{R}, \ \ \mathbb{E}\left[\exp(\lambda \epsilon_k) | \mathcal{D}_{k-1}\right] \leq \exp(\lambda^2 \sigma^2 / 2).$$

Assumption 5.4 (Bounded parameter vector)**.** For some $B_1 > 0$, $\|\boldsymbol{\theta}^*\|_1 \leq B_1$.

Assumption 5.5 (Hard sparsity)**.** For some $s \geq 1$, $\|\boldsymbol{\theta}^*\|_0 = s$.

Assumption 5.6 (Minimum signal)**.** For some constant $m > 0$, $|\theta_i^*| \geq m$ for all $i \in J$.

Assumption 5.7 (Bounded feature vectors)**.** For some constants $L_2 > 0$ and $L_\infty > 0$, $\|\phi(a)\|_2 \leq L_2$ and $\|\phi(a)\|_\infty \leq L_\infty$.

Assumption 5.8 (Bounded expected reward)**.** For some $C > 0$, $\phi(a)^\top \boldsymbol{\theta}^* \in [-C, C]$ for all $a \in \mathcal{A}$.

Assumption 5.9 (Feature vectors span $\mathbb{R}^d$)**.** $\mathcal{A}$ and $\phi : \mathcal{A} \to \mathbb{R}^d$ are such that the set of feature vectors $\{\phi(a) | a \in \mathcal{A}\}$ spans $\mathbb{R}^d$.

Assumptions 5.3, 5.7 and 5.8 are standard in the linear bandit literature. In analyses of linear bandit algorithms, it is commonly assumed that the norm of $\boldsymbol{\theta}^*$ is bounded (although more often the ℓ_2 norm than the ℓ_1 norm). The minimum signal assumption is fairly common when one wishes

to derive feature selection guarantees, although one could make the case that the combination of the hard sparsity and minimum signal assumptions is too restrictive to be an accurate way to model low-dimensional structure in real-world bandit problems. Note that Assumption 5.4 and Assumption 5.7 together imply that Assumption 5.8 must hold with $C \leq B_1 L_\infty$. We state Assumption 5.8 as a separate assumption because this leaves open the possibility that a better (than $B_1 L_\infty$) value of C is known.

5.6.1 Feature Selection Guarantees

Our first result is a feature selection guarantee for the Thresholded Lasso Feature Selection method in Algorithm 2. We prove Theorem 5.10 in Appendix C.2.

Theorem 5.10 (Feature Selection Guarantee)**.** *Suppose that assumptions 5.3, 5.5, 5.6, 5.7 and 5.9 all hold. Choose any $\delta \in (0, 1]$ and set threshold level equal to $\tau = m/2$. Choose any exploration distribution ρ, such that $\nu_{\min}(\boldsymbol{\Sigma}_\rho) > 0$. Set $\eta = 2\sigma L_\infty \sqrt{2\ln(4d/\delta)/T_1}$. Choose any T_1 such that*

$$T_1 \geq \max\left(\frac{3}{\xi^2}\ln(d), \frac{1}{\xi^2}\ln(1/\delta), \frac{2048 s^2 \sigma^2 L_\infty^2 \ln(4d/\delta)}{m^2 \nu_{\min}(\boldsymbol{\Sigma}_\rho)^2}\right), \quad \textit{where} \quad \xi = \min\left(\frac{1}{2}, \frac{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{256 s L_\infty^2}\right).$$

With probability at least $1 - \delta$, we have

$$\mathrm{supp}(\boldsymbol{\theta}^*) = \mathrm{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}).$$

Recall that $\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}$ is the thresholded Lasso estimate from Equation (5.7). Theorem 5.10 states that if we choose a large enough exploration length T_1 and suitable values for the regularisation parameter η and threshold level τ, then Algorithm 2 is guaranteed to return the exact support of $\boldsymbol{\theta}^*$. If $\nu_{\min}(\boldsymbol{\Sigma}_\rho)$ does not depend on the feature vector dimension d, then the minimum exploration length required is $\mathcal{O}(\ln(d))$. This suggests that when d is large, it may be possible to select from d features using fewer than d samples.

A similar result, which holds under the same assumptions, was recently proven for the standard Lasso estimate in (the proof of) Theorem 5.2 of [76]. However, the result for the standard Lasso does not guarantee exact support recovery. It only guarantees that $\mathrm{supp}(\boldsymbol{\theta}^*) \subseteq \mathrm{supp}(\widehat{\boldsymbol{\theta}})$ and $|\mathrm{supp}(\widehat{\boldsymbol{\theta}})| \leq \mathcal{O}(s\nu_{\max}(\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1})/\nu_{\min}(\boldsymbol{\Sigma}_\rho))$, which means that the Lasso will select all relevant features, but may also select many unnecessary features. Theorem 5.10 shows that a stronger guarantee is possible for the thresholded Lasso. In Theorem C.1 in Appendix C.2, we also show that a weaker feature selection guarantee (similar to one in [76]) can be obtained with other threshold levels, including $\tau = 0$ (i.e., the standard Lasso).

The result of Theorem 5.10 has some significant limitations, although these limitations are shared by other feature selection guarantees for sparse linear bandits [76, 85]. Though the minimum value of T_1 has acceptable growth rates in s and d, it is typically very large, due to several large multiplicative constants. Moreover, the minimum value of T_1 depends on the sparsity level s and the minimum signal level m, which are often unknown. Due to these limitations, one aim of our experiments is to determine the performance of Algorithm 2 when T_1 is set to much smaller values than the minimum value in Theorem 5.10.

5.6.2 Data-Dependent Regret Bounds

We provide data-dependent simple regret bounds for SCMM-UCB and RSCMM-UCB.

Theorem 5.11 (Data-Dependent Simple Regret Bound for SCMM-UCB)**.** *Suppose that assumptions 5.3-5.4 hold. For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, any $\delta \in (0, 1]$ and all $T \geq 1$, with probability at least $1 - \delta$, the simple regret of SCMM-UCB is bounded by*

$$\Delta_T \leq \max_{a \in \mathcal{A}} \left\{ \mathrm{UCB}_{\Theta_T^{\ell_1}}(a) \right\} - \max_{t \in [T]} \left\{ \mathrm{LCB}_{\Theta_T^{\ell_1}}(a_t) \right\}.$$

Since the maximum of the SCMM upper confidence bound can be computed or approximated efficiently, Theorem 5.11 provides an attractive way to certify the quality of the best action selected by SCMM-UCB. Since this simple regret bound holds uniformly over $T \geq 1$ it is highly suitable for use in stopping criteria. For example, as in [116], one could stop the SCMM-UCB algorithm as soon as the simple regret bound drops below some tolerance level. The proof of this simple regret bound is simple. We only need to use basic properties of upper and lower confidence bounds.

Proof.

$$\begin{aligned}
\Delta_T &= \phi(a^*)^\top \boldsymbol{\theta}^* - \max_{t \in [T]} \left\{ \phi(a_t)^\top \boldsymbol{\theta}^* \right\} \\
&\leq \mathrm{UCB}_{\Theta_T^{\ell_1}}(a^*) - \max_{t \in [T]} \left\{ \mathrm{LCB}_{\Theta_T^{\ell_1}}(a_t) \right\} \\
&\leq \max_{a \in \mathcal{A}} \left\{ \mathrm{UCB}_{\Theta_T^{\ell_1}}(a) \right\} - \max_{t \in [T]} \left\{ \mathrm{LCB}_{\Theta_T^{\ell_1}}(a_t) \right\}.
\end{aligned}$$

□

One can show that the RSCMM-UCB algorithm has a similar data-dependent simple regret bound.

Theorem 5.12 (Data-Dependent Simple Regret Bound for RSCMM-UCB)**.** *Suppose that assumptions 5.3, 5.5, 5.6, 5.7 and 5.9 all hold. Choose any $\delta \in (0, 1/2]$ and set threshold level equal to $\tau = m/2$. Choose any exploration distribution ρ, such that $\nu_{\min}(\boldsymbol{\Sigma}_\rho) > 0$. Set $\eta = 2\sigma L_\infty \sqrt{2\ln(4d/\delta)/T_1}$. Choose any T_1 such that*

$$T_1 \geq \max\left(\frac{3}{\xi^2} \ln(d), \frac{1}{\xi^2} \ln(1/\delta), \frac{2048 s^2 \sigma^2 L_\infty^2 \ln(4d/\delta)}{m^2 \nu_{\min}(\boldsymbol{\Sigma}_\rho)^2} \right), \quad \textit{where} \quad \xi = \min\left(\frac{1}{2}, \frac{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{256 s L_\infty^2} \right).$$

For any sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, with probability at least $1 - 2\delta$, the simple regret of RSCMM-UCB is bounded by

$$\Delta_T \leq \max_{a \in \mathcal{A}} \left\{ \mathrm{UCB}_{\widetilde{\Theta}_T^{\ell_1}}(a) \right\} - \max\left(\max_{t \in \{1, \ldots, T_1\}} \left\{ \mathrm{LCB}_{\widetilde{\Theta}_T^{\ell_1}}(a'_t) \right\}, \max_{t \in \{1, \ldots, T - T_1\}} \left\{ \mathrm{LCB}_{\widetilde{\Theta}_T^{\ell_1}}(a_t) \right\} \right).$$

The main difference here is that the simple regret bound for RSCMM-UCB is only valid when the restricted feature map $\widetilde{\phi}$ contains all the relevant features. To guarantee that this is the case, we use the feature selection guarantee in Theorem 5.10, which introduces the requirement that the

exploration length T_1 is sufficiently large. Apart from this detail, the proof of Theorem 5.12 is the same as the proof of Theorem 5.11, so we omit it.

Since the simple regret guarantee for RSCMM-UCB only holds uniformly for $T \geq T_1$, it is much less useful than the simple regret guarantee for SCMM-UCB in Theorem 5.11. Note that the requirement of a feature selection guarantee also means that Theorem 5.12 requires more assumptions than Theorem 5.11.

5.6.3 Data-Independent Regret Bounds

We now state data-independent cumulative regret bounds for our SCMM-UCB and RSCMM-UCB algorithms, which have explicit dependence on the dimension d and the sparsity level s. Unfortunately, we are not yet able to prove a data-independent regret bound for SCMM-UCB that improves upon the growth rate (in d) of the equivalent regret bound for CMM-UCB in Chapter 4. The best we can do is to show that SCMM-UCB performs no worse than CMM-UCB.

Theorem 5.13 (Data-Independent Cumulative Regret Bound for SCMM-UCB)**.** *Suppose that assumptions 5.3, 5.4, 5.7 and 5.8 hold. If for any $c > 0$, the sequence of mixture distributions is $P_t = \mathcal{N}(\mathbf{0}, c\Phi_t\Phi_t^\top)$, then for all $T \geq 1$, with probability at least $1 - \delta$, the cumulative regret of SCMM-UCB is bounded by*

$$\Delta_{1:T} \leq \frac{2}{\sqrt{\ln 2}} \max\left\{C, \sigma\sqrt{d\ln\left(1+\frac{cL_2^2T}{\sigma^2 d}\right)+\frac{B_1^2}{c}+2\ln\frac{1}{\delta}}\right\}\sqrt{dT\ln\left(1+\frac{cL_2^2T}{\sigma^2 d}\right)}.$$

This cumulative regret bound for SCMM-UCB is $\mathcal{O}(d\sqrt{T}\ln(T))$, which is the same as the regret bound for CMM-UCB in Theorem 4.9. Using Equation (5.5), this cumulative regret bound implies that the simple regret of SCMM-UCB is no worse than $\mathcal{O}(d\ln(T)/\sqrt{T})$.

We use the cumulative regret bound for CMM-UCB in Theorem 4.9 to prove the regret bound in Theorem 5.13. Since an ℓ_1 ball of radius B_1 is a subset of an ℓ_2 ball of radius B_1, the SCMM confidence set $\Theta_t^{\ell_1}$ is a subset of the CMM confidence set $\Theta_t^{\ell_2}$ from Corollary 4.2 (if we set the ℓ_2 norm upper bound B_2 in $\Theta_t^{\ell_2}$ to be equal to B_1). This means that

$$\mathrm{UCB}_{\Theta_t^{\ell_1}}(a) \leq \mathrm{UCB}_{\Theta_t^{\ell_2}}(a) \leq \mathrm{AUCB}_{\Theta_t^{\ell_2}}(a),$$

where $\mathrm{AUCB}_{\Theta_t^{\ell_2}}(a_t)$ is analytic UCB from Equation (4.11). A similar statement holds for the LCBs. As a result, we are able to obtain the same bound on the regret at each round for SCMM-UCB, i.e.

$$\Delta(a_t) \leq \mathrm{AUCB}_{\Theta_t^{\ell_2}}(a_t) - \mathrm{ALCB}_{\Theta_t^{\ell_2}}(a_t).$$

From here, we can follow the proof of Theorem 4.9 in Appendix B.4.2. Using the feature selection guarantee from Theorem 5.10, we can prove that, if we run the feature selection phase of RSCMM-UCB for long enough, the cumulative regret of the RSCMM-UCB algorithm grows at most logarithmically with d.

Theorem 5.14 (Data-Independent Cumulative Regret Bound for RSCMM-UCB)**.** *Suppose that assumptions 5.3-5.9 all hold. Choose any $\delta \in (0, 1/2]$ and and set threshold level equal to $\tau = m/2$. Choose any exploration distribution ρ, such that $\nu_{\min}(\mathbf{\Sigma}_\rho) > 0$. Set $\eta = 2\sigma L_\infty \sqrt{2\ln(4d/\delta)/T_1}$. Choose any T_1 such that*

$$T_1 \geq \max\left(\frac{3}{\xi^2}\ln(d), \frac{1}{\xi^2}\ln(1/\delta), \frac{2048 s^2 \sigma^2 L_\infty^2 \ln(4d/\delta)}{m^2 \nu_{\min}(\mathbf{\Sigma}_\rho)^2}\right), \quad \textit{where} \quad \xi = \min\left(\frac{1}{2}, \frac{\nu_{\min}(\mathbf{\Sigma}_\rho)}{256 s L_\infty^2}\right).$$

If for any $c > 0$, the sequence of mixture distributions is $P_t = \mathcal{N}(\mathbf{0}, c\widetilde{\Phi}_t \widetilde{\Phi}_t^\top)$, then for all $T \geq T_1$, with probability at least $1 - 2\delta$, the cumulative regret of RSCMM-UCB is bounded by

$$\Delta_{1:T} \leq 2CT_1 + \frac{2}{\sqrt{\ln 2}} \max\left\{C, \sigma\sqrt{s\ln\left(1+\frac{cL_2^2(T-T_1)}{\sigma^2 s}\right) + \frac{B_1^2}{c} + 2\ln\frac{1}{\delta}}\right\} \sqrt{s(T-T_1)\ln\left(1+\frac{cL_2^2(T-T_1)}{\sigma^2 s}\right)}$$

Theorem 5.14 states that if the time horizon T is greater than T_1, then the RSCMM-UCB algorithm has cumulative regret bounded by

$$\mathcal{O}\left(\frac{s^2\ln(d)}{m^2\nu_{\min}(\mathbf{\Sigma}_\rho)^2} + s\sqrt{T}\ln(T)\right).$$

If $\nu_{\min}(\mathbf{\Sigma}_\rho)$ is independent of the dimension d, then the cumulative regret bound for RSCMM-UCB is $\mathcal{O}(s^2\ln(d) + s\sqrt{T}\ln(T))$. This implies that the simple regret of RSCMM-UCB (for $T \geq T_1$) is no worse than $\mathcal{O}(s^2\ln(d)/T + s\ln(T)/\sqrt{T})$. In the high-dimensional regime, where d and T both tend to infinity and their ratio d/T is a fixed positive constant, the $\sqrt{T}$ terms dominate the regret bounds and we obtain sublinear $\mathcal{O}(s\sqrt{T}\ln(T))$ cumulative regret and decaying $\mathcal{O}(s\ln(T)/\sqrt{T})$ simple regret. This suggests that RSCMM-UCB ought to work well in sparse high-dimensional linear bandit problems.

Theorem 5.13 follows from the combination of the feature selection guarantee in 5.10 and the regret bound for SCMM-UCB in Theorem 5.13. Due to Assumption 5.8, the total regret suffered in the feature selection phase is no more than $2CT_1$. Theorem 5.10 guarantees that the restricted feature map contains all the relevant features and exactly s features in total, which means the original d-dimensional linear bandit problem is reduced to an s-dimensional linear bandit problem. Finally, Theorem 5.13 upper bounds the total regret suffered by running the SCMM-UCB algorithm in the reduced s-dimensional linear bandit problem for the remaining $T - T_1$ rounds.

5.7 Experiments

We evaluate the empirical behaviour of the Thresholded Lasso Feature Selection method in Algorithm 2, our SCMM and RSCMM confidence bounds, and our SCMM-UCB and RSCMM-UCB algorithms for sparse linear bandits.

5.7.1 Benchmark Problem

Sparse Linear Reward Function. Our benchmark uses randomly generated linear reward functions of the form $f^*(\boldsymbol{a}) = \phi(\boldsymbol{a})^\top \boldsymbol{\theta}^*$, with actions $\boldsymbol{a} \in \mathbb{R}^{d_\mathcal{A}}$ and $\boldsymbol{\theta}^* \in \mathbb{R}^d$. The rewards are corrupted by independent Gaussian noise with variance σ^2, so $r_t = \phi(\boldsymbol{a}_t)^\top \boldsymbol{\theta}^* + \epsilon_t$ and $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. This ensures that the noise variables $\epsilon_1, \epsilon_2, \ldots$ are all σ-sub-Gaussian. We always use $\sigma = 0.1$. For the feature map $\phi : \mathbb{R}^{d_\mathcal{A}} \to \mathbb{R}^d$, we use Random Fourier Features (cf. Algorithm 1 of [140]). With this choice of ϕ, we have $L_\infty \leq \sqrt{2/d}$. The parameter vector $\boldsymbol{\theta}^*$ is generated in two stages: (a) sample an s-dimensional random vector with elements drawn independently from a Rademacher distribution (± 1 with equal probability); (b) append $d - s$ zeros to obtain an s-sparse d-dimensional vector that satisfies the minimum signal condition with $m = 1$.

5.7.2 Feature Selection

We investigate the behaviour of the feature selection procedure in Algorithm 2 in the Sparse Linear Reward Function, where we can guarantee that assumptions 5.3-5.9 hold. We attempt to verify that feature selection with the thresholded Lasso estimate can still work well when T_1 is smaller than the theoretically motivated minimum value.

Compared Methods. We compare the feature selection procedure in Algorithm 2 where the estimate used is: (a) the thresholded Lasso estimate (as suggested in this chapter); (b) the Lasso estimate (as in [76]); (c) the PopArt estimate (as in [85]).

Experimental Setup. We conduct experiments in the Sparse Linear Reward Function benchmark. We always set $d_\mathcal{A} = 50$ and we vary d between 64 and 256. We use data sets $\{(\boldsymbol{a}_t, r_t)\}_{t=1}^T$ of size T, where T is between 1 and 500. The actions $\{\boldsymbol{a}_t\}_{t=1}^T$ are drawn independently from a uniform distribution over the hypercube $[0, 1]^{d_\mathcal{A}}$, i.e. $\boldsymbol{a}_t \sim \mathcal{U}([0, 1]^{d_\mathcal{A}})$. We use $s = 5$, which means that most elements of $\boldsymbol{\theta}^*$ are 0. For the Lasso and thresholded Lasso, we use the theoretically motivated value $\eta = 2\sigma L_\infty \sqrt{2 \ln(4d/\delta)/T_1}$. We vary the threshold level τ between 0.01 and 0.5. In this setting, the minimum signal level is $m = 1$, so the value of τ used in Theorem 5.10 is 0.5. For the PopArt estimate, we use the default values of all parameters.

Results. For each feature selection method, and for each value of T and d, we record the proportions of the indices in $\text{supp}(\boldsymbol{\theta}^*)$ and $\text{supp}(\boldsymbol{\theta}^*)^\mathsf{c}$ which are selected. Ideally, all the indices in $\text{supp}(\boldsymbol{\theta}^*)$ are selected (all relevant features are selected) and none of the indices in $\text{supp}(\boldsymbol{\theta}^*)^\mathsf{c}$ (no irrelevant features are selected).

In Figure 5.2, we observe that when the threshold level τ is 0.3 or 0.5, the thresholded Lasso estimate successfully selects the exact support of $\boldsymbol{\theta}^*$, for all values of d and where T is approximately ≥ 50. This suggests that it is sometimes possible to achieve exact support selection with a number of samples that is much smaller than the lower bound in Theorem 5.10. Adding a threshold to the Lasso estimate had a large impact on its sparsity. When the threshold was close 0.01 or 0.03 and T was large, the thresholded Lasso estimate sometimes selected more than half of the irrelevant features. We hypothesise that the proportion of $\text{supp}(\boldsymbol{\theta}^*)^\mathsf{c}$ initially increases with T for some threshold levels because the regularisation strength η decreases as T increases. For all values of d and where T, the standard Lasso estimate had no elements that were exactly equal to 0, which meant it always selected every feature.

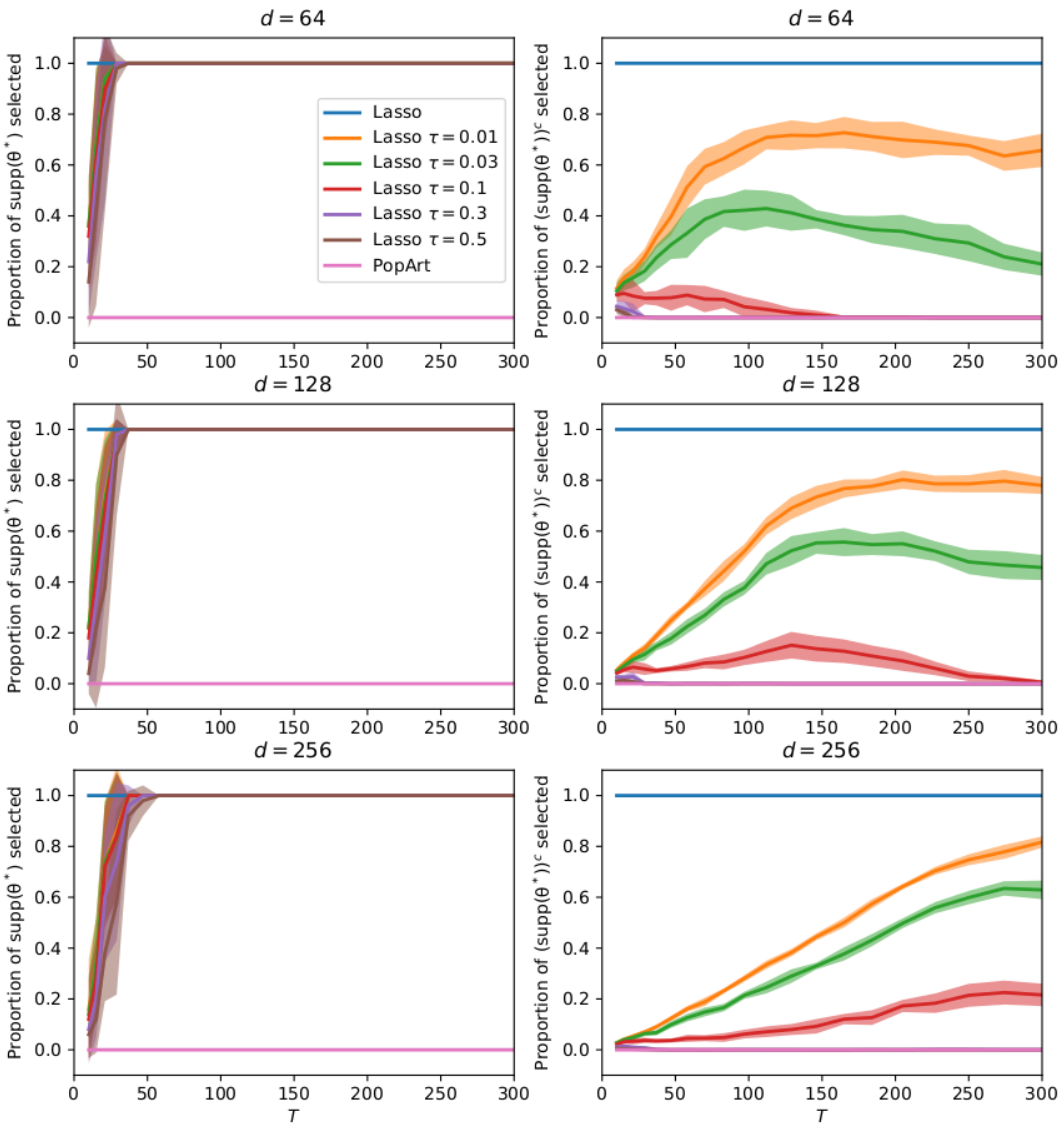


Figure 5.2: The proportion of the indices in $\text{supp}(\boldsymbol{\theta}^*)$ selected (left column) and the proportion of the indices in $\text{supp}(\boldsymbol{\theta}^*)^{\mathsf{c}}$ selected (right) for the (thresholded) Lasso estimate and the PopArt estimate. In each plot, the number of samples T varies along the x-axis. The top row shows results for $d = 64$, the middle row shows results for $d = 128$ and the bottom row shows results for $d = 256$.

For all values of d and where T shown in Figure 5.2, the PopArt estimate was the 0 vector, i.e.

it never selected any features. The PopArt estimate achieves sparsity by setting every component of the estimate that cannot be guaranteed to be non-zero (using an infinity norm estimation error bound) equal to 0. If the sample size T is large enough and the minimum signal assumption is satisfied, it can be shown that PopArt recovers the exact support of $\boldsymbol{\theta}^*$ (see Theorem 1 of [85]). However, we find that for smaller values of T, PopArt fails to perform effective feature selection.

5.7.3 Upper and Lower Confidence Bounds

We investigate the tightness and validity of the SCMM and Restricted SCMM upper and lower confidence bounds. We aim to investigate the growth rates in d of the widths of the SCMM and Restricted SCMM confidence bounds when the ground truth function is linear and sparse.

Compared Methods. We evaluate the following upper/lower confidence bounds: (a) CMM-UCB: our numerical UCBs/LCBs from Chapter 4; (b) SCMM-UCB: our UCBs/LCBs from Section 5.5.1; (c) RSCMM-UCB: our Restricted SCMM UCBs/LCBs from Section 5.5.2; (d) SCMM-UCB Oracle: our UCBs/LCBs from Section 5.5.1, except we use the s-dimensional feature map, which only contains features where $\boldsymbol{\theta}^*$ has support. CMM-UCB serves as a baseline method, which is not designed to exploit sparsity. The relative performance of our methods and SCMM-UCB oracle tells us the price of not knowing which s features to use in advance.

Experimental Setup. We conduct experiments in the Sparse Linear Reward Function benchmark. First, we produce plots of the different UCBs and LCBs (see Figure 5.1). We set $d_{\mathcal{A}} = 1$, $d = 100$ and $s = 5$. We use data sets $\{(\boldsymbol{a}_t, r_t)\}_{t=1}^T$ of size $T = 50$. For RSCMM-UCB, we set the exploration length to $T_1 = 30$. Half of the actions are drawn uniformly from the interval $[-4, -1]$ and the remaining half are drawn uniformly from the interval $[1, 4]$.

To investigate the growth rates of the widths of different confidence sets (see Figure 5.3), we set $d_{\mathcal{A}} = 50$, we vary d between 5 and 100, and we set $s = 5$. We use data sets $\{(\boldsymbol{a}_t, r_t)\}_{t=1}^T$ of size $T = 100$. The actions $\{\boldsymbol{a}_t\}_{t=1}^T$ are drawn independently from a uniform distribution over the hypercube $[0, 1]^{d_{\mathcal{A}}}$. For RSCMM-UCB, we set the exploration length to $T_1 = 50$. We record the average width (UCB minus LCB) at $T_{\text{test}} = 100$ randomly drawn test points $\{\boldsymbol{a}'_t\}_{t=1}^{T_{\text{test}}}$. Since the RSCMM-UCB confidence bounds do not necessarily contain the ground truth function with high probability (because the exploration length T_1 is smaller than the theoretically motivated minimum value), we also record the failure rate of each UCB/LCB. This is the frequency with which the function values $\phi(\boldsymbol{a}'_t)^\top \boldsymbol{\theta}^*$ at the test points $\boldsymbol{a}'_t$ lie outside the interval $[\text{LCB}(\boldsymbol{a}'_t), \text{UCB}(\boldsymbol{a}'_t)]$.

With each method, we use Gaussian mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$. For CMM-UCB and SCMM-UCB, we use the standard mixture distributions $P_t = \mathcal{N}(\boldsymbol{0}, \Phi_t \Phi_t^\top)$. For RSCMM-UCB, we the mixture distributions in Equation (5.20). SCMM-UCB Oracle use the mixture distributions $P_t = \mathcal{N}(\boldsymbol{0}, \Phi_t^* \Phi_t^{*\top})$, where Φ_t^* is the usual design matrix, except containing only the columns (features) on which $\boldsymbol{\theta}^*$ has support. In both experiments, we run RSCMM-UCB $\eta = 2\sigma L_\infty \sqrt{2\ln(4d/\delta)/T_1}$ and $\tau = 0.5$.

Results. Figure 5.3 shows the average width of the CMM-UCB, SCMM-UCB, RSCMM-UCB and SCMM-UCB Oracle confidence bounds. As we would expect, the SCMM-UCB Oracle confidence bounds appear to be have the smallest width. Their width does not grow with the feature vector

dimension. Surprisingly, the RSCMM-UCB confidence bounds appear to have almost exactly the same width as the SCMM-UCB Oracle confidence bounds and their width also did not grow when the feature vector dimension was increased from 5 to 100. The SCMM-UCB confidence bounds were wider than those of RSCMM-UCB, but much narrower than the CMM-UCB confidence bounds when d was large. In Figure 5.3 (right), we observe that the width SCMM-UCB confidence bounds grew approximately linearly in $\ln(d)$.

In Table 5.1, we see that CMM-UCB, SCMM-UCB and SCMM-UCB Oracle all achieved a failure rate of 0. However, when the feature vector dimension was $d = 100$, RSCMM-UCB had an average failure rate of 0.046, which suggests that the Thresholded Lasso feature selection sometimes fails to identify the support of $\boldsymbol{\theta}^*$ when the exploration length is too small relative to the dimension. In this setting, $T_1 = 50$ and $d = 100$.

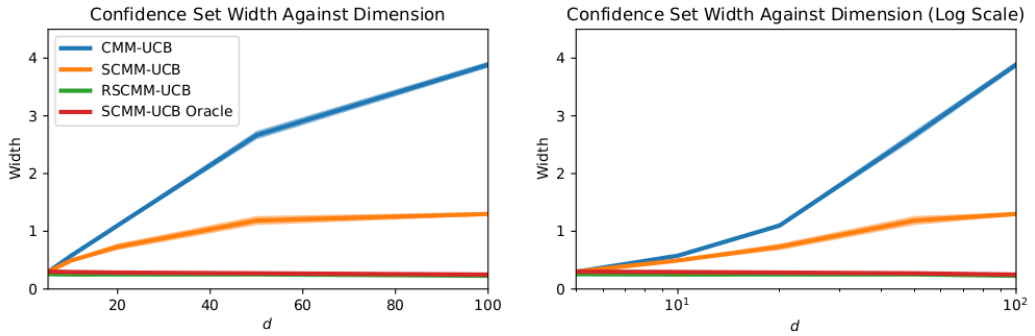


Figure 5.3: The confidence set width for different feature dimensions d. On the right, the dimension is displayed on a logarithmic scale. We show the mean and standard deviation of the widths over 10 runs.

Table 5.1: The confidence set width and the failure rate of the confidence bounds for different feature vector feature dimensions d. We show the mean and standard deviation over 10 runs.

	$d = 5$		$d = 20$		$d = 100$	
	Width	Failure Rate	Width	Failure Rate	Width	Failure Rate
CMM-UCB	0.299 ± 0.011	0.000 ± 0.000	1.101 ± 0.031	0.000 ± 0.000	3.886 ± 0.053	0.000 ± 0.000
SCMM-UCB	0.301 ± 0.010	0.000 ± 0.000	0.731 ± 0.057	0.000 ± 0.000	1.300 ± 0.033	0.000 ± 0.000
RSCMM-UCB	0.257 ± 0.030	0.000 ± 0.000	0.253 ± 0.029	0.000 ± 0.000	0.229 ± 0.023	0.046 ± 0.138
SCMM-UCB Oracle	0.301 ± 0.010	0.000 ± 0.000	0.285 ± 0.012	0.000 ± 0.000	0.250 ± 0.009	0.000 ± 0.000

5.7.4 Linear Bandits

We evaluate the SCMM-UCB and RSCMM-UCB algorithms in the Sparse Linear Reward Function benchmark. We aim to investigate the performance of our linear bandit algorithms and the tightness of their data-dependent simple regret bounds.

Compared Methods. The methods we compare are: (a) *SCMM-UCB:* our SCMM-UCB algorithm from Section 5.5.1; (b) *RSCMM-UCB:* our RSCMM-UCB algorithm from Section 5.5.2; (c) *CMM-UCB:* cf. Chapter 4; (d) *ESTC:* the Explore the Sparsity Then Commit (ESTC) algorithm

[76] (e) *ROFUL:* the Restricted Optimism in the Face of Uncertainty Linear bandit algorithm (ROFUL) [3, 76]. The original ROFUL algorithm uses the standard Lasso to select a restricted feature map. Since we found that this doesn't work well in the Sparse Linear Reward Function benchmark, we run ROFUL with our improved Thresholded Lasso feature selection algorithm to provide a tougher baseline for RSCMM-UCB.

Experimental Setup. We conduct experiments in the Sparse Linear Reward Function benchmark. For our first experiment, we set the action dimension to $d_{\mathcal{A}} = 5$, the feature vector dimension to $d = 50$, the sparsity level to $s = 5$ and the number of rounds to $T = 100$. Our second experiment uses a reward function of higher dimension. We set $d_{\mathcal{A}} = 5$, $d = 100$, $s = 5$ and $T = 200$. In both experiments, the action set is the unit hypercube, i.e. $\mathcal{A} = [0, 1]^{d_{\mathcal{A}}}$. We record the values of the data-dependent simple regret bounds at the final round, but only for the CMM-UCB, SCMM-UCB and SCMM-UCB Oracle methods. This is because the other methods only give valid simple regret guarantees when the exploration length is sufficiently large.

For each of our algorithms, we use Gaussian mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$. For CMM-UCB and SCMM-UCB, we use the standard mixture distributions $P_t = \mathcal{N}(\mathbf{0}, \Phi_t \Phi_t^\top)$. For RSCMM-UCB, we the mixture distributions in Equation (5.20). SCMM-UCB Oracle use the mixture distributions $P_t = \mathcal{N}(\mathbf{0}, \Phi_t^* {\Phi_t^*}^\top)$, where Φ_t^* is the usual design matrix, except containing only the columns (features) on which $\boldsymbol{\theta}^*$ has support. In both experiments, we use $\eta = 2\sigma L_\infty \sqrt{2\ln(4d/\delta)/T_1}$ in RSCMM-UCB. For the RSCMM-UCB, ESTC and ROFUL algorithms, which all have an initial exploration/feature selection phase, we set the exploration length to $T_1 = 40$ in the 50-dimensional problem and $T_1 = 60$ in the 100-dimensional problem.

Results. Figure 5.4 shows the average simple regret (left) and cumulative regret (right) for each method in the 50-dimensional problem. SCMM-UCB, RSCMM-UCB and ROFUL all appeared to reach almost 0 simple regret by approximately $T = 50$. The simple regret of CMM-UCB decayed to 0 at a slower rate, but reached almost 0 by $T = 100$. The simple regret of ESTC reached approximately 0.04 after the end of the exploration phase $T_1 = 40$, and then remained there.

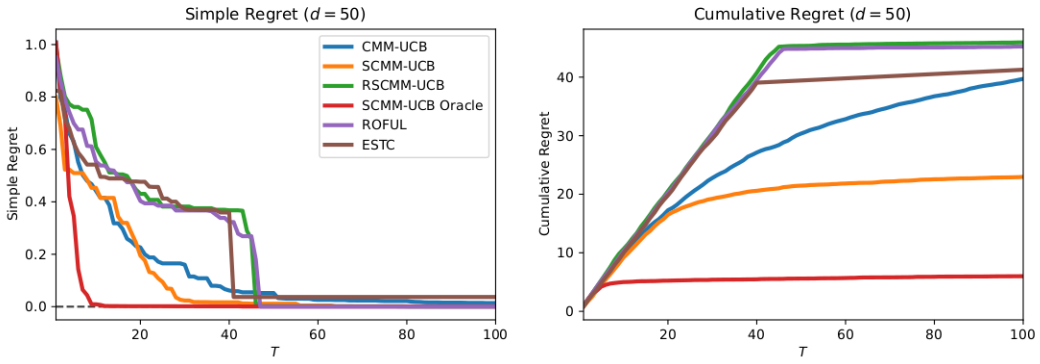


Figure 5.4: The simple regret (left) and cumulative regret (right) curves for our sparse UCB algorithms (SCMM-UCB, RSCMM-UCB) compared with CMM-UCB, ESTC, ROFUL and SCMM-UCB Oracle in the Sparse Linear Reward Function benchmark with feature vector dimension $d = 50$. We show the mean over 10 runs.

RSCMM-UCB, ROFUL and ESTC all suffered large cumulative regret during their exploration phases. As a result, even though it is not designed for sparse linear bandits, CMM-UCB had smaller cumulative regret that RSCMM-UCB, ROFUL and ESTC at $T = 100$. SCMM-UCB had by far the best cumulative regret (excluding the oracle baseline).

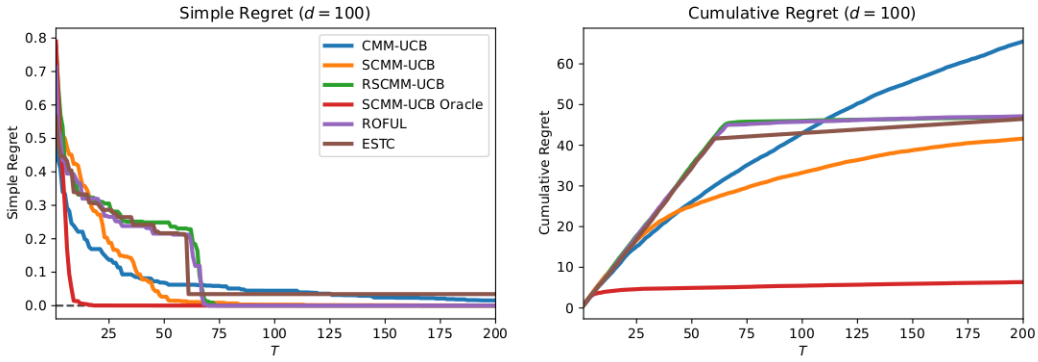


Figure 5.5: The simple regret (left) and cumulative regret (right) curves for our sparse UCB algorithms (SCMM-UCB, RSCMM-UCB) compared with CMM-UCB, ESTC, ROFUL and SCMM-UCB Oracle in the Sparse Linear Reward Function benchmark with feature vector dimension $d = 100$. We show the mean over 10 runs.

Figure 5.5 shows the average simple regret (left) and cumulative regret (right) for each method in the 50-dimensional problem. In the higher-dimensional problem, CMM-UCB had by far the highest cumulative regret at $T = 100$. SCMM-UCB still had the lowest cumulative regret (excluding the oracle baseline), although the gaps between SCMM-UCB and the three methods with exploration phases were much smaller, which suggests that SCMM-UCB may not scale up to high-dimensional problems as well as the other three methods.

Table 5.2: The cumulative regret (Cum. Regret), simple regret (Sim. Regret) and simple regret bound (Sim. Reg. Bnd.) for each algorithm in the Sparse Linear Reward Function benchmark with feature vector dimension $d = 50$ (left) and $d = 100$ (right). We show the mean $\pm$ standard deviation over 10 repetitions.

	$d = 50, T = 100$			$d = 100, T = 200$		
	Cum. Regret	Sim. Regret	Sim. Reg. Bnd.	Cum. Regret	Sim. Regret	Sim. Reg. Bnd.
CMM-UCB	39.705 $\pm$ 4.069	0.014 $\pm$ 0.011	1.024 $\pm$ 0.053	65.524 $\pm$ 7.491	0.016 $\pm$ 0.008	1.489 $\pm$ 0.067
SCMM-UCB	22.965 $\pm$ 4.410	0.000 $\pm$ 0.001	0.153 $\pm$ 0.038	41.851 $\pm$ 4.223	0.000 $\pm$ 0.001	0.189 $\pm$ 0.021
RSCMM-UCB	45.928 $\pm$ 2.777	0.000 $\pm$ 0.000	-	46.942 $\pm$ 2.510	0.000 $\pm$ 0.000	-
ESTC	41.268 $\pm$ 3.544	0.037 $\pm$ 0.028	-	46.459 $\pm$ 3.443	0.035 $\pm$ 0.024	-
ROFUL	45.225 $\pm$ 2.402	0.000 $\pm$ 0.001	-	47.121 $\pm$ 2.329	0.000 $\pm$ 0.000	-
SCMM-UCB Oracle	6.002 $\pm$ 1.150	0.000 $\pm$ 0.001	0.036 $\pm$ 0.009	6.356 $\pm$ 1.802	0.000 $\pm$ 0.000	0.027 $\pm$ 0.006

Table 5.2 shows the cumulative regret, the simple regret and the simple regret bounds for each method in the final round of each bandit problem. In these problems, the combination of Assumption 5.4 and Assumption 5.7 means that the simple regret of any action is trivially upper bounded by $B_1 L_\infty$. For the 50-dimensional and 100-dimensional problems, these trivial simple regret bounds

are 2 and $\sqrt{2} \approx 1.414$ respectively. In Table 5.2, we see that the SCMM-UCB simple regret bound is much tighter than the CMM-UCB simple regret bound, which is vacuous for the 100-dimensional problem.

5.8 Conclusion

In this chapter, we re-used our general-purpose tail bound for adaptive martingale mixtures to design new confidence sequences for sparse linear bandits. We proposed the SCMM-UCB algorithm, which selects actions by maximising upper confidence bounds constructed from martingale mixture confidence sets with ℓ_1 norm constraints. We also proposed the RSCMM-UCB algorithm, which adds an initial feature selection phase to the SCMM-UCB algorithm.

In our theoretical analysis, we proved that the feature selection phase of RSCMM-UCB, which uses the Thresholded Lasso estimate, is guaranteed to recover the exact support of a sparse parameter vector if a minimum signal condition is satisfied and the number of samples is sufficiently large. We proved data-dependent simple regret bounds for both of our algorithms as well as data-independent cumulative regret bounds, which have explicit dependence on the dimension d, the sparsity level s and the number of rounds T. For RSCMM-UCB, we proved an almost dimension-free $\mathcal{O}(s^2 \ln(d) + s\sqrt{T} \ln(T))$ cumulative regret bound.

In our experiments, we found that the feature selection phase of RSCMM-UCB is able perform successful feature selection even when the sample size is much lower than the theoretically motivated minimum value. We found that the widths of our SCMM-UCB and RSCMM-UCB confidence sets grows at much slower rate in d than the width of our CMM-UCB confidence set when the ground-truth function is sparse. Finally we found that our SCMM-UCB algorithm achieves better cumulative regret than existing sparse linear bandit algorithms and is able to provide non-vacuous simple regret guarantees, which could be used to design stopping criteria.

However, the algorithms proposed in this Chapter are not without limitations. The data-independent regret bound for SCMM-UCB does not have improved dependence on d. Moreover, while SCMM-UCB had by far the best cumulative regret in the 50 dimensional problem, the gap between the cumulative regret of SCMM-UCB and the competing sparse methods was much smaller in the 100 dimensional problem. This suggests that SCMM-UCB may have worse dependence on the feature vector dimension that the other sparse methods. We believe that this is due to our use of Gaussian mixture distributions, since from Lemma B.16, the radius $R_{\mathrm{MM},T}$ is $\mathcal{O}(\sqrt{d \ln(T)})$. In future work, we would like to investigate sparsity inducing mixture distributions (as described in Section 5.5.3).

Chapter 6

Conclusion

In this thesis, the central aim was to answer the question:

> *How can PAC-Bayesian theory be used to design bandit algorithms with performance guarantees?*

In Section 6.1, we discuss how each chapter of this thesis helps to answer this question. In Section 6.2, we describe some directions for future research.

6.1 Summary of Contributions

Chapter 3 reviewed existing PAC-Bayesian bandit algorithms and their performance guarantees. We found that previous works on PAC-Bayesian bandit algorithms focused on PAC-Bayes bounds for martingales and their application to importance sampling-based estimates of the reward or regret of a policy. A general principle for designing PAC-Bayesian bandit algorithms is to select a policy or sequence of policies that optimises one of these PAC-Bayes bounds. On the one hand, we found this approach can yield offline policy search algorithms with competitive performance and surprisingly tight performance guarantees - even for neural network-based policies. On the other hand, we found that online bandit algorithms designed in this way had sub-optimal empirical performance and sub-optimal regret bounds.

In Chapter 4, we showed that PAC-Bayes bounds can be combined with the "optimism in the face of uncertainty" principle, which reduces bandit problems to the construction of a confidence sequence for the unknown reward function. We developed a novel general-purpose PAC-Bayes-style tail bound for martingale mixtures, which we used to construct convex confidence sequences and upper confidence bounds (UCBs) for linear bandits. We showed that our PAC-Bayes-style UCBs are provably tighter than state-of-the-art UCBs constructed with the usual self-normalised bound for vector-valued martingales [4, 3], which is one of the main building blocks of UCB algorithms and has been used for over a decade without any substantial changes. We proposed the Convex Martingale Mixture UCB (CMM-UCB) algorithm, which is a PAC-Bayes-style UCB algorithm for linear bandits. We proved that CMM-UCB has a regret bound that is minimax optimal up to logarithmic factors and found that it outperformed competing linear bandit algorithms in several hyperparameter tuning tasks.

In Chapter 5, we built upon the approach taken in Chapter 4. We re-used our general-purpose PAC-Bayes-style tail bound for martingale mixtures to construct confidence sequences and UCBs for sparse linear bandits. We proposed the Sparse Convex Martingale Mixture UCB (SCMM-UCB) algorithm, which uses confidence sets with ℓ_1 norm constraints, and the Restricted SCMM-UCB (RSCMM-UCB) algorithm, which adds an initial feature selection phase to the SCMM-UCB algorithm. We demonstrated that when the reward function is sparse, the confidence bounds used by our SCMM-UCB and RSCMM-UCB algorithms can be much tighter than the confidence bounds used by CMM-UCB. We showed that SCMM-UCB in particular achieves better cumulative regret than competing sparse linear bandit algorithms.

In summary, this thesis reviewed previous works on PAC-Bayesian policy search algorithms and showed that PAC-Bayes bounds can be used to develop PAC-Bayes-style upper confidence bound algorithms, which have competitive empirical performance and come with performance guarantees.

6.2 Outlook

In Theorem 4.1, we presented a novel general-purpose tail bound for adaptive martingale mixtures. In this thesis, we specialised this result to (sparse) linear bandits. We feel that there are many interesting consequences of our general-purpose tail bound. Here, we highlight one of them.

6.2.1 PAC-Bayes-Style Confidence Sequences for Non-Linear Bandits

A limitation of our bandit algorithms from Chapter 4 and Chapter 5 is that they assume a linear expected reward function, which may not be a realistic assumption for real-world bandit problems. However, our general-purpose tail bound in Theorem 4.1 allows us to easily derive confidence sequences for non-linear reward functions by simply choosing $Z_t(g_t) = (g_t - f^*(a_t))\epsilon_t$, where f^* is the non-linear reward function. We briefly describe the case where f^* lies in a reproducing kernel Hilbert space (RKHS).

Suppose that the reward function f^* is a function in an RKHS $\mathcal{H}$, with reproducing kernel $k : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$, and that the RKHS norm of f^* is bounded, i.e. $\|f^*\|_{\mathcal{H}} \leq B_{\mathcal{H}}$. If in the specialisation to linear bandits setting described at the beginning of Section 4.5.2, we instead choose $Z_t(g_t) = (g_t - f^*(a_t))\epsilon_t$, then one can derive an equivalent to Equation (4.8), which is

$$\|\boldsymbol{f}_t^* - \boldsymbol{r}_t\|_2^2 \leq (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2}\right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{r}_t) + \sigma^2 \ln\left(\det\left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2}\right)\right) + 2\sigma^2 \ln\frac{1}{\delta} =: R_{\mathrm{MM},t}^2,$$

where $\boldsymbol{f}_t^* = [f^*(a_1), \ldots, f^*(a_t)]^\top$ is the vector of the first t ground-truth function values. Letting $\boldsymbol{f}_t = [f(a_1), \ldots, f(a_t)]^\top$ denote the first t values of an arbitrary $f \in \mathcal{H}$, we can construct a familiar-looking confidence sequence for f^*.

Corollary 6.1. *For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, it holds with probability at least $1 - \delta$ that for all $t \geq 1$ simultaneously f^* lies in the set*

$$\mathcal{F}_t = \left\{ f \in \mathcal{H} \,\middle|\, \|\boldsymbol{f}_t - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t} \quad \text{and} \quad \|f\|_{\mathcal{H}} \leq B_{\mathcal{H}} \right\}.$$

Each $\mathcal{F}_t$ is a possibly infinite-dimensional set and the corresponding upper confidence bound $\text{UCB}_{\mathcal{F}_t}(a) := \max_{f \in \mathcal{F}_t} \{f(a)\}$ is the solution of a possibly infinite-dimensional optimisation problem. However, due to the RKHS norm constraint in $\mathcal{F}_t$, one can use the representer theorem [93] to reduce this to a finite-dimensional optimisation problem in weight space (or an equivalent one over a finite-dimensional vector of function values).

Lemma 6.2. *With $\mathcal{F}_t$ as defined in Corollary 6.1,* $\text{UCB}_{\mathcal{F}_t}(a)$ *is the solution of the convex (conic) program*

$$\max_{\boldsymbol{w}_{t+1} \in \mathbb{R}^{t+1}} \boldsymbol{k}_{t+1}(a_{t+1})^\top \boldsymbol{w}_{t+1} \quad s.t. \quad \|\boldsymbol{K}_{t,t+1}\boldsymbol{w}_{t+1} - \boldsymbol{r}_t\|_2 \leq R_{\text{MM},t}, \quad \|\boldsymbol{L}_{t+1}\boldsymbol{w}_{t+1}\|_2 \leq B_{\mathcal{H}},$$

where $a_{t+1} = a$, $\boldsymbol{k}_{t+1}(a_{t+1}) = [k(a_{t+1}, a_1), \ldots, k(a_{t+1}, a_{t+1})]^\top$, $\boldsymbol{K}_{t,t+1}$ is the $t \times t+1$ kernel matrix with i, j^{th} element equal to $k(a_i, a_j)$ and $\boldsymbol{L}_{t+1}$ is any matrix satisfying $\boldsymbol{L}_{t+1}^\top \boldsymbol{L}_{t+1} = \boldsymbol{K}_{t+1}$ for the usual kernel matrix $\boldsymbol{K}_{t+1}$ (e.g. $\boldsymbol{L}_{t+1}$ could be the right Cholesky factor of $\boldsymbol{K}_{t+1}$).

From here, we could use differentiable convex optimisation to compute $\text{UCB}_{\mathcal{F}_t}(a)$ and its gradient (as in CMM-UCB) or use an analytic upper bound on $\text{UCB}_{\mathcal{F}_t}(a)$ (as in AMM-UCB). This is all we need to run a kernel bandit version of CMM-UCB or AMM-UCB. While the data-dependent regret bound in Theorem 4.8 would remain basically unaltered, the data-independent regret bound in Theorem 4.9 would need to be modified. The main challenge is that the regret bound must now depend on quantities like the effective dimension (see e.g. [177]) or the maximum information gain (see e.g. [168]) of the kernel instead of the dimension of the feature vectors, which is $d = \infty$ for interesting kernels (e.g. RBF or Matérn).

Bibliography

[1] Y. Abbasi-Yadkori. *Online learning for linearly parametrized control problems*. PhD thesis, University of Alberta, 2012.

[2] Y. Abbasi-Yadkori, A. Antos, and C. Szepesvári. Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*, volume 92, page 236, 2009.

[3] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[4] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

[5] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9. PMLR, 2012.

[6] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[7] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and J. Z. Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

[8] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

[9] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[10] M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun. Review and analysis of risk factor of maternal health in remote area using the internet of things (iot). In *InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019*, pages 357–365. Springer, 2020.

[11] P. Alquier. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

[12] P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14(1), 2013.

[13] P. Alquier and K. Lounici. Pac-bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, pages 127–145, 2011.

[14] A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter pac-bayes bounds. *Advances in neural information processing systems*, 19, 2006.

[15] R. Amit and R. Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *International Conference on Machine Learning*, pages 205–214. PMLR, 2018.

[16] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1):5–43, 2003.

[17] M. Anthony, P. L. Bartlett, P. L. Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

[18] K. Ariu, K. Abe, and A. Proutière. Thresholded lasso bandit. In *International Conference on Machine Learning*, pages 878–928. PMLR, 2022.

[19] J.-Y. Audibert, S. Bubeck, et al. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[20] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[21] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[22] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[23] J. A. Bagnell and J. Schneider. Covariant policy search. In *International Joint Conference on Artificial Intelligence*, 2003.

[24] M. Balázs. Lecture notes on martingale theory with applications, 2021.

[25] R. Bardenet, A. Doucet, and C. C. Holmes. On markov chain monte carlo methods for tall data. *Journal of Machine Learning Research*, 18(47), 2017.

[26] H. Bastani and M. Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

[27] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19:521–547, 2013.

[28] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

[29] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[30] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[31] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[32] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[33] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[34] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[35] R. Camilleri, K. Jamieson, and J. Katz-Samuels. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning*, pages 1227–1237. PMLR, 2021.

[36] O. Catoni. A PAC-Bayesian approach to adaptive classification. *preprint*, 840, 2003.

[37] O. Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, volume 1851. Springer Science & Business Media, 2004.

[38] O. Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *Lecture Notes-Monograph Series. Institute of Mathematical Statistics.*, 2007.

[39] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[40] S. Chakraborty, S. Roy, and A. Tewari. Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pages 3979–4008. PMLR, 2023.

[41] S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

[42] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[43] B. Chugg, H. Wang, and A. Ramdas. A unified recipe for deriving (time-uniform) pac-bayes bounds. *arXiv preprint arXiv:2302.03421*, 2023.

[44] I. Cinar, M. Koklu, and S. Tasdemir. Classification of raisin grains using machine vision and artificial intelligence methods. *Gazi Mühendislik Bilimleri Dergisi*, 6(3):200–209, 2020.

[45] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *International Conference on Learning Representations (ICLR)*, 2015.

[46] M. C. Cohen, I. Lobel, and R. Paes Leme. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

[47] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[48] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

[49] D. Darling and H. Robbins. Some further remarks on inequalities for sample sums. *Proceedings of the National Academy of Sciences*, 60(4):1175–1182, 1968.

[50] V. H. de la Peña, M. J. Klass, and T. Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3), 2004.

[51] V. H. de la Peña, T. L. Lai, and Q.-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

[52] S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[53] M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time—iii. *Communications on pure and applied Mathematics*, 29(4):389–461, 1976.

[54] D. Dua and C. Graff. UCI machine learning repository, 2017.

[55] M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

[56] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pages 67–82. PMLR, 2018.

[57] A. Durand, O.-A. Maillard, and J. Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *The Journal of Machine Learning Research*, 19(1):650–683, 2018.

[58] R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[59] G. K. Dziugaite, K. Hsu, W. Gharbieh, G. Arpino, and D. Roy. On the role of data in pac-bayes bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 604–612. PMLR, 2021.

[60] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Conference on Uncertainty in Artificial Intelligence [UAI]*, 2017.

[61] G. K. Dziugaite and D. M. Roy. Data-dependent pac-bayes priors via differential privacy. *Advances in neural information processing systems*, 31, 2018.

[62] H. Flynn, D. Reeb, M. Kandemir, and J. Peters. Pac-bayesian lifelong learning for multi-armed bandits. *Data Mining and Knowledge Discovery*, 36(2):841–876, 2022.

[63] A. Foong, W. Bruinsma, D. Burt, and R. Turner. How tight can pac-bayes be in the small data regime? *Advances in Neural Information Processing Systems*, 34:4093–4105, 2021.

[64] D. Foster and A. Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

[65] E. I. George and R. E. McCulloch. Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373, 1997.

[66] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14:729–769, 2013.

[67] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. Pac-bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 353–360, 2009.

[68] P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

[69] P. D. Grünwald. The e-posterior. *Philosophical Transactions of the Royal Society A*, 381(2247):20220146, 2023.

[70] B. Guedj. A primer on pac-bayesian learning. In *Proceedings of the French Mathematical Society*, volume 33, pages 391–414. Société Mathématique de France, 2019.

[71] B. Guedj and P. Alquier. Pac-bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

[72] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017.

[73] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[74] M. Haddouche and B. Guedj. Online PAC-Bayes learning. *Advances in Neural Information Processing Systems*, 35:25725–25738, 2022.

[75] M. Haddouche and B. Guedj. PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research [TMLR]*, 2023.

[76] B. Hao, T. Lattimore, and M. Wang. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.

[77] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[78] R. Herbrich and T. Graepel. A pac-bayesian margin bound for linear classifiers: Why svms work. *Advances in neural information processing systems*, 13, 2000.

[79] T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

[80] S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[81] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[82] E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[83] H. Ishibashi, M. Karasuyama, I. Takeuchi, and H. Hino. A stopping criterion for bayesian optimization by the gap of expected minimum simple regrets. In *International Conference on Artificial Intelligence and Statistics*, pages 6463–6497. PMLR, 2023.

[84] H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of Statistics*, 33:730–773, 2005.

[85] K. Jang, C. Zhang, and K.-S. Jun. Popart: Efficient sparse regression and experimental design for optimal sparse linear bandits. *Advances in Neural Information Processing Systems*, 35:2102–2114, 2022.

[86] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[87] E. T. Jaynes. Information theory and statistical mechanics. ii. *Physical review*, 108(2):171, 1957.

[88] S. T. Jose, O. Simeone, and G. Durisi. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization. *IEEE Transactions on Information Theory*, 68(1):474–501, 2021.

[89] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[90] S. M. Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[91] E. Kaufmann and W. M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *The Journal of Machine Learning Research*, 22(1):11140–11183, 2021.

[92] G.-S. Kim and M. C. Paik. Doubly-robust lasso bandit. *Advances in Neural Information Processing Systems*, 32, 2019.

[93] G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[94] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

[95] J. Kirschner and A. Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

[96] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

[97] I. Kuzborskij and C. Szepesvári. Efron-stein pac-bayesian inequalities. *arXiv preprint arXiv:1909.01931*, 2019.

[98] I. Kuzborskij, C. Vernade, A. Gyorgy, and C. Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pages 640–648. PMLR, 2021.

[99] T. L. Lai. On confidence sequences. *The Annals of Statistics*, pages 265–280, 1976.

[100] J. Langford and R. Caruana. (not) bounding the true error. *Advances in Neural Information Processing Systems*, 14, 2001.

[101] J. Langford and J. Shawe-Taylor. Pac-bayes & margins. *Advances in neural information processing systems*, 15, 2002.

[102] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

[103] T. Lattimore, K. Crammer, and C. Szepesvári. Linear multi-resource allocation with semi-bandit feedback. *Advances in Neural Information Processing Systems*, 28, 2015.

[104] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[105] T. Lattimore, C. Szepesvari, and G. Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

[106] F. Laviolette. A tutorial on pac-bayesian theory. In *Talk at the NIPS 2017 Workshop:(Almost)*, volume 50, 2017.

[107] G. Letarte, P. Germain, B. Guedj, and F. Laviolette. Dichotomize and generalize: Pac-bayesian binary activated deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[108] G. Lever, F. Laviolette, and J. Shawe-Taylor. Distribution-dependent pac-bayes priors. In *International Conference on Algorithmic Learning Theory*, pages 119–133. Springer, 2010.

[109] G. Lever, F. Laviolette, and J. Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

[110] C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[111] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[112] Z. Li and J. Scarlett. Gaussian process bandit optimization with few batches. In *International Conference on Artificial Intelligence and Statistics*, pages 92–107. PMLR, 2022.

[113] J. S. Liu and J. S. Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

[114] T. Liu, J. Lu, Z. Yan, and G. Zhang. Statistical generalization performance guarantee for meta-learning with data dependent prior. *Neurocomputing*, 465:391–405, 2021.

[115] B. London and T. Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, pages 4125–4133. PMLR, 2019.

[116] A. Makarova, H. Shen, V. Perrone, A. Klein, J. B. Faddoul, A. Krause, M. Seeger, and C. Archambeau. Automatic termination for hyperparameter optimization. In *International Conference on Automated Machine Learning*, pages 7–1. PMLR, 2022.

[117] J. Mary, R. Gaudel, and P. Preux. Bandits and recommender systems. In *International Workshop on Machine Learning, Optimization and Big Data*, pages 325–336. Springer, 2015.

[118] A. Maurer. A note on the pac bayesian theorem. *arXiv preprint cs/0411099*, 2004.

[119] D. McAllester. Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer, 2003.

[120] D. McAllester. A pac-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013.

[121] D. A. McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

[122] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[123] D. Meunier and P. Alquier. Meta-strategy for learning tuning parameters with guarantees. *Entropy*, 23(10):1257, 2021.

[124] K. Misra, E. M. Schwartz, and J. Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019.

[125] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

[126] I. S. Molchanov and I. S. Molchanov. *Theory of random sets*, volume 19. Springer, 2005.

[127] J. W. Mueller, V. Syrgkanis, and M. Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. *Advances in Neural Information Processing Systems*, 32, 2019.

[128] W. Neiswanger and A. Ramdas. Uncertainty quantification using martingales for misspecified gaussian processes. In *Algorithmic Learning Theory*, pages 963–982. PMLR, 2021.

[129] M.-h. Oh, G. Iyengar, and A. Zeevi. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR, 2021.

[130] L. Oneto, D. Anguita, and S. Ridella. Pac-bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis. *Pattern Recognition Letters*, 80:200–207, 2016.

[131] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. Pac-bayes bounds with data dependent priors. *The Journal of Machine Learning Research*, 13(1):3507–3531, 2012.

[132] A. Pentina and C. Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999. PMLR, 2014.

[133] M. Pérez-Ortiz, O. Rivasplata, B. Guedj, M. Gleeson, J. Zhang, J. Shawe-Taylor, M. Bober, and J. Kittler. Learning pac-bayes priors for probabilistic neural networks. *arXiv preprint arXiv:2109.10304*, 2021.

[134] M. Perez-Ortiz, O. Rivasplata, E. Parrado-Hernandez, B. Guedj, and J. Shawe-Taylor. Progress in self-certified neural networks. In *NeurIPS 2021 workshop: Bayesian Deep Learning*. NeurIPS, 2021.

[135] M. Pérez-Ortiz, O. Rivasplata, J. Shawe-Taylor, and C. Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22, 2021.

[136] J. Peters, K. Mulling, and Y. Altun. Relative entropy policy search. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[137] J. Peters, S. Vijayakumar, and S. Schaal. Reinforcement learning for humanoid robotics. In *Proceedings of the third IEEE-RAS international conference on humanoid robots*, pages 1–20, 2003.

[138] K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[139] M. S. Pinsker. *Information and information stability of random variables and processes*. Holden-Day, 1964.

[140] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[141] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Sciences*, 2023.

[142] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

[143] D. Reeb, A. Doerr, S. Gerwinn, and B. Rakitsch. Learning gaussian processes by minimizing pac-bayesian generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.

[144] O. Rivasplata, I. Kuzborskij, C. Szepesvári, and J. Shawe-Taylor. Pac-bayes analysis beyond the usual bounds. *Advances in Neural Information Processing Systems*, 33:16833–16845, 2020.

[145] O. Rivasplata, E. Parrado-Hernández, J. S. Shawe-Taylor, S. Sun, and C. Szepesvári. Pac-bayes bounds for stable algorithms with instance-dependent priors. *Advances in Neural Information Processing Systems*, 31, 2018.

[146] O. Rivasplata, V. M. Tankasali, and C. Szepesvári. Pac-bayes with backprop. *arXiv preprint arXiv:1908.07380*, 2019.

[147] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[148] H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

[149] H. Robbins and D. Siegmund. Boundary crossing probabilities for the wiener process and sample sums. *The Annals of Mathematical Statistics*, pages 1410–1429, 1970.

[150] J. Rothfuss, V. Fortuin, M. Josifoski, and A. Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*, pages 9116–9126. PMLR, 2021.

[151] J. Rothfuss, D. Heyn, A. Krause, et al. Meta-learning reliable priors in the function space. *Advances in Neural Information Processing Systems*, 34:280–293, 2021.

[152] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

[153] D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[154] O. Sakhi, P. Alquier, and N. Chopin. Pac-bayesian offline contextual bandits with guarantees. In *International Conference on Machine Learning*, pages 29777–29799. PMLR, 2023.

[155] S. Salgia, S. Vakili, and Q. Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. *Advances in Neural Information Processing Systems*, 34:28836–28847, 2021.

[156] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[157] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[158] M. Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of machine learning research*, 3(Oct):233–269, 2002.

[159] Y. Seldin, P. Auer, J. Shawe-taylor, R. Ortner, and F. Laviolette. Pac-bayesian analysis of contextual bandits. *Advances in neural information processing systems*, 24, 2011.

[160] Y. Seldin, N. Cesa-Bianchi, P. Auer, F. Laviolette, and J. Shawe-Taylor. Pac-bayes-bernstein inequality for martingales and its application to multiarmed bandits. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, pages 98–111. JMLR Workshop and Conference Proceedings, 2012.

[161] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

[162] Y. Seldin, F. Laviolette, J. Shawe-Taylor, J. Peters, and P. Auer. Pac-bayesian analysis of martingales and multiarmed bandits. *arXiv preprint arXiv:1105.2416*, 2011.

[163] Y. Seldin, C. Szepesvári, P. Auer, and Y. Abbasi-Yadkori. Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments. In *European Workshop on Reinforcement Learning*, pages 103–116. PMLR, 2013.

[164] Y. Seldin and N. Tishby. Pac-bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11(12), 2010.

[165] J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9, 1997.

[166] J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

[167] A. Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

[168] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. International Conference on Machine Learning (ICML)*, 2010.

[169] A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

[170] N. Thiemann, C. Igel, O. Wintenberger, and Y. Seldin. A strongly quasiconvex pac-bayesian bound. In *International Conference on Algorithmic Learning Theory*, pages 466–492. PMLR, 2017.

[171] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[172] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[173] M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[174] I. O. Tolstikhin and Y. Seldin. Pac-bayes-empirical-bernstein inequality. *Advances in Neural Information Processing Systems*, 26, 2013.

[175] Y. Tsuzuku, I. Sato, and M. Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In *International Conference on Machine Learning*, pages 9636–9647. PMLR, 2020.

[176] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[177] M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-Time Analysis of Kernelised Contextual Bandits. In *Uncertainty in Artificial Intelligence*, 2013.

[178] T. van Erven. Pac-bayes mini-tutorial: A continuous union bound. *arXiv preprint arXiv:1405.1580*, 2014.

[179] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. Openml: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

[180] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[181] J. Ville. Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939.

[182] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[183] J. Wang, B. Mazoure, G. McCracken, D. Venuto, and A. Durand. Pac-bayesian analysis of counterfactual risk in stochastic contextual bandits. In *Multi-disciplinary Conference on Reinforcement Learning and Decision Making*, 2019.

[184] X. Wang, M. Wei, and T. Yao. Minimax concave penalized multi-armed bandit model with high-dimensional covariates. In *International Conference on Machine Learning*, pages 5200–5208. PMLR, 2018.

[185] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

[186] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[187] D. Williams. *Probability with martingales*. Cambridge university press, 1991.

[188] J. Yang, S. Sun, and D. M. Roy. Fast-rate pac-bayes generalization bounds via shifted rademacher processes. *Advances in Neural Information Processing Systems*, 32, 2019.

[189] T. Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37:2109–2144, 2009.

[190] B. D. Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

Appendix A

Appendix for Chapter 3

A.1 Proofs

A.1.1 $M_T^{\mathrm{IS}}(\pi)$ is a martingale

Lemma A.1. *The sequence* $\{X_t^{\mathrm{IS}}(\pi)\}_{t=1}^T$ *defined as*

$$X_t^{\mathrm{IS}}(\pi) = \frac{\pi(a_t)}{b_t(a_i)} r_t - R(\pi),$$

is a martingale difference sequence with respect to $\{(a_t, r_t)\}_{t=1}^T$*. Moreover, if the importance weights* $\pi(a_t)/b_t(a_t)$ *are uniformly bounded above by* $1/\epsilon_T$*, then each* $X_t^{\mathrm{IS}}(\pi)$ *is uniformly bounded in the range* $[-R(\pi), 1/\epsilon_T - R(\pi)]$*, and the sum of the sequence is*

$$\sum_{t=1}^{T} X_t^{\mathrm{IS}}(\pi) = T(r^{\mathrm{IS}}(\pi, D_T) - R(\pi)).$$

Since $M_T^{\mathrm{IS}}(\pi) = \sum_{t=1}^T X_t^{\mathrm{IS}}(\pi)$, Lemma A.1 shows that $M_T^{\mathrm{IS}}(\pi)$ is a martingale.

Proof of Lemma A.1. We first verify that, for any $\pi \in \Pi$, $\{X_t^{\mathrm{IS}}(\pi)\}_{t=1}^T$ is a martingale difference sequence with respect to $\{(a_t, r_t)\}_{t=1}^T$.

$$\begin{aligned}
\mathbb{E}\left[X_t^{\mathrm{IS}}(\pi)\middle| D_{t-1}\right] &= \mathop{\mathbb{E}}_{a_t\sim b_t, r_t\sim\mathbb{P}_R(\cdot|a_t)}\left[\frac{\pi(a_t)}{b_t(a_t)} r_t - R(\pi)\right], \\
&= \mathop{\mathbb{E}}_{a_t\sim b_t}\left[\frac{\pi(a_t)}{b_t(a_t)} \mathop{\mathbb{E}}_{r_t\sim\mathbb{P}_R(\cdot|a_t)}[r_t]\right] - R(\pi), \\
&= \mathop{\mathbb{E}}_{a_t\sim \pi}\left[\mathop{\mathbb{E}}_{r_t\sim\mathbb{P}_R(\cdot|a_t)}[r_t]\right] - R(\pi), \\
&= R(\pi) - R(\pi) = 0.
\end{aligned}$$

Next, we verify that each $X_t^{\mathrm{IS}}(\pi)$ is bounded in the interval $[-R(\pi), 1/\epsilon_T - R(\pi)]$. If the importance weights $\pi(a_t)/b_t(a_t)$ are uniformly bounded above by $1/\epsilon_T$ and the rewards are bounded in $[0,1]$, then for any t, $(\pi(a_t)/b_t(a_t))r_t \in [0, 1/\epsilon_T]$. Therefore, $X_t^{\mathrm{IS}}(\pi) = (\pi(a_t)/b_t(a_t))r_t - R(\pi) \in [-R(\pi), 1/\epsilon_T - R(\pi)]$.

Finally, we verify that $\{X_t^{\mathrm{IS}}(\pi)\}_{t=1}^T$ sums to $T(r^{\mathrm{IS}}(\pi, D_T) - R(\pi))$.

$$\sum_{t=1}^{T} X_t^{\mathrm{IS}}(\pi) = \sum_{t=1}^{T} \left(\frac{\pi(a_t)}{b_t(a_t)} r_t - R(\pi) \right) = T \left(\frac{1}{T} \sum_{t=1}^{T} \frac{\pi(a_t)}{b_t(a_t)} r_t - R(\pi) \right) = T \left(r^{\mathrm{IS}}(\pi, D_T) - R(\pi) \right).$$

□

A.1.2 Bias of the CIS estimate

Lemma A.2 (Bias of the CIS estimate)**.** *The expected value of the CIS estimate satisfies*

$$R^{\mathrm{CIS}}(\rho) \leq R(\rho). \tag{A.1}$$

Proof of Lemma A.2. First, we show that $r^{\mathrm{IS}}(\pi, D_T)$ is an unbiased estimate of $R(\pi)$. For any $\pi \in \Pi$, any $t \in 1, \ldots, T$, and any history D_{t-1}, we have that

$$\mathop{\mathbb{E}}_{a_t \sim b_t, r_t \sim \mathbb{P}_R(\cdot|a_t)} \left[\frac{\pi(a_t)}{b_t(a_t)} r_t \right] = \mathop{\mathbb{E}}_{a_t \sim b_t} \left[\frac{\pi(a_t)}{b_t(a_t)} \mathop{\mathbb{E}}_{r_t \sim \mathbb{P}_R(\cdot|a_t)} [r_t] \right] = \mathop{\mathbb{E}}_{a_t \sim \pi(\cdot)} \left[\mathop{\mathbb{E}}_{r_t \sim \mathbb{P}_R(\cdot|a_t)} [r_t] \right] = R(\pi).$$

Therefore, we have

$$\mathop{\mathbb{E}}_{D_T} \left[r^{\mathrm{IS}}(\pi, D_T) \right] = \mathop{\mathbb{E}}_{D_T} \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\pi(a_t)}{b_t(a_t)} r_t \right] = \frac{1}{T} \sum_{t=1}^{T} R(\pi) = R(\pi).$$

Finally, we have

$$\mathop{\mathbb{E}}_{D_T} \left[r^{\mathrm{CIS}}(\pi, D_T) \right] = \mathop{\mathbb{E}}_{D_T} \left[\frac{1}{T} \sum_{t=1}^{T} \min \left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau} \right) r_t \right] \leq \mathop{\mathbb{E}}_{D_T} \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\pi(a_t)}{b(a_t)} r_t \right] = \mathop{\mathbb{E}}_{D_T} \left[r^{\mathrm{IS}}(\pi, D_T) \right] = R(\pi).$$

Taking the expected value $\mathbb{E}_{\pi \sim \rho}[\cdot]$ of both sides yields the statement of the lemma. The proof for the CB case is the same. □

A.1.3 Proof of Theorem 3.7

First we state and prove a one-sided version of the Hoeffding-Azuma inequality for supermartingale difference sequences.

Lemma A.3 (One-Sided Hoeffding-Azuma inequality)**.** *Let $X_1, \ldots, X_T$ be a supermartingale difference sequence (meaning $\mathbb{E}[X_t | X_1, \ldots, X_{t-1}] \leq 0$ for $t = 1, \ldots, T$) where each X_t is bounded in the interval $[a, b]$. Then for any $\lambda \geq 0$, we have*

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_t}\left[e^{\lambda \sum_{t=1}^{T} X_t}\right] \leq e^{\frac{T\lambda^2(b-a)^2}{8}}.$$

Proof of Lemma A.3. First, by Hoeffding's Lemma (see, for example, Lemma A.1 of [39]), for any random variable bounded in the interval $[a, b]$ and any $\lambda \in \mathbb{R}$

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda \mathbb{E}[X] + \frac{\lambda^2}{8}(b-a)^2}.$$

Now, for any $\lambda \geq 0$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{X_1,\ldots,X_T}\left[e^{\lambda \sum_{t=1}^{T} X_t}\right] &= \mathop{\mathbb{E}}_{X_1,\ldots,X_T}\left[\prod_{t=1}^{T} e^{\lambda X_t}\right], \\
&= \mathop{\mathbb{E}}_{X_1,\ldots,X_{T-1}}\left[\mathop{\mathbb{E}}_{X_T}\left[\prod_{t=1}^{T} e^{\lambda X_t} \;\middle|\; X_1, \ldots, X_{T-1}\right]\right], \\
&\leq \mathop{\mathbb{E}}_{X_1,\ldots,X_{T-1}}\left[e^{\lambda\, \mathbb{E}_{X_T}[X_T | X_1,\ldots,X_{T-1}] + \frac{\lambda^2}{8}(b-a)^2} \prod_{t=1}^{T-1} e^{\lambda X_t}\right], \\
&\leq e^{\frac{\lambda^2}{8}(b-a)^2} \mathop{\mathbb{E}}_{X_1,\ldots,X_{T-1}}\left[e^{\lambda \sum_{t=1}^{T-1} X_t}\right].
\end{aligned}$$

By iterating the above steps, we obtain

$$\mathop{\mathbb{E}}_{X_1,\ldots,X_T}\left[e^{\lambda \sum_{t=1}^{T} X_t}\right] \leq \prod_{t=1}^{T} e^{\frac{\lambda^2(b-a)^2}{8}} = e^{\frac{T\lambda^2(b-a)^2}{8}}.$$

□

Next, we show that the sequence $\{Y_t^{\mathrm{CIS}}(\pi)\}_{t=1}^{T}$, defined as

$$Y_t^{\mathrm{CIS}}(\pi) = \min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t - R(\pi),$$

is a supermartingale difference sequence with respect to $\{(a_t, r_t)\}_{t=1}^{T}$, and that each term $Y_t^{\mathrm{CIS}}(\pi)$

is bounded in the interval $[-R(\pi), 1/\tau - R(\pi)]$. First, we have

$$\begin{aligned}
\mathbb{E}\left[Y_t^{\mathrm{CIS}}(\pi)\middle| D_{t-1}\right] &= \mathop{\mathbb{E}}_{a_t\sim b_t, r_t\sim\mathbb{P}_R(\cdot|a_t)}\left[Y_t^{\mathrm{CIS}}(\pi)\right], \\
&= \mathop{\mathbb{E}}_{a_t\sim b_t, r_t\sim\mathbb{P}_R(\cdot|a_t)}\left[\min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t - R(\pi)\right], \\
&\leq \mathop{\mathbb{E}}_{a_t\sim b_t, r_t\sim\mathbb{P}_R(\cdot|a_t)}\left[\frac{\pi(a_t)}{b_t(a_t)} r_t - R(\pi)\right], \\
&= \mathop{\mathbb{E}}_{a_t\sim b_t}\left[\frac{\pi(a_t)}{b_t(a_t)} \mathop{\mathbb{E}}_{r_t\sim\mathbb{P}_R(\cdot|a_t)}[r_t]\right] - R(\pi), \\
&= \mathop{\mathbb{E}}_{a_t\sim\pi(\cdot)}\left[\mathop{\mathbb{E}}_{r_t\sim\mathbb{P}_R(\cdot|a_t)}[r_t]\right] - R(\pi), \\
&= R(\pi) - R(\pi) = 0.
\end{aligned}$$

Since $\min\left(\pi(a_t)/b(a_t|D_{t-1}), 1/\tau\right) r_t \in [0, 1/\tau]$, we have $Y_t^{\mathrm{CIS}}(\pi) \in [-R(\pi), 1/\tau - R(\pi)]$. Therefore, the sequence $\{Y_t^{\mathrm{CIS}}(\pi)\}_{t=1}^T$ is compatible with the one-sided Hoeffding-Azuma inequality in Lemma A.3, with $a = -R(\pi)$ and $b = 1/\tau - R(\pi)$. In addition, we have $(\lambda/n)\sum_{t=1}^T Y_t^{\mathrm{CIS}}(\pi) = \lambda(r^{\mathrm{CIS}}(\pi, D_T) - R(\pi))$.

To prove Theorem 3.7, we begin by using Lemma A.3 with $\lambda(r^{\mathrm{CIS}}(\pi, D_T) - R(\pi))$ and then use Toelli's theorem to swap expectations followed by the Donsker-Varadhan change of measure inequality (see Equation (2.4)).

$$\begin{aligned}
\mathop{\mathbb{E}}_{D_T}\left[\exp(\lambda(r^{\mathrm{CIS}}(\pi, D_T) - R(\pi)))\right] &\leq \exp\left(\frac{\lambda^2}{8T\tau^2}\right) \\
\mathop{\mathbb{E}}_{\pi\sim P}\mathop{\mathbb{E}}_{D_T}\left[\exp(\lambda(r^{\mathrm{CIS}}(\pi, D_T) - R(\pi)))\right] &\leq \exp\left(\frac{\lambda^2}{8T\tau^2}\right) \\
\mathop{\mathbb{E}}_{D_T}\mathop{\mathbb{E}}_{\pi\sim P}\left[\exp(\lambda(r^{\mathrm{CIS}}(\pi, D_T) - R(\pi)))\right] &\leq \exp\left(\frac{\lambda^2}{8T\tau^2}\right) \\
\mathop{\mathbb{E}}_{D_T}\left[\exp\left(\sup_{Q\in\mathcal{P}}\left\{\mathop{\mathbb{E}}_{\pi\sim Q}\left[\lambda(r^{\mathrm{CIS}}(\pi, D_T) - R(\pi))\right] - D_{\mathrm{KL}}(Q||P)\right\}\right)\right] &\leq \exp\left(\frac{\lambda^2}{8T\tau^2}\right)
\end{aligned}$$

Now, we use Markov's inequality and take the logarithm of both sides. This tells us that for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, we have

$$\mathop{\mathbb{E}}_{\pi\sim Q}\left[\lambda(r^{\mathrm{CIS}}(\pi, D_T) - R(\pi))\right] - D_{\mathrm{KL}}(Q||P) \leq \frac{\lambda^2}{8T\tau^2} + \ln(1/\delta).$$

This inequality can be rearranged into the statement of Theorem 3.7.

A.1.4 Variance of the CIS estimate

Lemma A.4 (Variance of the CIS estimate)**.** *The average variance of the CIS estimate (both the MAB and CB versions) satisfies*

$$V^{\mathrm{CIS}}(\pi, D_T) \leq \frac{1}{\tau}.$$

Proof of Lemma A.4. We let

$$X_t^{\mathrm{CIS}}(\pi) = \min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t - \mathop{\mathbb{E}}_{a_t \sim b_t, r_t \sim \mathbb{P}_R(\cdot|a_t)}\left[\min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t\right]. \tag{A.2}$$

To bound $V^{\mathrm{CIS}}(\pi, D_T)$, we use that fact that the rewards are bounded in the interval $[0, 1]$.

$$\begin{aligned}
V^{\mathrm{CIS}}(\pi, D_T) &= \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{a_t \sim b_t, r_t \sim \mathbb{P}_R(\cdot|a_t)}\left[\left(X_t^{\mathrm{CIS}}(\pi)\right)^2\right], \\
&= \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{a_t \sim b_t, r_t \sim \mathbb{P}_R(\cdot|a_t)}\left[\min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right)^2 {r_t}^2\right] - \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{a_t \sim b_t, r_t \sim \mathbb{P}_R(\cdot|a_t)}\left[\min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t\right]^2, \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{a_t \sim b_t}\left[\min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right)^2\right], \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\tau} \mathop{\mathbb{E}}_{a_t \sim b_t}\left[\frac{\pi(a_t)}{b_t(a_t)}\right], \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\tau} = \frac{1}{\tau}.
\end{aligned}$$

□

A.1.5 Proof of Theorem 3.8

First, we state Bernstein's inequality for martingales.

Lemma A.5 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_T$ be a martingale difference sequence where each X_t is bounded in the interval $[-b, b]$, for some $b > 0$. Then for all $\lambda \in [0, 1/b]$*

$$\mathop{\mathbb{E}}_{X_1, \ldots, X_T}\left[e^{\lambda \sum_{t=1}^{T} X_t - (e-2)\lambda^2 \sum_{t=1}^{T} \mathbb{E}_{X_t}[X_t^2 | X_1, \ldots, X_{t-1}]}\right] \leq 1.$$

For a proof, see Theorem 1 of [28]. Next, we show that the sequence $\{X_t^{\mathrm{CIS}}(\pi)\}_{t=1}^{T}$, defined in Equation A.2, is a martingale difference sequence with respect to $\{(a_t, r_t)\}_{t=1}^{T}$, and that each term

is bounded in the interval $[-1/\tau, 1/\tau]$. For any $\tau \in (0, 1]$, we have

$$\begin{aligned}\mathbb{E}\left[X_t^{\mathrm{CIS}}(\pi)\middle| D_{t-1}\right] &= \mathop{\mathbb{E}}_{\substack{a_t\sim b_t,\\ r_t\sim\mathbb{P}_R(\cdot|a_t)}}\left[X_t^{\mathrm{CIS}}(\pi)\right] \\ &= \mathop{\mathbb{E}}_{\substack{a_t\sim b_t,\\ r_t\sim\mathbb{P}_R(\cdot|a_t)}}\left[\min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t - \mathop{\mathbb{E}}_{\substack{a_t\sim b_t,\\ r_t\sim\mathbb{P}_R(\cdot|a_t)}}\left[\min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t\right]\right], \\ &= 0.\end{aligned}$$

For any $t \in \{1, \ldots, T\}$, we have

$$0 \leq \mathop{\mathbb{E}}_{a_t\sim b_t, r_t\sim\mathbb{P}_R(\cdot|a_t)}\left[\min\left(\frac{\pi(a_t)}{b_t(a_t)}, \frac{1}{\tau}\right) r_t\right] \leq \mathop{\mathbb{E}}_{a_t\sim b_t, r_t\sim\mathbb{P}_R(\cdot|a_t)}\left[\frac{\pi(a_t)}{b_t(a_t)} r_t\right] = R(\pi) \leq 1.$$

Since $\min\left(\pi(a_t)/b_t(a_t), 1/\tau\right) r_t \in [0, 1/\tau]$, we have $X_t^{\mathrm{CIS}}(\pi) \in [-1, 1/\tau] \subseteq [-1/\tau, 1/\tau]$. To prove Theorem 3.8, we can follow the final sequence of steps taken in the proof of Theorem 3.7, except starting from

$$\mathop{\mathbb{E}}_{D_T}\left[\exp\left(\lambda \sum_{t=1}^{T} X_t^{\mathrm{CIS}}(\pi) - \lambda^2 (e-2) \sum_{t=1}^{T} \mathbb{E}\left[(X_t^{\mathrm{CIS}}(\pi))^2 | D_{t-1}\right]\right)\right] \leq 1.$$

A.1.6 Proof of the Efron-Stein PAC-Bayes Bound (Theorem 3.10)

First, we recall the statement of the theorem. Let $f(\pi, D_T)$ be a real-valued function, let $F(\pi) = \mathbb{E}_{D_T}[f(\pi, D_T)]$ denote its expected value and let $V^{\mathrm{ES}}(\pi, D_T)$ denote its semi-empirical Efron-Stein variance proxy (see Equation (3.18)). If the data set D_T consists of independent random variables, then for any distribution P on Π, any $y > 0$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ (over the sampling of D_T), we have that for all $Q \in \mathcal{P}(\Pi)$

$$|f(Q, D_T) - F(Q)| \leq \sqrt{2\left(y + V^{\mathrm{ES}}(Q, D_T)\right)\left(D_{\mathrm{KL}}(Q||P) + \frac{1}{2}\ln\left(1 + V^{\mathrm{ES}}(Q, D_T)/y\right) + \ln(1/\delta)\right)}.$$

In the proof, we use some technical lemmas. The first technical lemma is an Efron-Stein concentration inequality by Kuzborskij and Szepesvári [97].

Lemma A.6 (Efron-Stein concentration inequality [97])**.** *Let $D_T = \{Z_t\}_{t=1}^T$ be a collection of independent random variables. Then for any $\pi \in \Pi$ and any $\lambda \in \mathbb{R}$, we have*

$$\mathop{\mathbb{E}}_{D_T}\left[e^{\lambda(f(\pi, D_T) - \mathbb{E}_{D_T}[f(\pi, D_T)]) - \frac{\lambda^2}{2} V^{\mathrm{ES}}(\pi, D_T)}\right] \leq 1.$$

The second technical lemma allows us to swap the order of a supremum and an exponentiation.

Lemma A.7. *For any set $A \subseteq \mathbb{R}$ where $\sup(A)$ exists, we have that*

$$e^{\sup(A)} = \sup\left(e^A\right),$$

where $e^A = \{e^a | a \in A\}$.

Proof of Lemma A.7. Let $\sup(A) = \alpha$. Suppose that $\sup\left(e^A\right) < e^\alpha$. Then there exists an $\epsilon > 0$ such that $\sup\left(e^A\right) \leq e^{\alpha-\epsilon}$. Therefore, for all $a \in A$, $e^a \leq e^{\alpha-\epsilon}$, and so $\alpha - \epsilon$ is an upper bound on A. This is a contradiction since α is the least upper bound on A.

Now, suppose that $\sup\left(e^A\right) > e^\alpha$. This means that e^α is not an upper bound for e^A. Therefore, there must exist an $a \in A$, where $e^a > e^\alpha$, and so $a > \alpha$. This is a contradiction, since $\sup(A) = \alpha$. Therefore we must have $e^{\sup(A)} = \sup\left(e^A\right)$. □

The third technical lemma allows us to upper bound the supremum of an Gaussian integral/expected value by the integral of the supremum.

Lemma A.8. *For any function $g : \mathcal{P}(\Pi) \times \mathbb{R} \to \mathbb{R}$, such that $\sup_{Q\in\mathcal{P}(\Pi)}\{g(Q,\lambda)\} < \infty$ for all $\lambda \in \mathbb{R}$, we have*

$$\sup_{Q\in\mathcal{P}(\Pi)}\left\{\int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2 y^2}{2}}g(Q,\lambda)\mathrm{d}\lambda\right\} \leq \int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2 y^2}{2}}\sup_{Q\in\mathcal{P}(\Pi)}\{g(Q,\lambda)\}\mathrm{d}\lambda.$$

Proof of Lemma A.8. For every $Q \in \mathcal{P}(\Pi)$ and $\lambda \in \mathbb{R}$

$$\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2 y^2}{2}}g(Q,\lambda) \leq \frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2 y^2}{2}}\sup_{Q\in\mathcal{P}(\Pi)}\{g(Q,\lambda)\}.$$

Therefore, for every $Q \in \mathcal{P}(\Pi)$

$$\int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2 y^2}{2}}g(Q,\lambda)\mathrm{d}\lambda \leq \int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2 y^2}{2}}\sup_{Q\in\mathcal{P}(\Pi)}\{g(Q,\lambda)\}\mathrm{d}\lambda.$$

Therefore, we have

$$\sup_{Q\in\mathcal{P}(\Pi)}\left\{\int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2 y^2}{2}}g(Q,\lambda)\mathrm{d}\lambda\right\} \leq \int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2 y^2}{2}}\sup_{Q\in\mathcal{P}(\Pi)}\{g(Q,\lambda)\}\mathrm{d}\lambda.$$

□

Proof of Theorem 3.10. Throughout the proof, let $A(\pi, D_T) = f(\pi, D_T) - F(\pi)$. Also, let $A(Q, D_T) = \mathbb{E}_{\pi\sim Q}[A(Q, D_T)]$ and let $V^{\mathrm{ES}}(Q, D_T) = \mathbb{E}_{\pi\sim Q}[V^{\mathrm{ES}}(\pi, D_T)]$.

Using the Donsker-Varadhan change of measure inequality (Equation (2.4)), we have that, for all $\lambda \in \mathbb{R}$

$$\sup_{Q\in\mathcal{P}(\Pi)}\left\{\lambda A(Q, D_T) - \frac{\lambda^2}{2}V^{\mathrm{ES}}(Q, D_T) - D_{\mathrm{KL}}(Q||P)\right\} = \ln\left(\mathop{\mathbb{E}}_{\pi\sim P}\left[e^{\lambda A(\pi,D_T)-\frac{\lambda^2}{2}V^{\mathrm{ES}}(\pi,D_T)}\right]\right).$$

We exponentiate and then take expected values (over D_T) of both sides to obtain

$$\mathop{\mathbb{E}}_{D_T}\left[e^{\sup_{Q\in\mathcal{P}(\Pi)}\left\{\lambda A(Q,D_T)-\frac{\lambda^2}{2}V^{\mathrm{ES}}(Q,D_T)-D_{\mathrm{KL}}(Q||P)\right\}}\right]=\mathop{\mathbb{E}}_{D_T}\left[\mathop{\mathbb{E}}_{\pi\sim P}\left[e^{\lambda A(\pi,D_T)-\frac{\lambda^2}{2}V^{\mathrm{ES}}(\pi,D_T)}\right]\right].$$

We use Tonelli's theorem to swap the order of the expectations on the right-hand-side. We then use the Efron-Stein Concentration Inequality in Lemma A.6 to obtain

$$\mathop{\mathbb{E}}_{D_T}\left[e^{\sup_{Q\in\mathcal{P}(\Pi)}\left\{\lambda A(Q,D_T)-\frac{\lambda^2}{2}V^{\mathrm{ES}}(Q,D_T)-D_{\mathrm{KL}}(Q||P)\right\}}\right]\leq 1.$$

We use Lemma A.7, multiply both sides by $(y/\sqrt{2\pi})e^{-\frac{\lambda^2y^2}{2}}$ for $y>0$, and then integrate w.r.t. λ from $-\infty$ to ∞, which gives

$$\int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2y^2}{2}}\mathop{\mathbb{E}}_{D_T}\left[\sup_{Q\in\mathcal{P}(\Pi)}\left\{e^{\lambda A(Q,D_T)-\frac{\lambda^2}{2}V^{\mathrm{ES}}(Q,D_T)-D_{\mathrm{KL}}(Q||P)}\right\}\right]\mathrm{d}\lambda\leq 1.$$

We use Tonelli's theorem to swap the order of the integral and the expected value.

$$\mathop{\mathbb{E}}_{D_T}\left[\int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2y^2}{2}}\sup_{Q\in\mathcal{P}(\Pi)}\left\{e^{\lambda A(Q,D_T)-\frac{\lambda^2}{2}V^{\mathrm{ES}}(Q,D_T)-D_{\mathrm{KL}}(Q||P)}\right\}\mathrm{d}\lambda\right]\leq 1.$$

Using Lemma A.8, we can move the integral inside the supremum.

$$\mathop{\mathbb{E}}_{D_T}\left[\sup_{Q\in\mathcal{P}(\Pi)}\left\{\int_{-\infty}^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{\lambda^2y^2}{2}}e^{\lambda A(Q,D_T)-\frac{\lambda^2}{2}V^{\mathrm{ES}}(Q,D_T)-D_{\mathrm{KL}}(Q||P)}\mathrm{d}\lambda\right\}\right]\leq 1.$$

We can now calculate the integral by rearranging the integrand to get a Gaussian density function.

$$\mathop{\mathbb{E}}_{D_T}\left[\sup_{Q\in\mathcal{P}(\Pi)}\left\{\frac{y}{\sqrt{y^2+V^{\mathrm{ES}}(Q,D_T)}}e^{\frac{A(Q,D_T)^2}{2(y^2+V^{\mathrm{ES}}(Q,D_T))}-D_{\mathrm{KL}}(Q||P)}\right\}\right]\leq 1.$$

This holds for all $P\in\mathcal{P}(\Pi)$ and $y>0$. Now, we fix P and y, and then use Markov's inequality. With probability at least $1-\delta$, and for all $Q\in\mathcal{P}(\Pi)$ simultaneously, we have

$$\frac{y}{\sqrt{y^2+V^{\mathrm{ES}}(Q,D_T)}}e^{\frac{A(Q,D_T)^2}{2(y^2+V^{\mathrm{ES}}(Q,D_T))}-D_{\mathrm{KL}}(Q||P)}\leq 1/\delta.$$

We can rearrange this inequality to obtain the following inequality, that holds with the same probability

$$|A(\rho,D_n)|\leq\sqrt{2\left(y^2+V^{\mathrm{ES}}(\rho,D_n)\right)\left(D_{\mathrm{KL}}(\rho||\mu)+\frac{1}{2}\ln\left(1+V^{\mathrm{ES}}(\rho,D_n)/y^2\right)+\ln(1/\delta)\right)}.$$

Since y was an arbitrary positive number, we can replace y^2 with y to recover the statement of the Theorem. □

A.1.7 Proof for the Localised PAC-Bayes Bernstein Bound (Theorem 3.25)

This proof follows steps in Section 1.3.4. of [38]. First, we recall the statement of the Theorem. For any $\lambda \in [0, T\epsilon_T]$, any β satisfying $0 \leq \beta < \lambda$, any $\delta \in (0, 1]$ and any probability distribution $P \in \mathcal{P}(\Pi)$, with probability at least $1 - \delta$, for all distributions $Q \in \mathcal{P}(\Pi)$ simultaneously

$$R(Q) \geq r^{\text{IS}}(Q, D_T) - \frac{(\lambda^2 + \beta^2)(e-2)}{(\lambda - \beta)T\epsilon_T} - \frac{D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}}) + 2\ln(1/\delta)}{\lambda - \beta}.$$

Also, recall that $P_{\beta R}$ and $P_{\beta r^{\text{IS}}}$ are Gibbs distributions, which are defined as

$$P_{\beta R}(\pi) = \frac{P(\pi)e^{\beta R(\pi)}}{\mathbb{E}_{\pi \sim P}\left[e^{\beta R(\pi)}\right]}, \quad P_{\beta r^{\text{IS}}}(\pi) = \frac{P(\pi)e^{\beta r^{\text{IS}}(\pi, D_T)}}{\mathbb{E}_{\pi \sim P}\left[e^{\beta r^{\text{IS}}(\pi, D_T)}\right]}.$$

Proof of Theorem 3.25. As an intermediate step in the proof of the original PAC-Bayes Bernstein bound, for all $\lambda \in [-T\epsilon_T, T\epsilon_T]$, we have that

$$\mathop{\mathbb{E}}_{D_T}\left[\exp\left(\sup_{Q \in \mathcal{P}(\Pi)}\left\{\lambda r^{\text{IS}}(Q, D_T) - \lambda R(Q) - D_{\text{KL}}(Q||P_{\beta R})\right\}\right)\right] \leq \exp\left(\frac{\lambda^2(e-2)}{T\epsilon_T}\right). \tag{A.3}$$

Next, we attempt to find a relationship between $D_{\text{KL}}(Q||P_{\beta R})$ and $D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}})$.

$$\begin{aligned}
D_{\text{KL}}(Q||P_{\beta R}) &= \mathop{\mathbb{E}}_{\pi \sim Q}\left[\ln\left(\frac{Q(\pi)}{P_{\beta R}(\pi)}\right)\right], \\
&= \mathop{\mathbb{E}}_{\pi \sim Q}\left[\ln\left(\frac{Q(\pi)}{P_{\beta r^{\text{IS}}}(\pi)}\frac{P_{\beta r^{\text{IS}}}(\pi)}{P_{\beta R}(\pi)}\right)\right], \\
&= D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}}) + \beta \mathop{\mathbb{E}}_{\pi \sim Q}\left[r^{\text{IS}}(\pi, D_T) - R(\pi)\right] + \ln\left(\mathop{\mathbb{E}}_{\pi \sim P}\left[e^{\beta R(\pi)}\right]\right) - \ln\left(\mathop{\mathbb{E}}_{\pi \sim P}\left[e^{\beta r^{\text{IS}}(\pi, D_T)}\right]\right).
\end{aligned}$$

The last thing we need is a bound on $\ln(\mathbb{E}_{\pi \sim P}[e^{\beta R(\pi)}]) - \ln(\mathbb{E}_{\pi \sim P}[e^{\beta r^{\text{IS}}(\pi, D_T)}])$. Using two applications of the Donsker-Varadhan change of measure inequality (Lemma 2.2), we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{D_T}\left[\exp\left(\ln\left(\mathop{\mathbb{E}}_{\pi \sim P}\left[e^{\beta R(\pi)}\right]\right) - \ln\left(\mathop{\mathbb{E}}_{\pi \sim P}\left[e^{\beta r^{\text{IS}}(\pi, D_T)}\right]\right)\right)\right] \\
&\quad = \mathop{\mathbb{E}}_{D_T}\left[\exp\left(\ln\left(\mathop{\mathbb{E}}_{\pi \sim P}\left[e^{\beta R(\pi)}\right]\right) + \inf_{Q \in \mathcal{P}(\Pi)}\left\{-\beta r^{\text{IS}}(Q, D_T) + D_{\text{KL}}(Q||P)\right\}\right)\right], \\
&\quad \leq \mathop{\mathbb{E}}_{D_T}\left[\exp\left(\ln\left(\mathop{\mathbb{E}}_{\pi \sim P}\left[e^{\beta R(\pi)}\right]\right) - \beta r^{\text{IS}}(P_{\beta R}, D_T) + D_{\text{KL}}(P_{\beta R}||P)\right)\right], \\
&\quad = \mathop{\mathbb{E}}_{D_T}\left[\exp\left(\beta R(P_{\beta R}) - D_{\text{KL}}(P_{\beta R}||P) - \beta r^{\text{IS}}(P_{\beta R}, D_T) + D_{\text{KL}}(P_{\beta R}||P)\right)\right], \\
&\quad = \mathop{\mathbb{E}}_{D_T}\left[\exp\left(\beta R(P_{\beta R}) - \beta r^{\text{IS}}(P_{\beta R}, D_T)\right)\right], \\
&\quad \leq \exp\left(\frac{\beta^2(e-2)}{T\epsilon_T}\right).
\end{aligned}$$

The final inequality is obtained by using Equation A.3 with $\lambda = -\beta$. This inequality and the one in Equation A.3 can be combined using the Cauchy-Schwarz inequality

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{D_T}\left[\exp\left(\frac{1}{2}\sup_{Q\in\mathcal{P}(\Pi)}\left\{(\lambda-\beta)(r^{\mathrm{IS}}(Q,D_T)-R(Q))-D_{\mathrm{KL}}(Q||P_{\beta r^{\mathrm{IS}}})\right\}\right)\right]\\
&=\mathop{\mathbb{E}}_{D_T}\left[\exp\left(\frac{1}{2}\sup_{Q\in\mathcal{P}(\Pi)}\left\{\lambda(r^{\mathrm{IS}}(Q,D_T)-R(Q))-D_{\mathrm{KL}}(Q||P_{\beta R})\right\}\right)\right.\\
&\left.\times\exp\left(\frac{1}{2}\left[\ln\left(\mathop{\mathbb{E}}_{\pi\sim P}\left[e^{\beta R(\pi)}\right]\right)-\ln\left(\mathop{\mathbb{E}}_{\pi\sim P}\left[e^{\beta r^{\mathrm{IS}}(\pi,D_T)}\right]\right)\right]\right)\right],\\
&\leq\mathop{\mathbb{E}}_{D_T}\left[\exp\left(\sup_{Q\in\mathcal{P}(\Pi)}\left\{\lambda(r^{\mathrm{IS}}(Q,D_T)-R(Q))-D_{\mathrm{KL}}(Q||P_{\beta R})\right\}\right)\right]^{1/2}\\
&\times\mathop{\mathbb{E}}_{D_T}\left[\exp\left(\ln\left(\mathop{\mathbb{E}}_{\pi\sim P}\left[e^{\beta R(\pi)}\right]\right)-\ln\left(\mathop{\mathbb{E}}_{\pi\sim P}\left[e^{\beta r^{\mathrm{IS}}(\pi,D_T)}\right]\right)\right)\right]^{1/2},\\
&\leq\exp\left(\frac{\lambda^2(e-2)}{T\epsilon_T}\right)^{1/2}\exp\left(\frac{\beta^2(e-2)}{T\epsilon_T}\right)^{1/2},\\
&=\exp\left(\frac{(\lambda^2+\beta^2)(e-2)}{2T\epsilon_T}\right).
\end{aligned}
$$

We use Markov's inequality and then rearrange the result to obtain the statement of the Theorem. □

A.2 Further Information About The Experiments

A.2.1 Details About Classification Data Sets

The four data sets (see Table A.1) came from either OpenML (OptDigits, PenDigits and Chars) or the UCI Machine Learning Repository (DriveDiag). In the UCI Repository, the DriveDiag data set can be found under its full name: "Dataset for Sensorless Drive Diagnosis Data Set".

Name	OptDigits	PenDigits	Chars	DriveDiag
OpenML ID	28	32	1459	n/a
Size (T)	4496	8793	8174	46807
Input dim (d)	64	16	7	48
Classes (K)	10	10	10	11

Table A.1: The OpenML ID number, size, input dimensionality and number of classes for all the data sets we use in the CB Classification benchmark.

In Table A.1, the reported size of each data set is approximately 80% of the size of the original data set. At the start of each repetition of each experiment, we perform a random 80:20 split. We use 80% of the data to generate the training data for the offline bandit problem, learn a policy and then evaluate the reward bound. The remaining 20% of the data are used to estimate the expected reward of the learned policy. Therefore, the reported data set size reflects the number of examples used to learn a policy and evaluate a lower bound on the expected reward.

A.2.2 Details About Bound Optimisation and Evaluation

In the MAB benchmark, we allow Q to be any distribution over $\mathcal{A}$, so each Q is an element of the standard K-simplex. Unless stated otherwise, P is a uniform prior. For bounds where the optimal Q is a Gibbs posterior (Hoeffding-Azuma and Bernstein), we use the optimal Gibbs posterior. For bounds where we do not have a closed-form expression for the optimal Q (Pinsker and kl^{-1}), we optimise them by gradient ascent. As shown by Reeb et al. [143], the derivatives of kl^{-1} can be calculated by differentiating the identity $kl(p||kl^{-1}(p, B)) = B$.

In the CB benchmark, we restrict Q to be a diagonal Gaussian distribution $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\sigma}^2\boldsymbol{I})$, where $\boldsymbol{m}$ and $\boldsymbol{\sigma}$ are $d \times K$-dimensional vectors, over the weight matrix of the linear softmax policy. Unless stated otherwise, P is a standard Gaussian prior over the weights. We optimise each bound with respect to Q by stochastic gradient ascent, using the local reparameterisation trick [94] to calculate stochastic gradients. For the bounds that are linear in the reward estimate (e.g. $\mathbb{E}_{\theta\sim\mathcal{N}(\boldsymbol{m},\boldsymbol{\sigma}^2\boldsymbol{I})}[r^{\mathrm{IS}}(\pi_\theta, D_T)]$), this procedure will converge to the mean and variance parameters that maximise the bound. However, when we approximate $\mathbb{E}_{\theta\sim\mathcal{N}(\boldsymbol{m},\boldsymbol{\sigma}^2\boldsymbol{I})}[r^{\mathrm{IS}}(\pi_\theta, D_T)]$ with a single sample in the kl^{-1} bound, this procedure will converge to the mean and variance parameters that maximise

$$\frac{1}{\epsilon_T} \mathop{\mathbb{E}}_{\theta\sim\mathcal{N}(\boldsymbol{m},\boldsymbol{\sigma}^2\boldsymbol{I})} \left[kl^{-1}\left(\epsilon_T r^{\mathrm{IS}}(\pi_\theta, D_T), \frac{D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{m}, \boldsymbol{\sigma}^2\boldsymbol{I})||P) + \ln(2\sqrt{T}/\delta)}{T} \right) \right],$$

which may not be the mean and variance parameters that maximise the kl^{-1} bound. Nevertheless, the resulting approximately optimal Gaussian posterior still results in a valid bound. In all our experiments, this approximation appeared to work well.

For the PAC-Bayes Hoeffding-Azuma and Bernstein bounds, we set λ to the (data-independent) value that would be optimal if $D_{\mathrm{KL}}(Q||P) = 0$. In the MAB benchmark we can calculate $\mathbb{E}_{\pi\sim Q}[r^{\mathrm{IS}}(\pi, D_T)]$ exactly. In the CB benchmark, when evaluating the bound value, we approximate $\mathbb{E}_{\pi\sim Q}[r^{\mathrm{IS}}(\pi, D_T)]$ by averaging over 100 samples from Q.

A.2.3 Details About Implementation of the Priors

We tested the sample splitting prior with a value of $m = T/2$, meaning half the data were used to learn a prior and the other half were used to optimise and evaluate a PAC-Bayes bound. To learn priors from the subset $D_{1:m}$ of the data, we first split $D_{1:m}$ into training data D_{tr} and validation data D_{val}. In the MAB benchmark, we used D_{tr} to calculate an empirical Gibbs prior

$$P_{\beta r^{\mathrm{IS}}}(\pi) \propto P(\pi)\exp(\beta\sqrt{T_{\mathrm{tr}}} r^{\mathrm{IS}}(\pi, D_{\mathrm{tr}})),$$

for each β in a grid. We selected the value of β where $r^{\mathrm{IS}}(P_{\beta r^{\mathrm{IS}}}, D_{\mathrm{val}})$ was the greatest and then calculated a final empirical Gibbs prior with this β, and using all the data in $D_{1:m}$. P was a uniform prior and we used the grid $\beta \in \{1, 5, 10\}$.

In the CB Linear benchmark, we followed the same procedure, but with some small modifications. We approximated $P_{\beta r^{\mathrm{IS}}}$ with a diagonal Gaussian for each β in the grid $\{10, 100, 1000\}$. P was a standard Gaussian prior over the weights of the linear softmax policy.

When using differentially private priors in the MAB benchmark, we used priors of the form $P_{\boldsymbol{w}}(a) = \exp(\boldsymbol{w}_a)/\exp(\sum_{a'=1}^{K} \boldsymbol{w}_{a'})$, parameterised by $\boldsymbol{w} \in \mathbb{R}^K$. $\boldsymbol{w}_a$ is the ath element of $\boldsymbol{w}$. We

used Preconditioned Stochastic Gradient Langevin Dynamics (PSGLD) [110] to draw $\boldsymbol{w}$ from the Gibbs distribution with density proportional to $p(\boldsymbol{w})\exp(\eta\mathbb{E}_{a\sim P_{\boldsymbol{w}}}[r^{\text{IS}}(a, D_T)])$. $p(\boldsymbol{w})$ was a standard Gaussian. Due to Corollary 5.2 of [61], $P_{\boldsymbol{w}}$ is $2\eta/(T\epsilon_T)$-differentially private with this $\boldsymbol{w}$.

In the CB Linear benchmark, we learned Gaussian priors $P_{\boldsymbol{w}}(\theta) = \mathcal{N}(\boldsymbol{w}, \boldsymbol{I})$ over the weight matrix θ of the linear softmax policy π_θ. We used PSGLD to draw $\boldsymbol{w}$ from the distribution with density proportional to $p(\boldsymbol{w})\exp(\eta r^{\text{IS}}(\pi_{\boldsymbol{w}}, D_T))$. $p(\boldsymbol{w})$ was a standard Gaussian, and $P_{\boldsymbol{w}}$ is $2\eta/(T\epsilon_T)$-differentially private with this choice of $\boldsymbol{w}$.

In both benchmarks we drew $\boldsymbol{w}$'s from Gibbs distributions with $\eta \in \{0.1\sqrt{T\epsilon_T}, 0.5\sqrt{T\epsilon_T}, \sqrt{T\epsilon_T}\}$ and used the one that gave the best bound value, which we justify with the union bound.

To evaluate the kl^{-1} Lever bound and the distribution stability bound we need to calculate or sample from the Gibbs distribution $P_{\beta r^{\text{IS}}}$. In the MAB benchmark, we can calculate $P_{\beta r^{\text{IS}}}$ in closed-form. In the CB benchmark, we drew samples from $P_{\beta r^{\text{IS}}}$ using PSGLD and approximated $\mathbb{E}_{\pi\sim P_{\beta r^{\text{IS}}}}[r^{\text{IS}}(\pi, D_T)]$ by averages over 100 samples from $P_{\beta r^{\text{IS}}}$. In both the MAB and CB benchmarks, and for both bounds, we evaluated the bounds for several Gibbs posteriors with $\beta \in \{0.1\sqrt{T\epsilon_T}, 0.5\sqrt{T\epsilon_T}, \sqrt{T\epsilon_T}\}$ and used the ones that gave the best bound values.

Using the Donsker-Varadhan change of measure inequality in Lemma 2.2, the optimal posterior for the localised PAC-Bayes Bernstein bound is the Gibbs distribution $P_{\lambda r^{\text{IS}}}$, regardless of the value of β. We evaluated the bound at $\lambda \in \{\tilde{\lambda}, 1.5\tilde{\lambda}, 2.0\tilde{\lambda}\}$, where $\tilde{\lambda} = \sqrt{2T\epsilon_T \ln(1/\delta)/(e-2)}$ is the optimal value of λ when $\beta = 0$ and $D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}}) = 0$. For each λ, we evaluated the bound with $\beta \in \{0, \lambda/4, \lambda/2\}$. We used the (λ, β) pair that gave the best bound value. The optimal posterior for the PAC-Bayes Hoeffding-Azuma bound with the empirical Gibbs prior is also the Gibbs distribution $P_{\lambda r^{\text{IS}}}$. We evaluated this bound with $\lambda = \sqrt{2T(\epsilon_T^2 \ln(1/\delta) + 2)}$ and $\beta \in \{0, \lambda/4, \lambda/2\}$. This value of λ is approximately the optimal value when $D_{\text{KL}}(Q||P_{\beta r^{\text{IS}}}) = 0$.

A.2.4 Description of the TPOEM and TL2 Baselines

Like the original POEM algorithm [169], TPOEM uses the sample variance of the CIS reward estimate to regularise the policy selection. Its objective function is

$$r^{\text{CIS}}(\pi_\theta, D_T) - \beta\sqrt{\frac{v^{\text{CIS}}(\pi_\theta, D_T)}{T}}, \tag{A.4}$$

where

$$v^{\text{CIS}}(\pi_\theta, D_T) = \frac{1}{T-1}\sum_{t=1}^{T}\left(\min\left(\frac{\pi_\theta(a_t|s_t)}{b(a_t|s_t)}, \frac{1}{\tau}\right) r_t - r^{\text{CIS}}(\pi_\theta, D_T)\right)^2,$$

is the sample variance of the CIS estimate. We split the data set into training data D_{tr} and validation data D_{val} such that D_{tr} contains four times the number of samples in D_{val}. We maximise Equation A.4 with respect to the weights θ using the training data, and for each $\beta \in \{10^{-k} | k \in \{0, \ldots, 5\}\}$. This gives us a set of policies Π_Θ^β with 6 elements (one for each β). Using the validation data, we evaluate the following bound which is essentially a simpler version of the original POEM bound that only holds for finite policy classes.

$$r^{\mathrm{CIS}}(\pi_\theta, D_{\mathrm{val}}) - \sqrt{\frac{2v^{\mathrm{CIS}}(\pi_\theta, D_{\mathrm{val}})\ln(2|\Pi_\Theta^\beta|/\delta)}{T_{\mathrm{val}}}} - \frac{7\ln(2|\Pi_\Theta^\beta|/\delta)}{\tau(T_{\mathrm{val}}-1)}, \tag{A.5}$$

where $T_{\mathrm{val}} = |D_{\mathrm{val}}|$ is the number of examples in the validation data set. We choose the policy $\pi_\theta \in \Pi_\Theta^\beta$ that maximises the bound in Equation (A.5). The TL2 baseline uses the ℓ_2 norm of the neural network weights to regularise the policy selection. It uses the objective function

$$r^{\mathrm{CIS}}(\pi_\theta, D_T) - \beta \left\|\theta\right\|_2^2, \tag{A.6}$$

We split the data set into D_{tr} and D_{val} with the same relative sizes as with TPOEM. We maximise Equation A.6 with respect to θ using $D_T = D_{\mathrm{tr}}$ and for each $\beta \in \{10^{-k} | k \in \{1, \ldots, 6\}\}$. This gives us a set of policies Π_Θ^β with 6 elements. Using D_{val}, we evaluate a PAC bound based on the Hoeffding-Azuma inequality.

$$r^{\mathrm{CIS}}(\pi_\theta, D_{\mathrm{val}}) - \frac{1}{\tau}\sqrt{\frac{\ln(|\Pi_\Theta^\beta|/\delta)}{2T_{\mathrm{val}}}}, \tag{A.7}$$

We choose the policy $\pi_\theta \in \Pi_\Theta^\beta$ that maximises the bound in Equation (A.7).

A.3 Additional Experiments

A.3.1 Experiments With The Efron-Stein PAC-Bayes Bound

We compare the Efron-Stein (ES) PAC-Bayes bound for the r^{WIS} estimate (in Equation (3.19)) against the Hoeffding-Azuma (Theorem 3.2), kl^{-1} (Equation (3.6)), Pinsker (Equation (3.7)), and Bernstein (Equation (3.8)) PAC-Bayes bounds for the r^{IS} estimate.

We compare the bounds in the offline MAB Binary benchmark, in which the policy class is the set of all deterministic policies (i.e. the set of actions). As in our experiments in Section 3.7.3, we optimise each bound with respect to the posterior Q and then report the value of the bound and the expected reward for this Q. Details about how we optimise the bounds for the r^{IS} estimate with respect to Q and then evaluate them can be found in Appendix A.2.2. For convenience, we re-state the RHS of the ES PAC-Bayes bound for the r^{WIS} estimate from Equation (3.19).

$$r^{\mathrm{WIS}}(Q, D_T) - \left|R^{\mathrm{WIS}}(Q) - R(Q)\right| - \sqrt{2\left(y + 2V^{\mathrm{WIS}}(Q, D_T)\right)\left(D_{\mathrm{KL}}(Q||P) + \frac{1}{2}\ln\left(1 + \frac{2V^{\mathrm{WIS}}(Q, D_T)}{y}\right) + \ln(1/\delta)\right)}.$$

Also, recall that $V^{\mathrm{WIS}}(\pi, D_T)$ was defined as

$$V^{\mathrm{WIS}}(\pi, D_T) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{D_T, D_T'}\left[\tilde{w}_{\pi,t}^2 + \tilde{u}_{\pi,t}^2 | D_t\right],$$

where

$$\tilde{w}_{\pi,t} = \frac{\frac{\pi(a_t)}{b(a_t)}}{\sum_{k=1}^{T} \frac{\pi(a_k)}{b(a_k)}}, \qquad \tilde{u}_{\pi,t} = \frac{\frac{\pi(a'_t)}{b(a'_t)}}{\frac{\pi(a'_t)}{b(a'_t)} + \sum_{k \neq t} \frac{\pi(a_k)}{b(a_k)}}.$$

a'_t is an independently sampled copy of a_t. To evaluate the ES PAC-Bayes bound, we first assume that the bias $\left|R^{\mathrm{WIS}}(Q) - R(Q)\right|$ is always equal to 0. Since the reward distribution $\mathbb{P}_R$ in the MAB benchmark is actually known, we can estimate $\left|R^{\mathrm{WIS}}(Q) - R(Q)\right|$ to arbitrary accuracy to check whether this is a reasonable assumption. Some rough estimates suggest that in the MAB benchmark with $T = 1000$, the bias of the WIS estimate is approximately 10^{-9} or smaller.

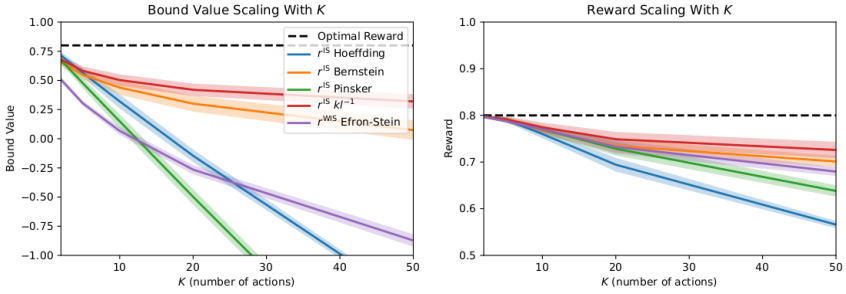


Figure A.1: The bound value (left) and expected reward (right) for the Efron-Stein WIS bound and each of the IS bounds in the MAB Binary benchmark. The number of actions K varies from 2 to 50 along the x axes.

Next, we replace the semi-empirical ES variance proxy $V^{\mathrm{WIS}}(\pi, D_T)$ with a fully empirical estimate. For t in $\{1, \ldots, T\}$, we draw $m = 1000$ actions $\{a'_{tk}\}_{k=1}^{m}$ and another m actions $\{a''_{tk}\}_{k=1}^{m}$ from the behaviour policy b. $\{a'_{tk}\}_{k=1}^{m}$ for $t = 1, \ldots, T$ are 1000 draws of the ghost sample D'_T and $\{a''_{tk}\}_{k=1}^{m}$ for $t = 1, \ldots, T$ are 1000 re-draws of the original sample D_T. Then for each policy $\pi \in \Pi$, we calculate

$$\widehat{w}_{\pi,t,k} = \frac{\frac{\pi(a_t)}{b(a_t)}}{\sum_{l=1}^{t} \frac{\pi(a_l)}{b(a_l)} + \sum_{l=t+1}^{T} \frac{\pi(a''_{lk})}{b(a''_{lk})}}, \quad \widehat{u}_{\pi,t,k} = \frac{\frac{\pi(a'_{tk})}{b(a'_{tk})}}{\frac{\pi(a'_{tk})}{b(a'_{tk})} + \sum_{l=1}^{t-1} \frac{\pi(a_l)}{b(a_l)} + \sum_{l=t+1}^{T} \frac{\pi(a''_{lk})}{b(a''_{lk})}}.$$

In the ES PAC-Bayes bound, we replace $V^{\mathrm{WIS}}(\pi, D_T)$ with the estimate:

$$\widehat{V}^{\mathrm{WIS}}(\pi, D_T) = \sum_{t=1}^{T} \frac{1}{m} \sum_{k=1}^{m} \widehat{w}^2_{\pi,t,k} + \widehat{u}^2_{\pi,t,k}$$

Note that we only have to calculate $\widehat{V}^{\mathrm{WIS}}(\pi, D_T)$ once before we optimise the ES PAC-Bayes bound with respect to Q. Since the policy class Π is finite with K elements, calculating $\widehat{V}^{\mathrm{WIS}}(\pi, D_T)$ for every $\pi \in \Pi$ is possible. However, this would obviously not be possible for infinite policy classes. Strictly speaking, we should replace $V^{\mathrm{WIS}}(\pi, D_T)$ with an upper bound rather than an estimate to obtain a valid bound, as is done in [98]. Using an estimate rather than an upper bound results in a (slightly) favourable evaluation of the ES PAC-Bayes bound.

We always use a data set of size $T = 1000$, and the data set is generated using a uniform behaviour policy. We varied the number of actions K from 2 to 50 to investigate how the bounds compare in MAB problems with different numbers of actions.

Figure A.1 shows the bound value (left) and the expected reward (right) for each of the bounds we compared at each K. In the left plot in Figure A.1, we observe that the value of the Efron-Stein WIS bound is the lowest for $K \leq 10$. As K increases above 10, the Efron-Stein WIS bound overtakes both the IS Pinsker and IS Hoeffding-Azuma bounds. However, for $K \geq 10$, the Efron-Stein WIS bound is vacuous (i.e. less than 0). On the bright side, the Efron-Stein WIS bound appears to work quite well as a learning objective. In the right plot of Figure A.1, we see that the policy learned by maximising the Efron-Stein WIS bound achieves close to the highest expected reward.

A.3.2 Insights About Choosing Bound Parameters

Now, we briefly explore why the fixed value of λ was almost as good as the optimal value. Both the PAC-Bayes Hoeffding-Azuma and PAC-Bayes Bernstein bounds can be written in the form

$$R(Q) \geq r(Q, D_T) - a\lambda - \frac{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}{\lambda}.$$

For bounds of this form, the optimal λ is the one that minimises $f(\lambda) = a\lambda + (D_{\mathrm{KL}}(Q||P) + \ln(1/\delta))/\lambda$. One can verify that $\lambda^* = \sqrt{(D_{\mathrm{KL}}(Q||P) + \ln(1/\delta))/a}$. We found that the value of the PAC-Bayes Bernstein bound at $\hat{\lambda} = \sqrt{\ln(1/\delta)/a}$ was almost the same as the bound value at λ^*. It can be shown that the second derivative of f evaluated at λ^* is

$$f''(\lambda^*) = 2a^{3/2}/\sqrt{D_{\mathrm{KL}}(Q||P) + \ln(1/\delta)}.$$

When a is close to 0, $f''(\lambda^*)$ will also be close to 0. Therefore, we can expect $f(\lambda)$ to be almost constant in the neighbourhood of λ^* when a is near 0. For the PAC-Bayes Bernstein r^{IS} bound, $a = (e-2)/T\epsilon_T$. In the MAB Binary benchmark considered in Section 3.7.3, we had $T = 1000$ and $\epsilon_T = 0.1$, so $a \approx 0.000718$. This may explain why the Bernstein bound value at $\hat{\lambda}$ was close to the Bernstein bound value at λ^*. In Figure A.2, we plot $f(\lambda)$ for the PAC-Bayes Bernstein bound. We set $n \in \{100, 1000, 10000\}$, and to match our earlier experiment in the MAB Binary benchmark, we set $\epsilon_T = 0.1$, $\delta = 0.05$ and $D_{\mathrm{KL}}(Q||P) = \ln(K)$, with $K = 10$. This is the maximum value of the KL divergence, which means the difference between $\hat{\lambda}$ and λ^* is maximised.

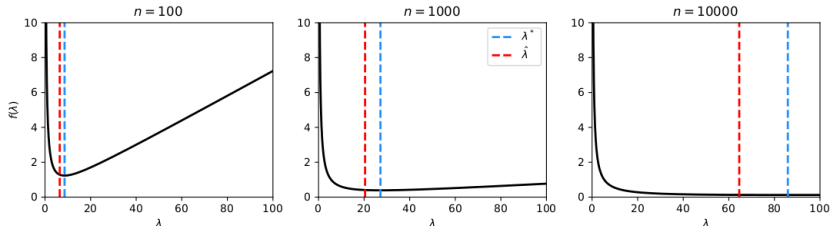


Figure A.2: $f(\lambda)$ for $a = (e-2)/(T\epsilon_T)$ with $\epsilon_T = 0.1$, $\delta = 0.05$ and $D_{\mathrm{KL}}(Q||P) = \ln(K)$. T is equal to 100 (left), 1000 (middle) and 10000 (right).

In Figure A.2, we see that as T increases (and a decreases) $f(\lambda)$ becomes almost constant in the neighbourhood of λ^*. The value of $f(\lambda)$ at $\hat{\lambda}$ and λ^* is almost the same even for $T = 100$.

Appendix B

Appendix for Chapter 4

B.1 Proof of the General-Purpose Tail Bound for Adaptive Martingale Mixtures

B.1.1 Verifying Martingale Properties

First, we recall the definition of $(M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)|t \in \mathbb{N})$ in Eq. (4.3). We are given a filtration $(\mathcal{D}_t|t \in \mathbb{N})$, a sequence of adapted random functions $(Z_t : \mathbb{R} \to \mathbb{R}|t \in \mathbb{N})$, and a sequence of predictable random variables $(\lambda_t|t \in \mathbb{N})$.

A filtration is an increasing sequence of σ-algebras $\mathcal{D}_0 \subseteq \mathcal{D}_1 \subseteq \mathcal{D}_2 \cdots$. Each σ-algebra $\mathcal{D}_t$ represents the information available at time t. $(Z_t : \mathbb{R} \to \mathbb{R}|t \in \mathbb{N})$ being a sequence of *adapted* (to the filtration $(\mathcal{D}_t|t \in \mathbb{N})$) functions means that, when conditioned on $\mathcal{D}_t$, Z_t is no longer random. $(\lambda_t|t \in \mathbb{N})$ being a sequence of *predictable* random variables means that, when conditioned on $\mathcal{D}_{t-1}$, λ_t is no longer random.

For a sequence of real numbers $(g_t : t \in \mathbb{N})$, we define

$$M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t) = \exp\left(\sum_{k=1}^{t} \lambda_k Z_k(g_k) - \psi_k(g_k, \lambda_k)\right),$$

where $\psi_t(g_t, \lambda_t)$ is the conditional cumulant generating function

$$\psi_t(g_t, \lambda_t) := \ln\left(\mathbb{E}\left[\exp(\lambda_t Z_t(g_t))|\mathcal{D}_{t-1}\right]\right).$$

Lemma B.1. *For any sequence of real numbers* $(g_t|t \in \mathbb{N})$, $(M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)|t \in \mathbb{N})$ *is a martingale and* $\mathbb{E}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)] = 1$ *for all* $t \in \mathbb{N}$.

Proof. For $t = 1$, we have

$$\begin{aligned}\mathbb{E}[M_1(\boldsymbol{g}_1, \boldsymbol{\lambda}_1)|\mathcal{D}_0] &= \mathbb{E}\left[\exp(\lambda_1 Z_1(g_1) - \psi_1(g_1, \lambda_1))|\mathcal{D}_0\right] \\ &= \mathbb{E}\left[\exp(\lambda_1 Z_1(g_1))|\mathcal{H}_0\right] / \exp(\psi_1(g_1, \lambda_1)) \\ &= \exp(\psi_1(g_1, \lambda_1)) / \exp(\psi_1(g_1, \lambda_1)) \\ &= 1.\end{aligned}$$

Using the tower rule of conditional expectation, we also have

$$\mathbb{E}[M_1(\boldsymbol{g}_1, \boldsymbol{\lambda}_1)] = \mathbb{E}[\mathbb{E}[M_1(\boldsymbol{g}_1, \boldsymbol{\lambda}_1)|\mathcal{D}_0]] = 1.$$

Now, we verify the martingale property. For any $t \geq 2$, we have

$$\begin{aligned}\mathbb{E}\left[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)|\mathcal{D}_{t-1}\right] &= \mathbb{E}\left[\exp\left(\sum_{k=1}^{t} \lambda_k Z_k(g_k) - \psi_k(g_k, \lambda_k)\right) \middle| \mathcal{D}_{t-1}\right] \\ &= \exp\left(\sum_{k=1}^{t-1} \lambda_k Z_k(g_k) - \psi_k(g_k, \lambda_k)\right) \mathbb{E}\left[\exp\left(\lambda_t Z_t(g_t) - \psi_t(g_t, \lambda_t)\right)|\mathcal{D}_{t-1}\right] \\ &= \exp\left(\sum_{k=1}^{t-1} \lambda_k Z_k(g_k) - \psi_k(g_k, \lambda_k)\right) \\ &= M_{t-1}(\boldsymbol{g}_{t-1}, \boldsymbol{\lambda}_{t-1}).\end{aligned}$$

Using the tower rule again, we have for any $t \geq 2$

$$\mathbb{E}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)] = \mathbb{E}[\mathbb{E}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)|\mathcal{D}_{t-1}]] = \mathbb{E}[M_{t-1}(\boldsymbol{g}_{t-1}, \boldsymbol{\lambda}_{t-1})].$$

Therefore, we have

$$\mathbb{E}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)] = \mathbb{E}[M_{t-1}(\boldsymbol{g}_{t-1}, \boldsymbol{\lambda}_{t-1})] = \cdots = \mathbb{E}[M_1(\boldsymbol{g}_1, \boldsymbol{\lambda}_1)] = 1.$$

□

Lemma B.2. *For any adaptive sequence of mixture distributions* $(P_t|t \in \mathbb{N})$, $(\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]|t \in \mathbb{N})$ *is a martingale and* $\mathbb{E}[\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]] = 1$ *for all* $t \in \mathbb{N}$.

Proof. For any $t \geq 1$, since $M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)$ is non-negative and P_t is $\mathcal{D}_{t-1}$-measurable, Tonelli's theorem implies

$$\mathbb{E}\left[\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]|\mathcal{D}_{t-1}\right] = \mathbb{E}_{\boldsymbol{g}_t \sim P_t}\left[\mathbb{E}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)|\mathcal{H}_{t-1}]\right].$$

The requirement that the distributions $P_1, P_2, \ldots$ have coinciding marginals, i.e. $\int P_t(\boldsymbol{g}_t)\mathrm{d}g_t = P_{t-1}(\boldsymbol{g}_{t-1})$, means that for all $t \geq 2$

$$\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_{t-1}(\boldsymbol{g}_{t-1}, \boldsymbol{\lambda}_{t-1})] = \mathbb{E}_{\boldsymbol{g}_{t-1} \sim P_{t-1}}[M_{t-1}(\boldsymbol{g}_{t-1}, \boldsymbol{\lambda}_{t-1})].$$

Using these two results, and the fact that $(M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)|t \in \mathbb{N})$ is a martingale for any sequence $(g_t|t \in \mathbb{N})$, we now verify that the martingale mixture $(\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]|t \in \mathbb{N})$ is a martingale with expected value 1. For $t = 1$, we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\boldsymbol{g}_1 \sim P_1}[M_1(\boldsymbol{g}_1, \boldsymbol{\lambda}_1)]|\mathcal{D}_0\right] &= \mathbb{E}_{\boldsymbol{g}_1 \sim P_1}\left[\mathbb{E}[M_1(\boldsymbol{g}_1, \boldsymbol{\lambda}_1)|\mathcal{D}_0]\right] \\ &= \mathbb{E}_{\boldsymbol{g}_1 \sim P_1}\left[1\right] \\ &= 1.\end{aligned}$$

Using the tower rule as before, this also means that $\mathbb{E}\left[\mathbb{E}_{\boldsymbol{g}_1 \sim P_1}[M_1(\boldsymbol{g}_1, \boldsymbol{\lambda}_1)]\right] = 1$. For any $t \geq 2$, we have

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]|\mathcal{D}_{t-1}\right] &= \mathbb{E}_{\boldsymbol{g}_t \sim P_t}\left[\mathbb{E}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)|\mathcal{D}_{t-1}]\right] \\ &= \mathbb{E}_{\boldsymbol{g}_t \sim P_t}\left[M_{t-1}(\boldsymbol{g}_{t-1}, \boldsymbol{\lambda}_{t-1})\right] \\ &= \mathbb{E}_{\boldsymbol{g}_{t-1} \sim P_{t-1}}\left[M_{t-1}(\boldsymbol{g}_{t-1}, \boldsymbol{\lambda}_{t-1})\right].\end{aligned}$$

Using the tower rule one more time, we have for any $t \geq 2$

$$\mathbb{E}[\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]] = \mathbb{E}[\mathbb{E}[\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]|\mathcal{D}_{t-1}]] = \mathbb{E}[\mathbb{E}_{\boldsymbol{g}_{t-1} \sim P_{t-1}}[M_{t-1}(\boldsymbol{g}_{t-1}, \boldsymbol{\lambda}_{t-1})]].$$

Therefore, we have

$$\mathbb{E}[\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]] = \mathbb{E}[\mathbb{E}_{\boldsymbol{g}_1 \sim P_1}[M_1(\boldsymbol{g}_1, \boldsymbol{\lambda}_1)]] = 1.$$

□

B.1.2 Proof of Theorem 4.1

To prove Thm. 4.1, we use Ville's inequality for non-negative (super)martingales [181] (see Lemma 2.1).

Proof of Thm. 4.1. We choose an arbitrary $\delta \in (0, 1]$. From Lemma B.2, for any adaptive sequence of mixture distributions $(P_t|t \in \mathbb{N})$, $(\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]|t \in \mathbb{N})$ is a martingale and $\mathbb{E}[\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]] = 1$. In addition, $(\mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)]|t \in \mathbb{N})$ is clearly non-negative. Therefore, using Lemma 2.1, with probability at least $1 - \delta$

$$\forall t \geq 1, \quad \mathbb{E}_{\boldsymbol{g}_t \sim P_t}[M_t(\boldsymbol{g}_t, \boldsymbol{\lambda}_t)] \leq 1/\delta.$$

Taking the logarithm of both sides yields the statement of Thm. 4.1. □

B.2 Closed-Form Gaussian Integration

Here, we calculate the integral in the inequality (see beginning of Sec. 4.5.2):

$$\mathop{\mathbb{E}}_{\boldsymbol{g}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)}\left[\exp\left\{\sum_{k=1}^{t} \lambda_k (g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)(r_k - \phi(a_k)^\top \boldsymbol{\theta}^*) - \frac{\sigma^2}{2}\lambda_k^2 (g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)^2\right\}\right] \leq \frac{1}{\delta}. \quad \text{(B.1)}$$

First, we rearrange the integrand into a more convenient form. For every k, using $r_k = \phi(a_k)^\top \boldsymbol{\theta}^* + \epsilon_k$, we have

$$\begin{aligned}(\phi(a_k)^\top \boldsymbol{\theta}^* - r_k)^2 - (g_k - r_k)^2 &= \epsilon_k^2 - (g_k - \phi(a_k)^\top \boldsymbol{\theta}^* - \epsilon_k)^2 \\ &= \epsilon_k^2 - (g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)^2 + 2(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)\epsilon_k - \epsilon_k^2 \\ &= 2(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)(r_k - \phi(a_k)^\top \boldsymbol{\theta}^*) - (g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)^2.\end{aligned}$$

Therefore, we have that

$$\begin{aligned}&\lambda_k(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)(r_k - \phi(a_k)^\top \boldsymbol{\theta}^*) - \frac{\sigma^2}{2}\lambda_k^2(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)^2 \\ &= \frac{\lambda_k}{2}(\phi(a_k)^\top \boldsymbol{\theta}^* - r_k)^2 - \frac{\lambda_k}{2}(g_k - r_k)^2 + \frac{1}{2}(\lambda_k - \sigma^2\lambda_k^2)(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)^2.\end{aligned}$$

Equation (B.1) can now be re-written as

$$\mathop{\mathbb{E}}_{\boldsymbol{g}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)} \left[\exp\left\{\sum_{k=1}^{t} \frac{\lambda_k}{2}(\phi(a_k)^\top \boldsymbol{\theta}^* - r_k)^2 - \frac{\lambda_k}{2}(g_k - r_k)^2 + \frac{1}{2}(\lambda_k - \sigma^2\lambda_k^2)(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)^2\right\}\right] \leq \frac{1}{\delta}. \tag{B.2}$$

In the special case where $\lambda_t \equiv 1/\sigma^2$, we have $\lambda_k - \sigma^2\lambda_k^2 = 0$, which means that $\frac{1}{2}(\lambda_k - \sigma^2\lambda_k^2)(g_k - \phi(a_k)^\top \boldsymbol{\theta}^*)^2$ disappears. In addition, and for any λ_k, $(\phi(a_k)^\top \boldsymbol{\theta}^* - r_k)^2$ does not depend on g_k, so it can be moved outside the integral.

B.2.1 General λ_t

Let $\boldsymbol{\Lambda}_t$ be the $t \times t$ diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \ldots, \lambda_t$. Starting from (B.2), taking the logarithm of both sides, rearranging terms and then writing everything in matrix notation, we arrive at

$$(\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t)\boldsymbol{\Lambda}_t(\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t) \leq 2\ln(1/\delta) \tag{B.3}$$

$$-2\ln\left(\mathop{\mathbb{E}}_{\boldsymbol{g}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)}\left[\exp\left(-\frac{1}{2}(\boldsymbol{g}_t - \boldsymbol{r}_t)^\top \boldsymbol{\Lambda}_t(\boldsymbol{g}_t - \boldsymbol{r}_t) + \frac{1}{2}(\boldsymbol{g}_t - \Phi_t\boldsymbol{\theta}^*)^\top \left(\boldsymbol{\Lambda}_t - \sigma^2\boldsymbol{\Lambda}_t^2\right)(\boldsymbol{g}_t - \Phi_t\boldsymbol{\theta}^*)\right)\right]\right)$$

The expected value inside the logarithm can be re-written as

$$\begin{aligned}\frac{1}{\sqrt{(2\pi)^t \det(\boldsymbol{T}_t)}} \int \exp\Bigg(&-\frac{1}{2}(\boldsymbol{g}_t - \boldsymbol{\mu}_t)^\top \boldsymbol{T}_t^{-1}(\boldsymbol{g}_t - \boldsymbol{\mu}_t) - \frac{1}{2}(\boldsymbol{g}_t - \boldsymbol{r}_t)^\top \boldsymbol{\Lambda}_t(\boldsymbol{g}_t - \boldsymbol{r}_t) \\ &+ \frac{1}{2}(\boldsymbol{g}_t - \Phi_t\boldsymbol{\theta}^*)^\top \left(\boldsymbol{\Lambda}_t - \sigma^2\boldsymbol{\Lambda}_t^2\right)(\boldsymbol{g}_t - \Phi_t\boldsymbol{\theta}^*)\Bigg)\mathrm{d}\boldsymbol{g}_t.\end{aligned} \tag{B.4}$$

We will calculate the integral by "completing the square", i.e. rewriting the exponent in the form $-\frac{1}{2}(\boldsymbol{g}_t - \boldsymbol{b})^\top \boldsymbol{A}(\boldsymbol{g}_t - \boldsymbol{b}) + c$, to recover the integral of a Gaussian density function. For a symmetric matrix $\boldsymbol{A}$, we have

$$-\frac{1}{2}(\boldsymbol{g}_t - \boldsymbol{b})^\top \boldsymbol{A}(\boldsymbol{g}_t - \boldsymbol{b}) + c = -\frac{1}{2}\boldsymbol{g}_t^\top \boldsymbol{A}\boldsymbol{g}_t + \boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{g}_t - \frac{1}{2}\boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{b} + c.$$

We also have

$$
\begin{aligned}
-\frac{1}{2}(\boldsymbol{g}_t-\boldsymbol{\mu}_t)^\top \boldsymbol{T}_t^{-1}(\boldsymbol{g}_t-\boldsymbol{\mu}_t) &= -\frac{1}{2}\boldsymbol{g}_t^\top \boldsymbol{T}_t^{-1}\boldsymbol{g}_t + \boldsymbol{\mu}_t^\top \boldsymbol{T}_t^{-1}\boldsymbol{g}_t - \frac{1}{2}\boldsymbol{\mu}_t^\top \boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t \\
-\frac{1}{2}(\boldsymbol{g}_t-\boldsymbol{r}_t)^\top \boldsymbol{\Lambda}_t(\boldsymbol{g}_t-\boldsymbol{r}_t) &= -\frac{1}{2}\boldsymbol{g}_t^\top \boldsymbol{\Lambda}_t\boldsymbol{g}_t + \boldsymbol{r}_t^\top \boldsymbol{\Lambda}_t\boldsymbol{g}_t - \frac{1}{2}\boldsymbol{r}_t^\top \boldsymbol{\Lambda}_t\boldsymbol{r}_t \\
\frac{1}{2}(\boldsymbol{g}_t-\Phi_t\boldsymbol{\theta}^*)^\top \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)(\boldsymbol{g}_t-\Phi_t\boldsymbol{\theta}^*) &= \frac{1}{2}\boldsymbol{g}_t^\top \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\boldsymbol{g}_t - \boldsymbol{\theta}^{*\top}\Phi_t^\top \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\boldsymbol{g}_t \\
&\quad + \frac{1}{2}\boldsymbol{\theta}^{*\top}\Phi_t^\top \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^*.
\end{aligned}
$$

We now equate coefficients to find $\boldsymbol{A}$, $\boldsymbol{b}$ and c. We find that $\boldsymbol{A}$ is

$$
\boldsymbol{A} = \boldsymbol{T}_t^{-1} + \sigma^2\boldsymbol{\Lambda}_t^2.
$$

Note that $\boldsymbol{A}$ is symmetric. We find that $\boldsymbol{b}$ is

$$
\begin{aligned}
\boldsymbol{b}^\top\boldsymbol{A} &= \boldsymbol{\mu}_t^\top \boldsymbol{T}_t^{-1} + \boldsymbol{r}_t^\top \boldsymbol{\Lambda}_t - \boldsymbol{\theta}^{*\top}\Phi_t^\top \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right) \\
\implies \boldsymbol{A}\boldsymbol{b} &= \boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t + \boldsymbol{\Lambda}_t\boldsymbol{r}_t - \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^* \\
\implies \boldsymbol{b} &= \left(\boldsymbol{T}_t^{-1}+\sigma^2\boldsymbol{\Lambda}_t^2\right)^{-1}\left(\boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t + \boldsymbol{\Lambda}_t\boldsymbol{r}_t - \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^*\right).
\end{aligned}
$$

Finally, we find that c is

$$
\begin{aligned}
c &= \frac{1}{2}\boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{b} - \frac{1}{2}\boldsymbol{\mu}_t^\top \boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t - \frac{1}{2}\boldsymbol{r}_t^\top \boldsymbol{\Lambda}_t\boldsymbol{r}_t + \frac{1}{2}\boldsymbol{\theta}^{*\top}\Phi_t^\top \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^* \\
&= \frac{1}{2}\left(\boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t + \boldsymbol{\Lambda}_t\boldsymbol{r}_t - \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^*\right)^\top \left(\boldsymbol{T}_t^{-1}+\sigma^2\boldsymbol{\Lambda}_t^2\right)^{-1}\left(\boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t + \boldsymbol{\Lambda}_t\boldsymbol{r}_t - \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^*\right) \\
&\quad - \frac{1}{2}\boldsymbol{\mu}_t^\top \boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t - \frac{1}{2}\boldsymbol{r}_t^\top \boldsymbol{\Lambda}_t\boldsymbol{r}_t + \frac{1}{2}\boldsymbol{\theta}^{*\top}\Phi_t^\top \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^*
\end{aligned}
$$

Now, we can rewrite and calculate the integral in (B.4) as

$$
\begin{aligned}
\frac{\exp(c)}{\sqrt{(2\pi)^t\det(\boldsymbol{T}_t)}}\int \exp\left(-\frac{1}{2}(\boldsymbol{g}_t-\boldsymbol{b})^\top\boldsymbol{A}(\boldsymbol{g}_t-\boldsymbol{b})\right)\mathrm{d}\boldsymbol{g}_t &= \frac{\exp(c)\sqrt{(2\pi)^t\det(\boldsymbol{A}^{-1})}}{\sqrt{(2\pi)^t\det(\boldsymbol{T}_t)}} \\
&= \exp(c)\sqrt{\frac{\det(\boldsymbol{A}^{-1})}{\det(\boldsymbol{T}_t)}}
\end{aligned}
$$

Substituting this into (B.3), we obtain the constraint

$$
\begin{aligned}
&(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)^\top\boldsymbol{\Lambda}_t(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t) \leq -2\ln\left(\exp(c)\sqrt{\frac{\det(\boldsymbol{A}^{-1})}{\det(\boldsymbol{T}_t)}}\right) + 2\ln(1/\delta) \\
&= -2c + \ln\left(\det(\boldsymbol{A}\boldsymbol{T}_t)\right) + 2\ln(1/\delta) \\
&= -\left(\boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t + \boldsymbol{\Lambda}_t\boldsymbol{r}_t - \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^*\right)^\top \left(\boldsymbol{T}_t^{-1}+\sigma^2\boldsymbol{\Lambda}_t^2\right)^{-1}\left(\boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t + \boldsymbol{\Lambda}_t\boldsymbol{r}_t - \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^*\right) \\
&+ \boldsymbol{\mu}_t^\top \boldsymbol{T}_t^{-1}\boldsymbol{\mu}_t + \boldsymbol{r}_t^\top \boldsymbol{\Lambda}_t\boldsymbol{r}_t - \boldsymbol{\theta}^{*\top}\Phi_t^\top \left(\boldsymbol{\Lambda}_t-\sigma^2\boldsymbol{\Lambda}_t^2\right)\Phi_t\boldsymbol{\theta}^* + \ln\left(\det(\boldsymbol{I}+\sigma^2\boldsymbol{\Lambda}_t^2\boldsymbol{T}_t)\right) + 2\ln(1/\delta)
\end{aligned}
$$

Note that $\boldsymbol{\theta}^*$ appears on both the left-hand-side and right-hand-side of this inequality. However, when $\boldsymbol{\Lambda}_t - \sigma^2\boldsymbol{\Lambda}_t^2$ is the zero matrix (e.g. when $\lambda_t \equiv 1/\sigma^2$), all the $\boldsymbol{\theta}^*$-dependent terms on the right-hand-side disappear.

B.2.2 The Special Case $\lambda_t \equiv 1/\sigma^2$

Starting from (B.2), choosing $\lambda_t \equiv 1/\sigma^2$, taking the logarithm of both sides and then rearranging terms, we arrive at

$$\|\Phi_t\boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 \leq -2\sigma^2 \ln\left(\mathop{\mathbb{E}}_{\boldsymbol{g}_t\sim\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{T}_t)}\left[\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{g}_t-\boldsymbol{r}_t)^\top(\boldsymbol{g}_t-\boldsymbol{r}_t)\right)\right]\right) + 2\sigma^2\ln(1/\delta). \quad \text{(B.5)}$$

For any $t \times t$ covariance matrix $\boldsymbol{T}$, let $Z(\boldsymbol{T})$ denote the normalising constant of a Gaussian distribution with covariance $\boldsymbol{T}$, so

$$Z(\boldsymbol{T}) = \sqrt{(2\pi)^t \det(\boldsymbol{T})}.$$

For any t-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{\mu}$, and any $t \times t$ covariance matrix $\boldsymbol{T}$, let $p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{T})$ denote the density function of a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{T}$, evaluated at $\boldsymbol{x}$. This means that

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{T}) = \frac{1}{Z(\boldsymbol{T})}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{T}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

We will use the product of Gaussians trick from [138] (Section 8.1.8, Equation 371), which states

$$p(\boldsymbol{x}|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)p(\boldsymbol{x}|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) = p(\boldsymbol{\mu}_1|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)p(\boldsymbol{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c), \quad \text{(B.6)}$$

where

$$\boldsymbol{\mu}_c = \left(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right), \qquad \boldsymbol{\Sigma}_c = \left(\boldsymbol{\Sigma}_1^{-1}+\boldsymbol{\Sigma}_2^{-1}\right)^{-1}.$$

We have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{g}_t\sim\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{T}_t)}\left[\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{g}_t-\boldsymbol{r}_t)^\top(\boldsymbol{g}_t-\boldsymbol{r}_t)\right)\right] &= \mathop{\mathbb{E}}_{\boldsymbol{g}_t\sim\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{T}_t)}\left[Z(\sigma^2\boldsymbol{I})p(\boldsymbol{g}_t|\boldsymbol{r}_t,\sigma^2\boldsymbol{I})\right] \\
&= Z(\sigma^2\boldsymbol{I})\int_{\mathbb{R}^t} p(\boldsymbol{g}_t|\boldsymbol{\mu}_t,\boldsymbol{T}_t)p(\boldsymbol{g}_t|\boldsymbol{r}_t,\sigma^2\boldsymbol{I})\mathrm{d}\boldsymbol{f}_t \\
&= Z(\sigma^2\boldsymbol{I})\int_{\mathbb{R}^t} p(\boldsymbol{\mu}_t|\boldsymbol{r}_t,\boldsymbol{T}_t+\sigma^2\boldsymbol{I})p(\boldsymbol{g}_t|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\mathrm{d}\boldsymbol{g}_t \\
&= Z(\sigma^2\boldsymbol{I})p(\boldsymbol{\mu}_t|\boldsymbol{r}_t,\boldsymbol{T}_t+\sigma^2\boldsymbol{I}) \\
&= \sqrt{\frac{\det(\sigma^2\boldsymbol{I})}{\det(\boldsymbol{T}_t+\sigma^2\boldsymbol{I})}}\exp\left(-\frac{1}{2}(\boldsymbol{\mu}_t-\boldsymbol{r}_t)^\top(\boldsymbol{T}_t+\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{\mu}_t-\boldsymbol{r}_t)\right)
\end{aligned}$$

Substituting this into (B.5), the constraint becomes

$$\begin{aligned}
\|\Phi_t\boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 &\leq \sigma^2(\boldsymbol{\mu}_t-\boldsymbol{r}_t)^\top(\boldsymbol{T}_t+\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{\mu}_t-\boldsymbol{r}_t) - 2\sigma^2\ln\left(\sqrt{\frac{\det(\sigma^2\boldsymbol{I})}{\det(\boldsymbol{T}_t+\sigma^2\boldsymbol{I})}}\right) + 2\sigma^2\ln\left(\frac{1}{\delta}\right). \\
&= (\boldsymbol{\mu}_t-\boldsymbol{r}_t)^\top\left(\boldsymbol{I}+\frac{\boldsymbol{T}_t}{\sigma^2}\right)^{-1}(\boldsymbol{\mu}_t-\boldsymbol{r}_t) + \sigma^2\ln\det\left(\boldsymbol{I}+\frac{\boldsymbol{T}_t}{\sigma^2}\right) + 2\sigma^2\ln\left(\frac{1}{\delta}\right).
\end{aligned}$$

B.3 Computing Upper Confidence Bounds

First, we state and prove some useful lemmas.

Lemma B.3. *For any $\alpha > 0$*

$$(\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t) + \alpha\boldsymbol{\theta}^\top\boldsymbol{\theta} - R^2_{\mathrm{MM},t} - \alpha B_2^2 = (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\alpha,t})^\top \left(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I}\right)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\alpha,t}) - R^2_{\mathrm{AMM},t},$$

where $R^2_{\mathrm{MM},t}$ is the squared radius quantity from Cor. 4.2 and

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\alpha,t} &= \left(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I}\right)^{-1}\Phi_t^\top\boldsymbol{r}_t,\\ R^2_{\mathrm{AMM},t} &= R^2_{\mathrm{MM},t} + \alpha B_2^2 - \boldsymbol{r}_t^\top\boldsymbol{r}_t + \boldsymbol{r}_t^\top\Phi_t\left(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I}\right)^{-1}\Phi_t^\top\boldsymbol{r}_t.\end{aligned}$$

Proof. For any symmetric matrix $\boldsymbol{A}$, we have

$$(\boldsymbol{\theta} - \boldsymbol{b})^\top\boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{b}) + c = \boldsymbol{\theta}^\top\boldsymbol{A}\boldsymbol{\theta} - 2\boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{b} + c.$$

We also have

$$(\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t) + \alpha\boldsymbol{\theta}^\top\boldsymbol{\theta} - R^2_{\mathrm{MM},t} - \alpha B_2^2 = \boldsymbol{\theta}^\top\left(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I}\right)\boldsymbol{\theta} - 2\boldsymbol{r}_t^\top\Phi_t\boldsymbol{\theta} + \boldsymbol{r}_t^\top\boldsymbol{r}_t - R^2_{\mathrm{MM},t} - \alpha B_2^2.$$

We can now find $\boldsymbol{A}$, $\boldsymbol{b}$ and c by equating coefficients. We find that

$$\boldsymbol{A} = \Phi_t^\top\Phi_t + \alpha\boldsymbol{I},$$

which is a symmetric matrix. We have

$$\begin{aligned}&\boldsymbol{b}^\top\boldsymbol{A} = \boldsymbol{r}_t^\top\Phi_t\\ \implies\ &\boldsymbol{A}\boldsymbol{b} = \Phi_t^\top\boldsymbol{r}_t\\ \implies\ &\boldsymbol{b} = \left(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I}\right)^{-1}\Phi_t^\top\boldsymbol{r}_t = \widehat{\boldsymbol{\theta}}_{\alpha,t}.\end{aligned}$$

Finally, we have

$$\begin{aligned}c &= -R^2_{\mathrm{MM},t} - \alpha B_2^2 + \boldsymbol{r}_t^\top\boldsymbol{r}_t - \boldsymbol{b}^\top\boldsymbol{A}\boldsymbol{b}\\ &= -R^2_{\mathrm{MM},t} - \alpha B_2^2 + \boldsymbol{r}_t^\top\boldsymbol{r}_t - \boldsymbol{r}_t^\top\Phi_t\left(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I}\right)^{-1}\Phi_t^\top\boldsymbol{r}_t\\ &= -R^2_{\mathrm{AMM},t}.\end{aligned}$$

Therefore, we have shown that

$$(\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t) + \alpha\boldsymbol{\theta}^\top\boldsymbol{\theta} - R^2_{\mathrm{MM},t} - \alpha B_2^2 = (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\alpha,t})^\top \left(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I}\right)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\alpha,t}) - R^2_{\mathrm{AMM},t}.$$

□

Lemma B.4. *For any symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, any $R > 0$, and any $\eta < 0$*

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{ \boldsymbol{a}^\top \boldsymbol{\theta} + \eta \left((\boldsymbol{\theta} - \boldsymbol{b})^\top \boldsymbol{A} (\boldsymbol{\theta} - \boldsymbol{b}) - R^2 \right) \right\} = \boldsymbol{a}^\top \boldsymbol{b} - \frac{1}{4\eta} \boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a} - \eta R^2.$$

Proof. Let

$$f(\boldsymbol{\theta}) = \boldsymbol{a}^\top \boldsymbol{\theta} + \eta \left((\boldsymbol{\theta} - \boldsymbol{b})^\top \boldsymbol{A} (\boldsymbol{\theta} - \boldsymbol{b}) - R^2 \right)$$

The gradient and Hessian of f are

$$\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{\theta}) = \boldsymbol{a} + 2\eta \boldsymbol{A} (\boldsymbol{\theta} - \boldsymbol{b}), \qquad \frac{\partial^2}{\partial \boldsymbol{\theta}^2} f(\boldsymbol{\theta}) = 2\eta \boldsymbol{A}.$$

Since $\boldsymbol{A}$ is positive definite and $\eta < 0$, $\frac{\partial^2}{\partial \boldsymbol{\theta}^2} f(\boldsymbol{\theta})$ is negative definite for all $\boldsymbol{\theta} \in \mathbb{R}^d$. Therefore, any solution $\boldsymbol{\theta}^*$ of $\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{\theta}) = 0$ must be a maximiser of $f(\boldsymbol{\theta})$. There is a unique solution, which is

$$\boldsymbol{\theta}^* = \boldsymbol{b} - \frac{1}{2\eta} \boldsymbol{A}^{-1} \boldsymbol{a}.$$

The maximum is

$$f(\boldsymbol{\theta}^*) = \boldsymbol{a}^\top \boldsymbol{b} - \frac{1}{4\eta} \boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a} - \eta R^2.$$

□

Lemma B.5. *For any symmetric, positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, and any $R > 0$*

$$\min_{\eta<0} \left\{ \boldsymbol{a}^\top \boldsymbol{b} - \frac{1}{4\eta} \boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a} - \eta R^2 \right\} = \boldsymbol{a}^\top \boldsymbol{b} + R\sqrt{\boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a}}.$$

Proof. Let

$$g(\eta) = \boldsymbol{a}^\top \boldsymbol{b} - \frac{1}{4\eta} \boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a} - \eta R^2.$$

The first and second derivatives of g are

$$\frac{\mathrm{d}}{\mathrm{d}\eta} g(\eta) = \frac{1}{4\eta^2} \boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a} - R^2, \qquad \frac{\mathrm{d}^2}{\mathrm{d}\eta^2} g(\eta) = -\frac{1}{2\eta^3} \boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a}.$$

Since $\boldsymbol{A}$ is positive definite, $\frac{\mathrm{d}^2}{\mathrm{d}\eta^2} g(\eta)$ is positive for all $\eta < 0$. Therefore, any negative solution η^* of $\frac{\mathrm{d}}{\mathrm{d}\eta} g(\eta) = 0$ must be a minimiser of $g(\eta)$. There is a unique (negative) solution, which is

$$\eta^* = -\frac{1}{2R} \sqrt{\boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a}}.$$

The minimum is

$$g(\eta^*) = \boldsymbol{a}^\top \boldsymbol{b} + R\sqrt{\boldsymbol{a}^\top \boldsymbol{A}^{-1} \boldsymbol{a}}.$$

□

B.3.1 Analytic UCBs

Here, we prove Theorem 4.3, which states that for all $\alpha > 0$:

$$\max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\} \;\leq\; \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{AMM},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)}, \tag{B.7}$$

$$\text{where} \qquad \widehat{\boldsymbol{\theta}}_{\alpha,t} = \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \Phi_t^\top \boldsymbol{r}_t,$$

$$R_{\mathrm{AMM},t}^2 = R_{\mathrm{MM},t}^2 + \alpha B_2^2 - \boldsymbol{r}_t^\top \boldsymbol{r}_t + \boldsymbol{r}_t^\top \Phi_t \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \Phi_t^\top \boldsymbol{r}_t.$$

Θ_t is the confidence set at time t in our confidence sequence from Cor. 4.2 and $R_{\mathrm{MM},t}$ is the radius quantity from Cor. 4.2. As well as proving this statement, we will also show that if Θ_t has an interior point, then when the right-hand-side of (B.7) is optimised with respect to $\alpha > 0$, the inequality in (B.7) becomes an equality, i.e.

$$\max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\} = \min_{\alpha > 0} \left\{ \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{AMM},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)} \right\}. \tag{B.8}$$

Proof of Thm. 4.3. We use weak Lagrangian duality to prove the upper bound and strong Lagrangian duality to prove the second part. The convex optimisation problem $\max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\}$ can be stated as

$$\max_{\boldsymbol{\theta} \in \mathbb{R}^d} \;\; \phi(a)^\top \boldsymbol{\theta} \quad \text{s.t. } (\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t)^\top (\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t) \leq R_{\mathrm{MM},t}^2 \quad \text{and} \quad \boldsymbol{\theta}^\top \boldsymbol{\theta} \leq B_2^2. \tag{B.9}$$

Rewriting both constraints in the form $f(\boldsymbol{\theta}) \leq 0$, we can see that the Lagrangian for this problem is

$$L(\boldsymbol{\theta}, \eta_1, \eta_2) = \phi(a)^\top \boldsymbol{\theta} + \eta_1 \left((\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t)^\top (\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t) - R_{\mathrm{MM},t}^2 \right) + \eta_2 \left(\boldsymbol{\theta}^\top \boldsymbol{\theta} - B_2^2 \right).$$

η_1 and η_2 are called the Lagrange multipliers. The Lagrange dual function (or just dual function) is

$$g(\eta_1, \eta_2) = \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ L(\boldsymbol{\theta}, \eta_1, \eta_2) \right\}.$$

By weak duality, for any $\eta_1, \eta_2 \leq 0$, the dual function is an upper bound on the solution of the primal problem in (B.9), i.e. for any $\eta_1, \eta_2 \leq 0$

$$\max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\} \leq g(\eta_1, \eta_2). \tag{B.10}$$

Alternatively, (B.10) can be verified by starting from the inequality $\phi(a)^\top \boldsymbol{\theta} \leq L(\boldsymbol{\theta}, \eta_1, \eta_2)$ for all $\boldsymbol{\theta} \in \Theta_t, \eta_1 \leq 0$, and $\eta_2 \leq 0$. The challenge is to set the Lagrange multipliers such that the dual function has a closed-form expression while being as close as possible to its minimum value $\min_{\eta_1, \eta_2 \leq 0} \{ g(\eta_1, \eta_2) \}$. We will now show that for any $\alpha > 0$, $\min_{\eta \leq 0} \{ g(\eta, \alpha\eta) \}$ has a closed-form solution, which is the right-hand-side of (B.7). The Lagrangian, evaluated with the Lagrange multipliers η and $\alpha\eta$, is

$$L(\boldsymbol{\theta}, \eta, \alpha\eta) = \phi(a)^\top \boldsymbol{\theta} + \eta \left((\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t)^\top (\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t) + \alpha \boldsymbol{\theta}^\top \boldsymbol{\theta} - R_{\mathrm{MM},t}^2 - \alpha B_2^2 \right).$$

Using Lemma B.3, the Lagrangian can be rewritten as

$$L(\boldsymbol{\theta}, \eta, \alpha\eta) = \phi(a)^\top \boldsymbol{\theta} + \eta \left((\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\alpha,t})^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right) (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\alpha,t}) - R_{\mathrm{AMM},t}^2 \right).$$

Using Lemma B.4, the dual function evaluated at η and $\alpha\eta$ is

$$\begin{aligned} g(\eta, \alpha\eta) &= \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \phi(a)^\top \boldsymbol{\theta} + \eta \left((\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\alpha,t})^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right) (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\alpha,t}) - R_{\mathrm{AMM},t}^2 \right) \right\} \\ &= \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} - \frac{1}{4\eta} \phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a) - \eta R_{\mathrm{AMM},t}^2. \end{aligned}$$

Using Lemma B.5, we have

$$\begin{aligned} \min_{\eta \leq 0} \{ g(\eta, \alpha\eta) \} &= \min_{\eta \leq 0} \left\{ \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} - \frac{1}{4\eta} \phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a) - \eta R_{\mathrm{AMM},t}^2 \right\} \\ &= \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{AMM},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)}. \end{aligned}$$

This concludes the proof of Theorem 4.3. To prove (B.8), we use strong duality. Clearly

$$\min_{\alpha > 0} \min_{\eta \leq 0} \{ g(\eta, \alpha\eta) \} = \min_{\eta_1, \eta_2 \leq 0} \{ g(\eta_1, \eta_2) \},$$

so if we optimise the upper bound in (B.7) with respect to α, then we will recover the minimum of the dual function. If strong duality holds, then the minimum of the dual function is equal to $\max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\}$. Since $\max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\}$ is a convex optimisation problem, we can use Slater's condition to obtain a sufficient condition for strong duality to hold. In particular, if $\Theta_t^{\ell_2}$ has an interior point, then strong duality holds, which means

$$\begin{aligned} \max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \{ \phi(a)^\top \boldsymbol{\theta} \} &= \min_{\alpha > 0} \min_{\eta \leq 0} \{ g(\eta, \alpha\eta) \} \\ &= \min_{\alpha > 0} \left\{ \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{AMM},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)} \right\}. \end{aligned}$$

□

One can follow the same steps, with a few minor modifications, to prove a similar statement for lower confidence bounds. For all $\alpha > 0$

$$\min_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\} \geq \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} - R_{\mathrm{AMM},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)}.$$

If Θ_t has an interior point, then

$$\min_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\} = \max_{\alpha > 0} \left\{ \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} - R_{\mathrm{AMM},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)} \right\}.$$

B.3.2 OFUL vs AMM-UCB (and CMM-UCB)

We will show that for any value of the parameter α, we can choose a sequence of Gaussian mixture distributions such that the confidence bounds of AMM-UCB (and therefore also CMM-UCB) are always tighter than the confidence bounds of OFUL [3].

To do this, we will use the following lemma.

Lemma B.6. *For any $\gamma > 0$, $\boldsymbol{v} \in \mathbb{R}^t$ and $\boldsymbol{M} \in \mathbb{R}^{t \times d}$, we have*

$$\boldsymbol{v}^\top \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{M} \left(\boldsymbol{M}^\top \boldsymbol{M} + \gamma \boldsymbol{I} \right)^{-1} \boldsymbol{M}^\top \boldsymbol{v} = \boldsymbol{v}^\top \left(\frac{1}{\gamma} \boldsymbol{M} \boldsymbol{M}^\top + \boldsymbol{I} \right)^{-1} \boldsymbol{v}.$$

Proof. We start with the identity

$$\boldsymbol{M} \left(\boldsymbol{M}^\top \boldsymbol{M} + \gamma \boldsymbol{I} \right) = \left(\boldsymbol{M} \boldsymbol{M}^\top + \gamma \boldsymbol{I} \right) \boldsymbol{M}.$$

By post-multiplying both sides with $\left(\boldsymbol{M}^\top \boldsymbol{M} + \gamma \boldsymbol{I} \right)^{-1}$ and pre-multiplying both sides with $\left(\boldsymbol{M} \boldsymbol{M}^\top + \gamma \boldsymbol{I} \right)^{-1}$, we obtain

$$\left(\boldsymbol{M} \boldsymbol{M}^\top + \gamma \boldsymbol{I} \right)^{-1} \boldsymbol{M} = \boldsymbol{M} \left(\boldsymbol{M}^\top \boldsymbol{M} + \gamma \boldsymbol{I} \right)^{-1}. \tag{B.11}$$

Now, using (B.11), we have

$$\begin{aligned}
\boldsymbol{v}^\top \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{M} \left(\boldsymbol{M}^\top \boldsymbol{M} + \gamma \boldsymbol{I} \right)^{-1} \boldsymbol{M}^\top \boldsymbol{v} &= \boldsymbol{v}^\top \boldsymbol{v} - \boldsymbol{v}^\top \left(\boldsymbol{M} \boldsymbol{M}^\top + \gamma \boldsymbol{I} \right)^{-1} \boldsymbol{M} \boldsymbol{M}^\top \boldsymbol{v} \\
&= \boldsymbol{v}^\top \boldsymbol{v} - \boldsymbol{v}^\top \left(\boldsymbol{M} \boldsymbol{M}^\top + \gamma \boldsymbol{I} \right)^{-1} \left(\boldsymbol{M} \boldsymbol{M}^\top + \gamma \boldsymbol{I} - \gamma \boldsymbol{I} \right) \boldsymbol{v} \\
&= \boldsymbol{v}^\top \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{v} + \gamma \boldsymbol{v}^\top \left(\boldsymbol{M} \boldsymbol{M}^\top + \gamma \boldsymbol{I} \right)^{-1} \boldsymbol{v} \\
&= \boldsymbol{v}^\top \left(\frac{1}{\gamma} \boldsymbol{M} \boldsymbol{M}^\top + \boldsymbol{I} \right)^{-1} \boldsymbol{v}.
\end{aligned}$$

□

With $\boldsymbol{v} = \boldsymbol{r}_t$ and $\boldsymbol{M} = \Phi_t$, we obtain

$$\boldsymbol{r}_t^\top \boldsymbol{r}_t - \boldsymbol{r}_t^\top \Phi_t \left(\Phi_t^\top \Phi_t + \gamma \boldsymbol{I} \right)^{-1} \Phi_t^\top \boldsymbol{r}_t = \boldsymbol{r}_t^\top \left(\frac{1}{\gamma} \Phi_t \Phi_t^\top + \boldsymbol{I} \right)^{-1} \boldsymbol{r}_t.$$

We will also use the fact that, due to the Weinstein–Aronszajn identity, for any $\gamma > 0$

$$\det(\gamma \Phi_t^\top \Phi_t + \boldsymbol{I}) = \det(\gamma \Phi_t \Phi_t^\top + \boldsymbol{I}). \tag{B.12}$$

For any $\alpha > 0$, the OFUL UCB states that

$$\phi(a)^\top \boldsymbol{\theta}^* \leq \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{OFUL},t} \sqrt{\phi(a)^\top \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I} \right)^{-1} \phi(a)},$$

$$\text{where} \quad R_{\mathrm{OFUL},t} = \sigma \sqrt{\ln \left(\det \left(\frac{1}{\alpha} \Phi_t^\top \Phi_t + \boldsymbol{I} \right) \right) + 2 \ln(1/\delta)} + \sqrt{\alpha} B_2.$$

Without loss of generality, suppose we choose $\alpha = \sigma^2/c$, for some $c > 0$. With this choice, the OFUL radius is

$$R_{\mathrm{OFUL},t} = \sigma \left(\sqrt{\ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t^\top \Phi_t + \boldsymbol{I}\right)\right) + 2\ln(1/\delta)} + \frac{B_2}{\sqrt{c}} \right).$$

For any $\alpha > 0$ and a Gaussian mixture distribution $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$, the squared AMM-UCB radius is

$$\begin{aligned}
R_{\mathrm{AMM},t}^2 &= R_{\mathrm{MM},t}^2 + \alpha B_2^2 - \boldsymbol{r}_t^\top \boldsymbol{r}_t + \boldsymbol{r}_t^\top \Phi_t \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I}\right)^{-1} \Phi_t^\top \boldsymbol{r}_t \\
&= (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2}\right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{r}_t) + \sigma^2 \ln\left(\det\left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2}\right)\right) + 2\sigma^2 \ln\left(\frac{1}{\delta}\right) \\
&\quad + \alpha B_2^2 - \boldsymbol{r}_t^\top \boldsymbol{r}_t + \boldsymbol{r}_t^\top \Phi_t \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I}\right)^{-1} \Phi_t^\top \boldsymbol{r}_t.
\end{aligned}$$

For AMM-UCB, we will use $\alpha = \sigma^2/c$ and the mixture distribution $P_t = \mathcal{N}(\boldsymbol{0}_t, c\Phi_t\Phi_t^\top)$ for each t. One can verify that this choice satisfies the conditions required for a sequence of adaptive mixture distributions, so the AMM-UCB is valid with these mixture distributions. With these choices, and using Lemma B.6 and (B.12), the squared AMM-UCB radius is

$$\begin{aligned}
R_{\mathrm{AMM},t}^2 &= \boldsymbol{r}_t^\top \left(\frac{c}{\sigma^2}\Phi_t\Phi_t^\top + \boldsymbol{I}\right)^{-1} \boldsymbol{r}_t - \boldsymbol{r}_t^\top \boldsymbol{r}_t + \boldsymbol{r}_t^\top \Phi_t \left(\Phi_t^\top \Phi_t + \frac{\sigma^2}{c}\boldsymbol{I}\right)^{-1} \Phi_t^\top \boldsymbol{r}_t \\
&\quad + \sigma^2 \ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t\Phi_t^\top + \boldsymbol{I}\right)\right) + 2\sigma^2 \ln\left(\frac{1}{\delta}\right) + \frac{\sigma^2 B_2^2}{c} \\
&= \sigma^2 \left(\ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t^\top \Phi_t + \boldsymbol{I}\right)\right) + 2\ln\left(\frac{1}{\delta}\right) + \frac{B_2^2}{c} \right).
\end{aligned} \tag{B.13}$$

Using the basic inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, we have

$$\begin{aligned}
R_{\mathrm{AMM},t} &= \sigma \sqrt{\ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t^\top \Phi_t + \boldsymbol{I}\right)\right) + 2\ln\left(\frac{1}{\delta}\right) + \frac{B_2^2}{c}} \\
&\le \sigma \left(\sqrt{\ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t^\top \Phi_t + \boldsymbol{I}\right)\right) + 2\ln\left(\frac{1}{\delta}\right)} + \frac{B_2}{\sqrt{c}} \right) \\
&= R_{\mathrm{OFUL},t}.
\end{aligned}$$

Therefore, the confidence bounds of AMM-UCB, with $\alpha = \sigma^2/c$ and $P_t = \mathcal{N}(\boldsymbol{0}, c\Phi_t\Phi_t^\top)$, are never looser than the confidence bounds of OFUL with an arbitrary $\alpha = \sigma^2/c$. Since $\ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t^\top \Phi_t + \boldsymbol{I}\right)\right) + 2\ln\left(\frac{1}{\delta}\right)$ and B_2^2/c are strictly positive, there is actually a strict inequality. This means that the AMM-UCB (and CMM-UCB) confidence bounds are always strictly tighter than the OFUL confidence bounds.

Note that $P_t = \mathcal{N}(\boldsymbol{0}_t, c\Phi_t\Phi_t^\top)$ is not necessarily the best choice for the mixture distribution. With a better choice of the mixture distribution, e.g. a mixture distribution that is chosen using some prior knowledge about the expected reward function and/or refined using previously observed rewards, $R_{\mathrm{AMM},t}$ will be smaller and the gap between AMM-UCB and OFUL will be greater.

B.4 Cumulative Regret Bounds

In this section, we prove the cumulative regret bounds stated in Section 4.7. We prove the data-dependent regret bound in Thm. 4.8. We also prove the data-independent regret bound in Thm. 4.9 and another data-independent regret bound, which holds for more general choices of the mixture distributions and the α parameter.

For convenience, we use some more compact notation in this section. For a symmetric positive semi-definite matrix $\boldsymbol{A}$ and vector $\boldsymbol{x}$, let

$$\|\boldsymbol{x}\|_{\boldsymbol{A}} := \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}.$$

Before presenting the proof of the main results, we state some useful lemmas.

Lemma B.7 (Determinant-Trace Inequality [3])**.** *If assumption 4.6 holds (i.e. $\|\phi(a)\|_2 \le L_2$), then for any $\gamma > 0$*

$$\ln\left(\det\left(\gamma\Phi_t^\top\Phi_t + \boldsymbol{I}\right)\right) \le d\ln\left(1 + \gamma t L_2^2/d\right). \tag{B.14}$$

The Determinant-Trace Inequality in Lemma 10 of [3] is stated in the form

$$\det\left(\Phi_t^\top\Phi_t + \frac{1}{\gamma}\boldsymbol{I}\right) \le (1/\gamma + tL_2^2/d)^d. \tag{B.15}$$

Since $\det\left(\gamma\Phi_t^\top\Phi_t + \boldsymbol{I}\right) = \det\left(\Phi_t^\top\Phi_t + (1/\gamma)\boldsymbol{I}\right) / \det\left((1/\gamma)\boldsymbol{I}\right)$, the statement in (B.14) follows from (B.15).

Lemma B.8. *For any $\sigma > 0$ and any $\sigma_0 > 0$, define $\boldsymbol{\Sigma}_t = (\frac{1}{\sigma^2}\Phi_t^\top\Phi_t + \frac{1}{\sigma_0^2}\boldsymbol{I})^{-1}$. We have*

$$\mathrm{tr}(\Phi_t^\top\Phi_t\boldsymbol{\Sigma}_t) = \sigma^2 d - \frac{\sigma^2}{\sigma_0^2}\mathrm{tr}(\boldsymbol{\Sigma}_t) \le \sigma^2 d.$$

Proof. Since $\boldsymbol{\Sigma}_t$ is positive-definite, its trace is positive. Now

$$\begin{aligned}
\mathrm{tr}(\Phi_t^\top\Phi_t\boldsymbol{\Sigma}_t) &= \sigma^2\mathrm{tr}\left(\frac{1}{\sigma^2}\Phi_t^\top\Phi_t\left(\frac{1}{\sigma^2}\Phi_t^\top\Phi_t + \frac{1}{\sigma_0^2}\boldsymbol{I}\right)^{-1}\right) \\
&= \sigma^2\mathrm{tr}\left(\left(\frac{1}{\sigma^2}\Phi_t^\top\Phi_t + \frac{1}{\sigma_0^2}\boldsymbol{I}\right)\left(\frac{1}{\sigma^2}\Phi_t^\top\Phi_t + \frac{1}{\sigma_0^2}\boldsymbol{I}\right)^{-1} - \frac{1}{\sigma_0^2}\left(\frac{1}{\sigma^2}\Phi_t^\top\Phi_t + \frac{1}{\sigma_0^2}\boldsymbol{I}\right)^{-1}\right) \\
&= \sigma^2 d - \frac{\sigma^2}{\sigma_0^2}\mathrm{tr}(\boldsymbol{\Sigma}_t) \\
&\le \sigma^2 d.
\end{aligned}$$

□

Lemma B.9. *For any $\sigma > 0$ and any $\sigma_0 > 0$, the matrix $\boldsymbol{\Sigma}_t = (\frac{1}{\sigma^2}\Phi_t^\top \Phi_t + \frac{1}{\sigma_0^2}\boldsymbol{I})^{-1}$ satisfies*

$$\operatorname{tr}(\boldsymbol{\Sigma}_t) \le \frac{d}{\sigma_0^2}.$$

Proof. Let $\{\gamma_i\}_{i=1}^d$ denote the eigenvalues of $\Phi_t^\top \Phi_t$. Since $\Phi_t^\top \Phi_t$ is positive semi-definite, its eigenvalues are real and non-negative. From the definition of eigenvalues, one can verify that the eigenvalues of $\boldsymbol{\Sigma}_t$ are $\{\frac{\sigma^2}{\gamma_i + \sigma^2/\sigma_0^2}\}_{i=1}^d$. Using this, we have

$$\operatorname{tr}(\boldsymbol{\Sigma}_t) = \sum_{i=1}^d \frac{\sigma^2}{\gamma_i + \sigma^2/\sigma_0^2} \le \sum_{i=1}^d \frac{\sigma^2}{\sigma^2/\sigma_0^2} = \frac{d}{\sigma_0^2}.$$

□

Lemma B.10. *Let $\boldsymbol{\epsilon}_t$ denote the vector containing the first t noise variables (so $\boldsymbol{r}_t = \Phi_t \boldsymbol{\theta}^* + \boldsymbol{\epsilon}_t$). For any $\alpha > 0$, we have*

$$(\Phi_t\boldsymbol{\theta}^* - \boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}^* - \boldsymbol{r}_t) - \boldsymbol{r}_t^\top \boldsymbol{r}_t + \boldsymbol{r}_t^\top \Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{r}_t \le \left\|\Phi_t^\top \boldsymbol{\epsilon}_t\right\|^2_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}} + 2\alpha \left\|\boldsymbol{\theta}^*\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}} \left\|\Phi_t^\top \boldsymbol{\epsilon}_t\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}.$$

Proof. Using $\boldsymbol{r}_t = \Phi_t\boldsymbol{\theta}^* + \boldsymbol{\epsilon}_t$, we have

$$(\Phi_t\boldsymbol{\theta}^* - \boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}^* - \boldsymbol{r}_t) = \boldsymbol{\epsilon}_t^\top \boldsymbol{\epsilon}_t,$$

and

$$\begin{aligned}
-\boldsymbol{r}_t^\top \boldsymbol{r}_t + \boldsymbol{r}_t^\top \Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{r}_t &= -(\Phi_t\boldsymbol{\theta}^* + \boldsymbol{\epsilon}_t)^\top(\Phi_t\boldsymbol{\theta}^* + \boldsymbol{\epsilon}_t) \\
&\quad + (\Phi_t\boldsymbol{\theta}^* + \boldsymbol{\epsilon}_t)^\top \Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top(\Phi_t\boldsymbol{\theta}^* + \boldsymbol{\epsilon}_t) \\
&= -\boldsymbol{\epsilon}_t^\top \boldsymbol{\epsilon}_t - 2\boldsymbol{\theta}^{*\top}\Phi_t^\top \boldsymbol{\epsilon}_t - \boldsymbol{\theta}^{*\top}\Phi_t^\top\Phi_t\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*\top}\Phi_t^\top\Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top\Phi_t\boldsymbol{\theta}^* \\
&\quad + 2\boldsymbol{\theta}^{*\top}\Phi_t^\top\Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_t^\top\Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{\epsilon}_t \\
&\le -\boldsymbol{\epsilon}_t^\top \boldsymbol{\epsilon}_t - 2\boldsymbol{\theta}^{*\top}\Phi_t^\top \boldsymbol{\epsilon}_t + 2\boldsymbol{\theta}^{*\top}\Phi_t^\top\Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{\epsilon}_t \\
&\quad + \boldsymbol{\epsilon}_t^\top\Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{\epsilon}_t \\
&= -\boldsymbol{\epsilon}_t^\top \boldsymbol{\epsilon}_t - 2\alpha\boldsymbol{\theta}^{*\top}(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_t^\top\Phi_t(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{\epsilon}_t.
\end{aligned}$$

Using the Cauchy-Schwarz inequality, we have

$$|\boldsymbol{\theta}^{*\top}(\Phi_t^\top\Phi_t + \alpha\boldsymbol{I})^{-1}\Phi_t^\top \boldsymbol{\epsilon}_t| \le \left\|\boldsymbol{\theta}^*\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}} \left\|\Phi_t^\top \boldsymbol{\epsilon}_t\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}.$$

Therefore

$$-2\alpha{\boldsymbol{\theta}^*}^\top(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}\Phi_t^\top\boldsymbol{\epsilon}_t \le 2\alpha\left\|\boldsymbol{\theta}^*\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}\left\|\Phi_t^\top\boldsymbol{\epsilon}_t\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}},$$

and

$$\begin{aligned}(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)-\boldsymbol{r}_t^\top\boldsymbol{r}_t+\boldsymbol{r}_t^\top\Phi_t(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}\Phi_t^\top\boldsymbol{r}_t &\le \boldsymbol{\epsilon}_t^\top\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_t^\top\boldsymbol{\epsilon}_t+\left\|\Phi_t^\top\boldsymbol{\epsilon}_t\right\|^2_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}\\ &\quad+2\alpha\left\|\boldsymbol{\theta}^*\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}\left\|\Phi_t^\top\boldsymbol{\epsilon}_t\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}\\ &=\left\|\Phi_t^\top\boldsymbol{\epsilon}_t\right\|^2_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}+2\alpha\left\|\boldsymbol{\theta}^*\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}\left\|\Phi_t^\top\boldsymbol{\epsilon}_t\right\|_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}.\end{aligned}$$

□

Theorem B.11 (Self-Normalised Bound for Vector-Valued Martingales (Theorem 1 of [3]))**.** *Let $(\mathcal{D}_t|t\ge 0)$ be a filtration. Let $(\epsilon_t|t\ge 1)$ be a real-valued stochastic process such that ϵ_t is $\mathcal{H}_t$-measurable and ϵ_t is conditionally σ-sub-Gaussian for some $\sigma>0$. Let $(\phi(a_t)|t\ge 1)$ be an $\mathbb{R}^d$-valued stochastic process such that $\phi(a_t)$ is $\mathcal{D}_{t-1}$-measurable. For any $\delta\in(0,1]$ and any $\alpha>0$, with probability at least $1-\delta$*

$$\forall t\ge 0,\qquad \left\|\Phi_t^\top\boldsymbol{\epsilon}_t\right\|^2_{(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I})^{-1}}\le\sigma^2\ln\left(\det\left(\frac{1}{\alpha}\Phi_t^\top\Phi_t+\boldsymbol{I}\right)\right)+2\sigma^2\ln(1/\delta).$$

Lemma B.12. *For any symmetric positive semi-definite matrix $\boldsymbol{A}$ with largest eigenvalue $\nu_{\max}$, we have*

$$\|\boldsymbol{x}\|_{\boldsymbol{A}}^2\le\nu_{\max}\|\boldsymbol{x}\|_2^2.$$

Proof. Let $\{\nu_i\}_{i=1}^d$ and $\{\boldsymbol{v}_i\}_{i=1}^d$ be the eigenvalues and eigenvectors of $\boldsymbol{A}$. Since $\{\boldsymbol{v}_i\}_{i=1}^d$ form an orthonormal basis, there are constants $\{c_i\}_{i=1}^d$ such that $\boldsymbol{x}=\sum_{i=1}^d c_i\boldsymbol{v}_i$. We have

$$\|\boldsymbol{x}\|_{\boldsymbol{A}}^2=\sum_{i=1,j=1}^d c_ic_j\boldsymbol{v}_i^\top\boldsymbol{A}\boldsymbol{v}_j=\sum_{i=1,j=1}^d\nu_jc_ic_j\boldsymbol{v}_i^\top\boldsymbol{v}_j\le\nu_{\max}\sum_{i=1,j=1}^d c_ic_j\boldsymbol{v}_i^\top\boldsymbol{v}_j=\nu_{\max}\|\boldsymbol{x}\|_2^2.$$

□

Lemma B.13. *For all $x\ge 0$,*

$$\min(1,x)\le\frac{1}{\ln(2)}\ln(1+x).$$

Proof. Since $\ln(1+x)/\ln(2)$ is monotonically increasing in x, we only need to prove that $x\le\ln(1+x)/\ln(2)$ for all $x\in[0,1]$. For any positive constant a, the function $a\ln(1+x)$ is concave on the domain $[0,1]$. Therefore, if $x\le a\ln(1+x)$ at the end points $x=0$ and $x=1$, then $x\le a\ln(1+x)$ for every $x\in[0,1]$. At $x=0$, we have $a\ln(1+x)=0$ for any a, which means we can choose the smallest a such that $1\le a\ln(1+1)$. By rearranging this inequality, we obtain $a\ge 1/\ln(2)$. □

B.4.1 Data-Dependent Regret Bound

First, we show that the cumulative regret of both of our algorithms can be upper bounded by the sum of the widths of the UCB/LCBs that they use. Let

$$\mathrm{UCB}_{\Theta_t^{\ell_2}}(a) = \max_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\}, \quad \text{and} \quad \mathrm{LCB}_{\Theta_t^{\ell_2}}(a) = \min_{\boldsymbol{\theta} \in \Theta_t^{\ell_2}} \left\{ \phi(a)^\top \boldsymbol{\theta} \right\}.$$

In words, $\mathrm{UCB}_{\Theta_t^{\ell_2}}(a)$ and $\mathrm{LCB}_{\Theta_t^{\ell_2}}(a)$ are the upper and lower confidence bounds used by CMM-UCB (evaluated at a). Similarly, let

$$\begin{aligned}
\mathrm{AUCB}_{\Theta_t^{\ell_2}}(a) &= \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} + R_{\mathrm{AMM},t} \left\| \phi(a) \right\|_{(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I})^{-1}}, \\
\mathrm{ALCB}_{\Theta_t^{\ell_2}}(a) &= \phi(a)^\top \widehat{\boldsymbol{\theta}}_{\alpha,t} - R_{\mathrm{AMM},t} \left\| \phi(a) \right\|_{(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I})^{-1}}.
\end{aligned}$$

$\mathrm{AUCB}_{\Theta_t}(a)$ and $\mathrm{ALCB}_{\Theta_t}(a)$ are the analytic upper and lower confidence bounds used by AMM-UCB. Lemma B.14 shows that the cumulative regret of CMM-UCB and AMM-UCB can be upper bounded by the sum of the widths (UCB minus LCB) of the confidence bounds that they use.

Lemma B.14. *Suppose the actions $a_1, a_2, \ldots$ are selected by the CMM-UCB algorithm. For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$ and any $\delta \in (0, 1]$, with probability at least $1 - \delta$*

$$\forall T \geq 1, \qquad \sum_{t=1}^{T} \Delta(a_t) \leq \sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}^{\ell_2}}(a_t) - \mathrm{LCB}_{\Theta_{t-1}^{\ell_2}}(a_t). \tag{B.16}$$

Suppose the actions $a_1, a_2, \ldots$ are selected by the AMM-UCB algorithm. For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$ and any $\delta \in (0, 1]$, with probability at least $1 - \delta$

$$\forall \alpha > 0, T \geq 1, \qquad \sum_{t=1}^{T} \Delta(a_t) \leq \sum_{t=1}^{T} \mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a_t) - \mathrm{ALCB}_{\Theta_{t-1}^{\ell_2}}(a_t). \tag{B.17}$$

Proof. Using Cor. 4.2 (i.e. the fact that $\Theta_1, \Theta_2, \ldots$ is a confidence sequence), for any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$ and any $\delta \in (0, 1]$, with probability at least $1 - \delta$

$$\forall a \in \mathcal{A}, t \geq 1, \qquad \mathrm{LCB}_{\Theta_{t-1}^{\ell_2}}(a) \leq \phi(a)^\top \boldsymbol{\theta}^* \leq \mathrm{UCB}_{\Theta_{t-1}^{\ell_2}}(a).$$

Using Thm. 4.3, this implies

$$\forall \alpha > 0, a \in \mathcal{A}, t \geq 1, \qquad \mathrm{ALCB}_{\Theta_{t-1}^{\ell_2}}(a) \leq \phi(a)^\top \boldsymbol{\theta}^* \leq \mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a).$$

Let $a_1, a_2, \ldots$ be the actions selected by CMM-UCB, i.e. $a_t = \mathrm{argmax}_{a \in \mathcal{A}_t} \left\{ \mathrm{UCB}_{\Theta_{t-1}^{\ell_2}}(a) \right\}$. Then,

with probability at least $1-\delta$, we have

$$\begin{aligned}\sum_{t=1}^{T}\Delta(a_t) &= \sum_{t=1}^{T}\phi(a_t^*)^\top\boldsymbol{\theta}^* - \phi(a_t)^\top\boldsymbol{\theta}^* \\ &\leq \sum_{t=1}^{T}\mathrm{UCB}_{\Theta_{t-1}^{\ell_2}}(a_t^*) - \mathrm{LCB}_{\Theta_{t-1}^{\ell_2}}(a_t) \\ &\leq \sum_{t=1}^{T}\mathrm{UCB}_{\Theta_{t-1}^{\ell_2}}(a_t) - \mathrm{LCB}_{\Theta_{t-1}^{\ell_2}}(a_t).\end{aligned}$$

Now, let $a_1, a_2, \ldots$ be the actions selected by AMM-UCB, i.e. $a_t = \mathrm{argmax}_{a\in\mathcal{A}_t}\left\{\mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a)\right\}$. Then, with probability at least $1-\delta$, we have

$$\begin{aligned}\sum_{t=1}^{T}\Delta(a_t) &= \sum_{t=1}^{T}\phi(a_t^*)^\top\boldsymbol{\theta}^* - \phi(a_t)^\top\boldsymbol{\theta}^* \\ &\leq \sum_{t=1}^{T}\mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a_t^*) - \mathrm{ALCB}_{\Theta_{t-1}^{\ell_2}}(a_t) \\ &\leq \sum_{t=1}^{T}\mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a_t) - \mathrm{ALCB}_{\Theta_{t-1}^{\ell_2}}(a_t).\end{aligned}$$

□

Since $\forall\alpha>0, a\in\mathcal{A}$ and $t\geq 1$, $\mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a) \geq \mathrm{UCB}_{\Theta_{t-1}^{\ell_2}}(a)$ and $\mathrm{ALCB}_{\Theta_{t-1}^{\ell_2}}(a) \leq \mathrm{LCB}_{\Theta_{t-1}^{\ell_2}}(a)$, (B.16) implies that (B.17) also holds when $a_1, a_2, \ldots$ are the actions selected by CMM-UCB.

Proof of Theorem 4.8. We start by using Lemma B.14. Suppose $a_1, a_2, \ldots$ are the actions selected by CMM-UCB or AMM-UCB. For any adaptive sequence of mixture distributions $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{T}_t)$ and any $\delta\in(0,1]$, with probability at least $1-\delta$

$$\forall\alpha>0, T\geq 1, \qquad \sum_{t=1}^{T}\Delta(a_t) \leq \sum_{t=1}^{T}\mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a_t) - \mathrm{ALCB}_{\Theta_{t-1}^{\ell_2}}(a_t).$$

Using the definitions of $\mathrm{AUCB}_{\Theta_{t-1}^{\ell_2}}(a_t)$ and $\mathrm{ALCB}_{\Theta_{t-1}^{\ell_2}}(a_t)$, we have

$$\forall\alpha>0, T\geq 1, \qquad \sum_{t=1}^{T}\Delta(a_t) \leq \sum_{t=1}^{T} 2R_{\mathrm{AMM},t-1}\left\|\phi(a_t)\right\|_{(\Phi_{t-1}^\top\Phi_{t-1}+\alpha\boldsymbol{I})^{-1}}.$$

□

B.4.2 Data-Independent Regret Bound

To establish data-independent regret bounds, we first prove data-independent upper bounds on the radius $R_{\mathrm{AMM},t}$ and the "norms" $\|\phi(a_t)\|_{(\Phi_{t-1}^\top\Phi_{t-1}+\alpha\boldsymbol{I})^{-1}}$. Then, we take the data-dependent regret bound in Lemma B.14 and substitute in these bounds on the radius and the norms.

Bounding the Radius

Lemma B.15. *If, for any $c > 0$, the sequence of mixture distributions is $P_t = \mathcal{N}(\mathbf{0}_t, c\Phi_t\Phi_t^\top)$ and $\alpha = c/\sigma^2$, then*

$$R_{\mathrm{AMM},t}^2 \leq \sigma^2 d \ln\left(1 + \frac{ctL_2^2}{\sigma^2 d}\right) + \frac{\sigma^2 B_2^2}{c} + 2\sigma^2 \ln(1/\delta). \tag{B.18}$$

Proof. In Equation (B.13), we already saw that for this choice of α and the sequence of mixture distributions, we have

$$R_{\mathrm{AMM},t}^2 = \sigma^2 \ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t^\top\Phi_t + \boldsymbol{I}\right)\right) + \frac{\sigma^2 B_2^2}{c} + 2\sigma^2\ln(1/\delta).$$

To obtain a data-independent upper bound on the radius, all that remains is to upper bound $\ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t^\top\Phi_t + \boldsymbol{I}\right)\right)$ by a data-independent quantity. Using Lemma B.7, we have

$$\ln\left(\det\left(\frac{c}{\sigma^2}\Phi_t^\top\Phi_t + \boldsymbol{I}\right)\right) \leq d\ln\left(1 + \frac{ctL_2^2}{\sigma^2 d}\right).$$

Therefore

$$R_{\mathrm{AMM},t}^2 \leq \sigma^2 d \ln\left(1 + \frac{ctL_2^2}{\sigma^2 d}\right) + \frac{\sigma^2 B_2^2}{c} + 2\sigma^2\ln(1/\delta).$$

□

Lemma B.16. *If, for any $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ and any $\sigma_0 > 0$, the sequence of mixture distributions is $P_t = \mathcal{N}(\Phi_t\boldsymbol{\theta}_0, \sigma_0^2\Phi_t\Phi_t^\top)$, then for any $\delta \in (0,1]$ and any $\alpha > 0$, with probability at least $1-\delta$, for all $t \geq 1$*

$$\begin{aligned} R_{\mathrm{AMM},t}^2 \leq{}& \sigma^2 d + \frac{\sigma^2}{\sigma_0^2}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\right\|_2^2 + \sigma^2 d\ln\left(1 + \frac{t\sigma_0^2 L_2^2}{\sigma^2 d}\right) + \alpha B_2^2 + 4\sigma^2\ln(1/\delta) \\ &+ \sigma^2 d\ln\left(1 + tL_2^2/(\alpha d)\right) + 2\sqrt{\alpha}\left\|\boldsymbol{\theta}^*\right\|_2\sqrt{\sigma^2 d\ln\left(1 + tL_2^2/(\alpha d)\right) + 2\sigma^2\ln(1/\delta)}. \end{aligned}$$

Proof. In App. B.2.2 (see Equation B.5), we saw that the squared radius $R_{\mathrm{MM},t}^2$ can be written as

$$R_{\mathrm{MM},t}^2 = -2\sigma^2 \ln\left(\mathop{\mathbb{E}}_{\boldsymbol{g}_t \sim \mathcal{N}(\Phi_t\boldsymbol{\theta}_0, \sigma_0^2\Phi_t\Phi_t^\top)}\left[\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{g}_t - \boldsymbol{r}_t)^\top(\boldsymbol{g}_t - \boldsymbol{r}_t)\right)\right]\right) + 2\sigma^2\ln(1/\delta). \tag{B.19}$$

Using the substitution $\Phi_t\boldsymbol{\theta} = \boldsymbol{g}_t$, (B.19) is equivalent to

$$R_{\mathrm{MM},t}^2 = -2\sigma^2 \ln\left(\mathop{\mathbb{E}}_{\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma_0^2\boldsymbol{I})}\left[\exp\left(-\frac{1}{2\sigma^2}(\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta} - \boldsymbol{r}_t)\right)\right]\right) + 2\sigma^2\ln(1/\delta). \tag{B.20}$$

Using the Donsker-Varadhan change of measure inequality (see Lemma 2.2, and particularly Equation (2.6)), the first term on the right-hand-side of (B.20) is equal to

$$\inf_{Q\in\mathcal{P}(\mathbb{R}^d)}\left\{\mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim Q}\left[(\Phi_t\boldsymbol{\theta}-\boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}-\boldsymbol{r}_t)\right]+2\sigma^2 D_{\mathrm{KL}}(Q||\mathcal{N}(\boldsymbol{\theta}_0,\sigma_0^2\boldsymbol{I}))\right\}.$$

If we evaluate this at any specific distribution Q, we obtain an upper bound on the infimum over Q. We choose $Q=\mathcal{N}(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_t)$, where $\boldsymbol{\Sigma}_t=(\frac{1}{\sigma^2}\Phi_t^\top\Phi_t+\frac{1}{\sigma_0^2}\boldsymbol{I})^{-1}$. Combining everything so far, we have

$$\begin{aligned}
R_{\mathrm{MM},t}^2 &\leq \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim\mathcal{N}(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_t)}\left[(\Phi_t\boldsymbol{\theta}-\boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}-\boldsymbol{r}_t)\right]+2\sigma^2 D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_t)||\mathcal{N}(\boldsymbol{\theta}_0,\sigma_0^2\boldsymbol{I}))+2\sigma^2\ln(1/\delta)\\
&=(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)+\mathrm{tr}(\Phi_t^\top\Phi_t\boldsymbol{\Sigma}_t)+\frac{\sigma^2}{\sigma_0^2}\mathrm{tr}(\boldsymbol{\Sigma}_t)\\
&\quad-\sigma^2 d+\frac{\sigma^2}{\sigma_0^2}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\right\|_2^2+\sigma^2\ln\left(\frac{\det(\boldsymbol{\Sigma}_t^{-1})}{\det((1/\sigma_0^2)\boldsymbol{I})}\right)+2\sigma^2\ln(1/\delta). \qquad \text{(B.21)}
\end{aligned}$$

Using Lemma B.8, we have

$$\mathrm{tr}(\Phi_t^\top\Phi_t\boldsymbol{\Sigma}_t)\leq\sigma^2 d.$$

Using Lemma B.9, we have

$$\frac{\sigma^2}{\sigma_0^2}\mathrm{tr}(\boldsymbol{\Sigma}_t)\leq\sigma^2 d.$$

Using Lemma B.7, we have

$$\ln\left(\frac{\det(\boldsymbol{\Sigma}_t^{-1})}{\det((1/\sigma_0^2)\boldsymbol{I})}\right)=\ln\left(\det\left(\frac{\sigma_0^2}{\sigma^2}\Phi_t^\top\Phi_t+\boldsymbol{I}\right)\right)\leq d\ln\left(1+\frac{\sigma_0^2 tL_2^2}{\sigma^2 d}\right).$$

The bound on $R_{\mathrm{MM},t}^2$ in (B.21) becomes

$$R_{\mathrm{MM},t}^2\leq(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)+\frac{\sigma^2}{\sigma_0^2}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\right\|_2^2+\sigma^2 d+\sigma^2 d\ln\left(1+\frac{\sigma_0^2 tL_2^2}{\sigma^2 d}\right)+2\sigma^2\ln(1/\delta).$$

This means that

$$\begin{aligned}
R_{\mathrm{AMM},t}^2 &\leq(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}^*-\boldsymbol{r}_t)+\frac{\sigma^2}{\sigma_0^2}\left\|\boldsymbol{\theta}^*-\boldsymbol{\theta}_0\right\|_2^2+\sigma^2 d+\sigma^2 d\ln\left(1+\frac{\sigma_0^2 tL_2^2}{\sigma^2 d}\right)+2\sigma^2\ln(1/\delta)\\
&\quad+\alpha B_2^2-\boldsymbol{r}_t^\top\boldsymbol{r}_t+\boldsymbol{r}_t^\top\Phi_t\left(\Phi_t^\top\Phi_t+\alpha\boldsymbol{I}\right)^{-1}\Phi_t^\top\boldsymbol{r}_t. \qquad \text{(B.22)}
\end{aligned}$$

Finally, using Lemma B.10, then Theorem B.11 and Lemma B.12, and then Lemma B.7, for any

$\delta \in (0,1]$ and any $\alpha > 0$, with probability at least $1-\delta$, for all $t \geq 0$ simultaneously

$$
\begin{aligned}
&(\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t)^\top (\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t) - \boldsymbol{r}_t^\top \boldsymbol{r}_t + \boldsymbol{r}_t^\top \Phi_t \left(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I}\right)^{-1} \Phi_t^\top \boldsymbol{r}_t \\
&\leq \left\| \Phi_t^\top \boldsymbol{\epsilon}_t \right\|^2_{(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I})^{-1}} + 2\alpha \left\| \boldsymbol{\theta}^* \right\|_{(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I})^{-1}} \left\| \Phi_t^\top \boldsymbol{\epsilon}_t \right\|_{(\Phi_t^\top \Phi_t + \alpha \boldsymbol{I})^{-1}} \\
&\leq \sigma^2 \ln\left(\det\left(\frac{1}{\alpha}\Phi_t^\top \Phi_t + \boldsymbol{I}\right)\right) + 2\sigma^2 \ln(1/\delta) + 2\sqrt{\alpha}\left\|\boldsymbol{\theta}^*\right\|_2 \sigma \sqrt{\ln\left(\det\left(\frac{1}{\alpha}\Phi_t^\top \Phi_t + \boldsymbol{I}\right)\right) + 2\ln(1/\delta)} \\
&\leq \sigma^2 d \ln\left(\det\left(1 + \frac{tL_2^2}{\alpha d}\right)\right) + 2\sigma^2 \ln(1/\delta) + 2\sqrt{\alpha}\left\|\boldsymbol{\theta}^*\right\|_2 \sigma \sqrt{d \ln\left(\det\left(1 + \frac{tL_2^2}{\alpha d}\right)\right) + 2\ln(1/\delta)}.
\end{aligned}
$$

Substituting this into (B.22), we have

$$
\begin{aligned}
R^2_{\mathrm{AMM},t} \leq \sigma^2 d + \frac{\sigma^2}{\sigma_0^2}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\right\|_2^2 + \sigma^2 d \ln\left(1 + \frac{t\sigma_0^2 L_2^2}{\sigma^2 d}\right) + \alpha B_2^2 + 4\sigma^2 \ln(1/\delta) \\
+ \sigma^2 d \ln\left(1 + tL_2^2/(\alpha d)\right) + 2\sqrt{\alpha}\left\|\boldsymbol{\theta}^*\right\|_2 \sqrt{\sigma^2 d \ln\left(1 + tL_2^2/(\alpha d)\right) + 2\sigma^2 \ln(1/\delta)}.
\end{aligned}
$$

□

Bounding the Sum of Norms

We use the following upper bound on the sum of the squared norms.

Lemma B.17 (Lemma 11 of [3])**.** *For any $\alpha > 0$, we have*

$$
\sum_{t=1}^{T} \min\left(1, \left\|\phi(a_t)\right\|^2_{\left(\Phi_{t-1}^\top \Phi_{t-1} + \alpha \boldsymbol{I}\right)^{-1}}\right) \leq \frac{1}{\ln(2)} d \ln\left(1 + \frac{TL_2^2}{\alpha d}\right).
$$

In Lemma 11 of [3], $1/\ln(2) \approx 1.44$ is replaced with 2. We achieve an improved constant by using Lemma B.13 instead of the looser bound $\min(1,x) \leq 2\ln(1+x)$, for $x \geq 0$.

Regret Bounds

We are now ready to prove our data-independent regret bounds.

Proof of Theorem 4.9. Following the same steps as in the proof of Lemma B.14, we can also obtain the following data-dependent bound on the per-round regret for actions selected by CMM-UCB or AMM-UCB. For the mixture distributions $P_t = \mathcal{N}(\boldsymbol{0}_t, c\Phi_t\Phi_t^\top)$, $\alpha = \sigma^2/c$ and any $\delta \in (0,1]$, with probability at least $1-\delta$

$$
\forall t \geq 1, \qquad \Delta(a_t) \leq 2R_{\mathrm{AMM},t-1}\left\|\phi(a_t)\right\|_{\left(\Phi_{t-1}^\top \Phi_{t-1} + \frac{\sigma^2}{c}\boldsymbol{I}\right)^{-1}}. \tag{B.23}
$$

From Assumption 4.7 ($\phi(a)^\top \boldsymbol{\theta}^* \in [-C, C]$), we have another bound on the per-round regret

$$
\Delta(a_t) \leq 2C. \tag{B.24}
$$

The combination of (B.23) and (B.24) yields

$$\begin{aligned}\Delta(a_t) &\leq \min(2C, 2R_{\mathrm{AMM},t-1} \left\|\phi(a_t)\right\|_{(\Phi_{t-1}^\top \Phi_{t-1} + \frac{\sigma^2}{c}\boldsymbol{I})^{-1}}) \\ &\leq 2\max(C, R_{\mathrm{AMM},t-1}) \min(1, \left\|\phi(a_t)\right\|_{(\Phi_{t-1}^\top \Phi_{t-1} + \frac{\sigma^2}{c}\boldsymbol{I})^{-1}}).\end{aligned}$$

Starting with the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\sum_{t=1}^{T} \Delta(a_t) &\leq \sqrt{T \sum_{t=1}^{T} \Delta(a_t)^2} \\ &\leq \sqrt{T \sum_{t=1}^{T} 4\max\left(C^2, R_{\mathrm{AMM},t-1}^2\right) \min\left(1, \left\|\phi(a_t)\right\|^2_{(\Phi_{t-1}^\top \Phi_{t-1} + \frac{\sigma^2}{c}\boldsymbol{I})^{-1}}\right)}.\end{aligned} \tag{B.25}$$

We will now use the upper bound on $R_{\mathrm{AMM},t-1}^2$ from Lemma B.15. Let $U_{\mathrm{AMM},t-1}^2$ denote this upper bound (i.e. the right-hand-side of (B.18)). We have

$$\begin{aligned}\sum_{t=1}^{T} \Delta(a_t) &\leq \sqrt{T \sum_{t=1}^{T} 4\max\left(C^2, U_{\mathrm{AMM},t-1}^2\right) \min\left(1, \left\|\phi(a_t)\right\|^2_{(\Phi_{t-1}^\top \Phi_{t-1} + \frac{\sigma^2}{c}\boldsymbol{I})^{-1}}\right)} \\ &\leq 2\max\left(C, U_{\mathrm{AMM},T-1}\right) \sqrt{T \sum_{t=1}^{T} \min\left(1, \left\|\phi(a_t)\right\|^2_{(\Phi_{t-1}^\top \Phi_{t-1} + \frac{\sigma^2}{c}\boldsymbol{I})^{-1}}\right)}\end{aligned}$$

Finally, using the bound on the sum of norms in Lemma B.17, we have

$$\sum_{t=1}^{T} \Delta(a_t) \leq \frac{2}{\sqrt{\ln(2)}} \max\left(C, \sigma\sqrt{d\ln\left(1 + \frac{c(T-1)L_2^2}{\sigma^2 d}\right) + \frac{B_2^2}{c} + 2\ln\left(\frac{1}{\delta}\right)}\right) \sqrt{dT\ln\left(1 + \frac{cTL_2^2}{\sigma^2 d}\right)}.$$

□

Now, we state and prove a cumulative regret bound that holds for more general choices of the mixture distributions and the parameter α.

Theorem B.18. *Suppose that assumptions 4.4-4.7 hold. If, for any $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ and any $\sigma_0 > 0$, the sequence of mixture distributions is $P_t = \mathcal{N}(\Phi_t \boldsymbol{\theta}_0, \sigma_0^2 \Phi_t \Phi_t^\top)$, then for any $\delta \in (0, 1/2]$ and any $\alpha > 0$, with probability at least $1 - 2\delta$, for all $T \geq 1$ simultaneously, the cumulative regret of CMM-UCB and AMM-UCB is bounded by*

$$\Delta_{1:T} \leq \frac{2}{\sqrt{\ln 2}} \max\left\{C, U_{\mathrm{AMM},T-1}\right\} \sqrt{dT\ln\left(1 + \frac{L_2^2 T}{\alpha d}\right)} = \mathcal{O}(d\sqrt{T}\ln(T)),$$

where

$$\begin{aligned}U_{\mathrm{AMM},T-1}^2 \leq \sigma^2 d + \frac{\sigma^2}{\sigma_0^2}\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\right\|_2^2 + \sigma^2 d\ln\left(1 + \frac{(T-1)\sigma_0^2 L_2^2}{\sigma^2 d}\right) + \alpha B_2^2 + 4\sigma^2\ln(1/\delta) \\ + \sigma^2 d\ln\left(1 + \frac{(T-1)L_2^2}{\alpha d}\right) + 2\sqrt{\alpha}\left\|\boldsymbol{\theta}^*\right\|_2 \sqrt{\sigma^2 d\ln\left(1 + \frac{(T-1)L_2^2}{\alpha d}\right) + 2\sigma^2\ln(1/\delta)}.\end{aligned} \tag{B.26}$$

Proof. Following the proof of Theorem 4.9, we obtain (with high probability)

$$\sum_{t=1}^{T} \Delta(a_t) \leq \sqrt{T \sum_{t=1}^{T} 4 \max\left(C^2, R^2_{\mathrm{AMM},t-1}\right) \min\left(1, \|\phi(a_t)\|^2_{(\Phi_{t-1}^\top \Phi_{t-1} + \alpha \boldsymbol{I})^{-1}}\right)}.$$

This time, we use the bound on the radius from Lemma B.16. Let $U_{\mathrm{AMM},T-1}$ denote this bound on the radius (i.e. the square root of the right-hand-side of (B.26)). Also, note that this bound on the radius holds with probability at $1-\delta$. Since $U_{\mathrm{AMM},T-1}$ is monotonically increasing with T, we have

$$\sum_{t=1}^{T} \Delta(a_t) \leq 2 \max\left(C, U_{\mathrm{AMM},T-1}\right) \sqrt{T \sum_{t=1}^{T} \min\left(1, \|\phi(a_t)\|^2_{(\Phi_{t-1}^\top \Phi_{t-1} + \alpha \boldsymbol{I})^{-1}}\right)}.$$

Finally, we use Lemma B.17 to obtain

$$\sum_{t=1}^{T} \Delta(a_t) \leq \frac{2}{\sqrt{\ln 2}} \max\left\{C, U_{\mathrm{AMM},T-1}\right\} \sqrt{dT \ln\left(1 + \frac{L_2^2 T}{\alpha d}\right)}.$$

We used two inequalities that each hold with probability at least $1-\delta$. By a union bound argument, the cumulative regret bound holds with probability at least $1-2\delta$. □

B.5 Additional Experimental Details

In this section, we present some extra information about the experimental setups.

B.5.1 Bayesian Credible Interval Construction

We attempt to construct a Bayesian credible interval that holds with high probability for all rounds $t \geq 0$, which leads to the most fair comparison with CMM-UCB. However, whilst the CMM-UCB confidence set holds with high probability over the random draw of the data $a_1, r_1, a_2, r_2, \ldots$, the Bayesian credible interval holds with high probability over the random draw of $\boldsymbol{\theta}^*$ from a prior (for fixed data $a_1, r_1, a_2, r_2, \ldots$).

For the Bayesian credible interval, we use a Gaussian prior $\boldsymbol{\theta}^* \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. We assume a Gaussian likelihood function, i.e. rewards are of the form $r_t = \phi(a_t)^\top \boldsymbol{\theta}^* + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. The Bayesian posterior for $\boldsymbol{\theta}^*$ is another Gaussian $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, where

$$\boldsymbol{\mu}_t = \boldsymbol{\Sigma}_t \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \Phi_t^\top \boldsymbol{r}_t\right), \qquad \boldsymbol{\Sigma}_t = \left(\frac{1}{\sigma^2} \Phi_t^\top \Phi_t + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}.$$

Using Bayes' rule, at any round t, we have $\boldsymbol{\theta^*} \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. Therefore

$$(\boldsymbol{\mu}_t - \boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{\mu}_t - \boldsymbol{\theta}^*) \sim \chi^2(d),$$

where $\chi^2(d)$ is a chi-squared distribution with d degrees of freedom. Let $Q_d(\cdot)$ be the quantile function of the chi-squared distribution with d degrees of freedom. With probability at least $1-\delta_t$ (over the random draw of $\boldsymbol{\theta}^*$ from $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$), we have

$$(\boldsymbol{\mu}_t - \boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{\mu}_t - \boldsymbol{\theta}^*) \leq Q_d(1-\delta_t). \tag{B.27}$$

Using a union bound argument, if $\delta_t = \frac{6\delta}{(t+1)^2\pi^2}$, then $\sum_{t=0}^{\infty} \delta_t = \delta$, which means (B.27) holds with probability at least $1-\delta$ for all $t \geq 0$. Therefore, if $\boldsymbol{\theta}^* \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, then with high probability, the following credible sets contain $\boldsymbol{\theta}^*$ for all $t \geq 0$ simultaneously.

$$\Theta_t = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \middle| (\boldsymbol{\mu}_t - \boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{\mu}_t - \boldsymbol{\theta}^*) \leq Q_d\left(1 - \frac{6\delta}{(t+1)^2\pi^2}\right) \right\}.$$

The upper limit of the credible interval at a for this credible set (and the one we use in Figure 4.3) is

$$\max_{\boldsymbol{\theta} \in \Theta_t} \left\{ \phi(a)^T \boldsymbol{\theta} \right\} = \phi(a)^T \boldsymbol{\mu}_t + \sqrt{Q_d\left(1 - \frac{6\delta}{(t+1)^2\pi^2}\right)} \sqrt{\phi(a)^\top \boldsymbol{\Sigma}_t \phi(a)}.$$

Appendix C

Appendix for Chapter 5

C.1 Analytic SCMM Confidence Bounds

The value of the SCMM upper confidence bound $\mathrm{UCB}_{\Theta_t^{\ell_1}}(a)$ at a is the solution of the convex program

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \ \phi(a)^\top\boldsymbol{\theta} \quad \text{s.t. } (\Phi_t\boldsymbol{\theta}-\boldsymbol{r}_t)^\top(\Phi_t\boldsymbol{\theta}-\boldsymbol{r}_t)\le R_{\mathrm{MM},t}^2 \quad \text{and} \quad \|\boldsymbol{\theta}\|_1\le B_1. \tag{C.1}$$

This is equivalent to

$$\begin{aligned}\max_{\boldsymbol{\theta}^+,\boldsymbol{\theta}^-\in\mathbb{R}^d} \ \phi(a)^\top(\boldsymbol{\theta}^+-\boldsymbol{\theta}^-) \quad \text{s.t. } & (\Phi_t(\boldsymbol{\theta}^+-\boldsymbol{\theta}^-)-\boldsymbol{r}_t)^\top(\Phi_t(\boldsymbol{\theta}^+-\boldsymbol{\theta}^-)-\boldsymbol{r}_t)\le R_{\mathrm{MM},t}^2 \\ & \text{and } \sum_{i=1}^d(\theta_i^+ + \theta_i^-)\le B_1 \\ & \text{and } \theta_i^+\ge 0 \quad \text{and } \theta_i^-\ge 0\end{aligned} \tag{C.2}$$

By weak Lagrangian duality, the solution of (C.2) is upper bounded by the solution of the dual problem, which is

$$\min_{\substack{a\ge 0,b\ge 0\\ \boldsymbol{c}\ge 0,\boldsymbol{d}\ge 0}} \max_{\boldsymbol{\theta}^+,\boldsymbol{\theta}^-\in\mathbb{R}^d}\{L(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,a,b,\boldsymbol{c},\boldsymbol{d})\},$$

where

$$\begin{aligned}L(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,a,b,\boldsymbol{c},\boldsymbol{d}) = \phi(a)^\top(\boldsymbol{\theta}^+-\boldsymbol{\theta}^-) + a\left[R_{\mathrm{MM},t}^2-(\Phi_t(\boldsymbol{\theta}^+-\boldsymbol{\theta}^-)-\boldsymbol{r}_t)^\top(\Phi_t(\boldsymbol{\theta}^+-\boldsymbol{\theta}^-)-\boldsymbol{r}_t)\right] \\ + b(B_1-\mathbf{1}^\top(\boldsymbol{\theta}^++\boldsymbol{\theta}^-)) + \boldsymbol{c}^\top\boldsymbol{\theta}^+ + \boldsymbol{d}^\top\boldsymbol{\theta}^-,\end{aligned}$$

is the Lagrangian, and a, b, $\boldsymbol{c}$ and $\boldsymbol{d}$ are the Lagrange multipliers. First, we find the maximum $\max_{\boldsymbol{\theta}^+,\boldsymbol{\theta}^-\in\mathbb{R}^d}\{L(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,a,b,\boldsymbol{c},\boldsymbol{d})\}$. We set

$$\nabla_{\boldsymbol{\theta}^+}L = \phi(a) - 2a\Phi_t^\top(\Phi_t(\boldsymbol{\theta}^+-\boldsymbol{\theta}^-)-\boldsymbol{r}_t) - b\mathbf{1} + \boldsymbol{c} = 0 \tag{C.3}$$

$$\nabla_{\boldsymbol{\theta}^-}L = -\phi(a) + 2a\Phi_t^\top(\Phi_t(\boldsymbol{\theta}^+-\boldsymbol{\theta}^-)-\boldsymbol{r}_t) - b\mathbf{1} + \boldsymbol{d} = 0. \tag{C.4}$$

From the KKT condition (or by adding together (C.3) and (C.4)), we obtain

$$2b\mathbf{1} = \boldsymbol{c} + \boldsymbol{d}. \tag{C.5}$$

At the optimal $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ (where $\nabla_{\boldsymbol{\theta}^+} L$ and $\nabla_{\boldsymbol{\theta}^-} L$ are both zero), we have $(\nabla_{\boldsymbol{\theta}^+} L)^\top \boldsymbol{\theta}^+ = 0$ and $(\nabla_{\boldsymbol{\theta}^-} L)^\top \boldsymbol{\theta}^- = 0$. Therefore, from (C.3) and (C.4), we have

$$\phi(a)^\top \boldsymbol{\theta}^+ - 2a(\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t)^\top \Phi_t \boldsymbol{\theta}^+ - b\mathbf{1}^\top \boldsymbol{\theta}^+ + \boldsymbol{c}^\top \boldsymbol{\theta}^+ = 0 \tag{C.6}$$

$$-\phi(a)^\top \boldsymbol{\theta}^- + 2a(\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t)^\top \Phi_t \boldsymbol{\theta}^- - b\mathbf{1}^\top \boldsymbol{\theta}^- + \boldsymbol{d}^\top \boldsymbol{\theta}^- = 0. \tag{C.7}$$

By adding (C.6) and (C.7), we obtain

$$\phi(a)^\top (\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - b\mathbf{1}^\top (\boldsymbol{\theta}^+ + \boldsymbol{\theta}^-) + \boldsymbol{c}^\top \boldsymbol{\theta}^+ + \boldsymbol{d}^\top \boldsymbol{\theta}^- = 2a(\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t)^\top \Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-). \tag{C.8}$$

Using (C.8), the Lagrangian at the optimal $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ (for any fixed values of the Lagrange multipliers) can be re-written as

$$\begin{aligned}
L(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, a, b, \boldsymbol{c}, \boldsymbol{d}) &= aR^2_{\mathrm{MM},t} + bB_1 - a(\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t)^\top (\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t) \\
&\quad + 2a(\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t)^\top \Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) \\
&= aR^2_{\mathrm{MM},t} + bB_1 + a(\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t)^\top \left[2\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - (\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t)\right] \\
&= aR^2_{\mathrm{MM},t} + bB_1 + a(\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t)^\top (\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) + \boldsymbol{r}_t).
\end{aligned} \tag{C.9}$$

Using (C.3), we can find an expression for the difference $(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-)$ at the optimal $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$.

$$\begin{aligned}
\text{(C.3)} &\implies \phi(a) - 2a\Phi_t^\top(\Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) - \boldsymbol{r}_t) - b\mathbf{1} + \boldsymbol{c} = 0 \\
&\implies \Phi_t^\top \Phi_t(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) = \Phi_t^\top \boldsymbol{r}_t + \tfrac{1}{2a}(\phi(a) - b\mathbf{1} + \boldsymbol{c}) \\
&\implies \boldsymbol{\theta}^+ - \boldsymbol{\theta}^- = (\Phi_t^\top \Phi_t)^{-1}\Phi_t^\top \boldsymbol{r}_t + \tfrac{1}{2a}(\Phi_t^\top \Phi_t)^{-1}(\phi(a) - b\mathbf{1} + \boldsymbol{c}).
\end{aligned} \tag{C.10}$$

Combining (C.9) and (C.10), for all valid values of $a, b, \boldsymbol{c}$ and $\boldsymbol{d}$, we have

$$\begin{aligned}
\mathrm{UCB}_{\Theta_t^{\ell_1}}(a) &\le \max_{\boldsymbol{\theta}^+, \boldsymbol{\theta}^- \in \mathbb{R}^d} \{L(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, a, b, \boldsymbol{c}, \boldsymbol{d})\} \\
&= aR^2_{\mathrm{MM},t} + bB_1 + a\left[(\Phi_t(\Phi_t^\top \Phi_t)^{-1}\Phi_t - \boldsymbol{I})\boldsymbol{r}_t + \tfrac{1}{2a}\Phi_t(\Phi_t^\top \Phi_t)^{-1}(\phi(a) - b\mathbf{1} + \boldsymbol{c})\right]^\top \\
&\quad \cdot \left[(\Phi_t(\Phi_t^\top \Phi_t)^{-1}\Phi_t + \boldsymbol{I})\boldsymbol{r}_t + \tfrac{1}{2a}\Phi_t(\Phi_t^\top \Phi_t)^{-1}(\phi(a) - b\mathbf{1} + \boldsymbol{c})\right] \\
&=: g(a, b, \boldsymbol{c}).
\end{aligned}$$

Now we can optimise (minimise) this expression with respect to a. Let

$$\begin{aligned}
\boldsymbol{u} &= \Phi_t(\Phi_t^\top \Phi_t)^{-1}\Phi_t - \boldsymbol{I})\boldsymbol{r}_t \\
\boldsymbol{w} &= \Phi_t(\Phi_t^\top \Phi_t)^{-1}\Phi_t + \boldsymbol{I})\boldsymbol{r}_t \\
\boldsymbol{v} &= \tfrac{1}{2}\Phi_t(\Phi_t^\top \Phi_t)^{-1}(\phi(a) - b\mathbf{1} + \boldsymbol{c}) \\
\implies g(a, b, \boldsymbol{c}) &= aR^2_{\mathrm{MM},t} + bB_1 + a(\boldsymbol{u} + \tfrac{1}{a}\boldsymbol{v})^\top (\boldsymbol{w} + \tfrac{1}{a}\boldsymbol{v}) \\
&= aR^2_{\mathrm{MM},t} + bB_1 + a\boldsymbol{u}^\top \boldsymbol{w} + \tfrac{1}{a}\boldsymbol{v}^\top \boldsymbol{v} + \boldsymbol{v}^\top (\boldsymbol{u} + \boldsymbol{w}).
\end{aligned} \tag{C.11}$$

We set

$$\frac{\mathrm{d}g(a,b,\boldsymbol{c})}{\mathrm{d}a} = R_{\mathrm{MM},t}^2 + \boldsymbol{u}^\top \boldsymbol{w} - \tfrac{1}{a^2}\boldsymbol{v}^\top \boldsymbol{v} = 0. \tag{C.12}$$

Solving for (positive) a, we obtain

$$a = \sqrt{\frac{\boldsymbol{v}^\top \boldsymbol{v}}{R_{\mathrm{MM},t}^2 + \boldsymbol{u}^\top \boldsymbol{w}}}. \tag{C.13}$$

Substituting (C.13) into (C.11), we obtain

$$\begin{aligned}\mathrm{UCB}_{\Theta_t^{\ell_1}}(a) &\le \min_{a\ge 0}\{g(a,b,\boldsymbol{c})\} \\ &= \sqrt{(R_{\mathrm{MM},t}^2 + \boldsymbol{r}_t^\top(\Phi_t(\Phi_t^\top\Phi_t)^{-1}\Phi_t - \boldsymbol{I})\boldsymbol{r}_t)(\phi(a) - b\mathbf{1} + \boldsymbol{c})^\top(\Phi_t^\top\Phi_t)^{-1}(\phi(a) - b\mathbf{1} + \boldsymbol{c})} \\ &\quad + \boldsymbol{r}_t^\top\Phi_t(\Phi_t^\top\Phi_t)^{-1}(\phi(a) - b\mathbf{1} + \boldsymbol{c}) + bB_1. \end{aligned} \tag{C.14}$$

Through the same procedure, one can also deduce that

$$\begin{aligned}\mathrm{LCB}_{\Theta_t^{\ell_1}}(a) &\ge -\sqrt{(R_{\mathrm{MM},t}^2 + \boldsymbol{r}_t^\top(\Phi_t(\Phi_t^\top\Phi_t)^{-1}\Phi_t - \boldsymbol{I})\boldsymbol{r}_t)(\phi(a) + b\mathbf{1} - \boldsymbol{c})^\top(\Phi_t^\top\Phi_t)^{-1}(\phi(a) + b\mathbf{1} - \boldsymbol{c})} \\ &\quad + \boldsymbol{r}_t^\top\Phi_t(\Phi_t^\top\Phi_t)^{-1}(\phi(a) + b\mathbf{1} - \boldsymbol{c}) - bB_1. \end{aligned} \tag{C.15}$$

We can freely set b and $\boldsymbol{c}$ in (C.14), as long as $b \ge 0$, $\boldsymbol{c} \ge 0$ and $2b\mathbf{1} - \boldsymbol{c} \ge 0$ (due to (C.5)). Sadly, there is no closed-form solution for the (constrained) minimum of the right-hand-side of (C.14) with respect to b and $\boldsymbol{c}$. Moreover, it is not clear how to obtain useful upper bounds on $\mathrm{UCB}_{\Theta_t^{\ell_1}}(a)$ through specific choices of b and $\boldsymbol{c}$.

One way to set b and $\boldsymbol{c}$ is as follows. Choose any $b \ge 0$ and set choose $\boldsymbol{c} = b\mathbf{1}$. With these choices, (C.14) becomes

$$\mathrm{UCB}_{\Theta_t^{\ell_1}}(a) \le \sqrt{(R_{\mathrm{MM},t}^2 + \boldsymbol{r}_t^\top(\Phi_t(\Phi_t^\top\Phi_t)^{-1}\Phi_t - \boldsymbol{I})\boldsymbol{r}_t)\phi(a)^\top(\Phi_t^\top\Phi_t)^{-1}\phi(a)} + \boldsymbol{r}_t^\top\Phi_t(\Phi_t^\top\Phi_t)^{-1}\phi(a) + bB_1.$$

This would be (essentially) the exact solution of (C.1) if we removed the ℓ_1 norm constraint in (C.1). This is not very appealing, because we get an upper confidence bound that completely ignores the bound on ℓ_1 norm of $\boldsymbol{\theta}^*$ (which is probably not very good).

There is at least one more way to set b and $\boldsymbol{c}$ such the upper bound in (C.14) is simplified. The constraints $b \ge 0$ and $0 \le \boldsymbol{c} \le 2b\mathbf{1}$ mean that $-b\mathbf{1} + \boldsymbol{c}$ can be any vector $\boldsymbol{v}$ with $\|\boldsymbol{v}\|_\infty \le b$. If we choose $b = \|\phi(a)\|_\infty$, then we can choose $\boldsymbol{c}$ such that $-b\mathbf{1} + \boldsymbol{c} = -\phi(a)$. With this choice, the bound on the UCB becomes

$$\mathrm{UCB}_{\Theta_t^{\ell_1}}(a) \le \|\phi(a)\|_\infty B_1.$$

This upper bound is not very useful, because we could obtain the same upper (or lower) bound on $\phi(a)^\top\boldsymbol{\theta}^*$ by Hölder's inequality.

$$|\phi(a)^\top\boldsymbol{\theta}^*| \le \|\phi(a)\|_\infty \|\boldsymbol{\theta}^*\|_1 \le \|\phi(a)\|_\infty B_1.$$

In other words, we get an upper confidence bound that completely ignores the data-dependent constraint in (C.1).

C.2 Feature Selection Guarantees

Throughout this section, let $J = \text{supp}(\boldsymbol{\theta}^*)$. We begin by stating an alternative version of Theorem 5.10, which allows for threshold levels $0 \leq \tau < m$ (which includes the standard Lasso at $\tau = 0$), but gives a weaker, approximate feature selection guarantee.

Theorem C.1 (Approximate Feature Selection Guarantee)**.** *Suppose that assumptions 5.3, 5.5, 5.6, 5.7 and 5.9 all hold. Choose any $\delta \in (0, 1]$ and any threshold level τ that satisfies $0 \leq \tau < m$. Choose any exploration distribution ρ, such that $\nu_{\min}(\boldsymbol{\Sigma}_\rho) > 0$. Set $\eta = 2\sigma L_\infty \sqrt{2\ln(4d/\delta)/T_1}$. Choose any T_1 such that*

$$T_1 \geq \max\left(\frac{3}{\xi^2}\ln(d), \frac{1}{\xi^2}\ln(1/\delta), \frac{512 s^2 \sigma^2 L_\infty^2 \ln(4d/\delta)}{(m-\tau)^2 \nu_{\min}(\boldsymbol{\Sigma}_\rho)^2}\right), \quad \textit{where} \quad \xi = \min\left(\frac{1}{2}, \frac{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{256 s L_\infty^2}\right).$$

With probability at least $1 - \delta$, we have

$$\text{supp}(\boldsymbol{\theta}^*) \subseteq \text{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}), \quad \textit{and} \quad |\text{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau})| \leq \frac{72 s \nu_{\max}(\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1})}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}.$$

In contrast to Theorem 5.10, Theorem C.1 does not guarantee that the (thresholded) Lasso estimate removes all irrelevant features. To prove Theorem 5.10 and Theorem C.1, we establish several important intermediate results in Section C.2.1 to Section C.2.6. In Section C.2.7 and Section C.2.8, we combine all of these intermediate results.

C.2.1 Part 1: Useful Properties of Lasso

For $S = J$ and large enough η, the Lasso estimate satisfies the constraint that appears in the definition of the compatibility condition (see Equation 5.8)).

Lemma C.2 (Lemma 11.1 of [77])**.** *Suppose that $\eta > \frac{2}{T_1}\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty$. The error $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ satisfies*

$$\|\widehat{\boldsymbol{\theta}}_{J^c} - \boldsymbol{\theta}^*_{J^c}\|_1 \leq 3\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1. \tag{C.16}$$

We now state a bound on $\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty$, which holds in our setting with conditionally sub-Gaussian noise variables (i.e. possibly non-i.i.d. noise).

Lemma C.3. *For any $\delta \in (0, 2]$ and any $T_1 \geq 1$, with probability at least $1 - \delta/2$*

$$\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty \leq \sqrt{2 T_1 \sigma^2 L_\infty^2 \ln(4d/\delta)}.$$

Proof. We can re-write the quantity to be bounded as

$$\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty = \max_{1 \leq i \leq d}\left|\sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t\right|,$$

where $\phi_i(a_t)$ is the i^{th} element of the d-dimensional feature vector $\phi(a_t)$. Since $\epsilon_1, \ldots, \epsilon_{T_1}$ are conditionally σ-sub-Gaussian, for any $\lambda > 0$ and any i, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t\right)\right] &= \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T_1-1} \phi_i(a_t)\epsilon_t\right) \mathbb{E}\left[\exp\left(\lambda \phi_i(a_{T_1})\epsilon_{T_1}\right) | \mathcal{D}_{T_1-1}\right]\right] \\
&\leq \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T_1-1} \phi_i(a_t)\epsilon_t\right) \exp\left(\frac{\lambda^2\sigma^2\phi_i(a_{T_1})^2}{2}\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T_1-1} \phi_i(a_t)\epsilon_t\right) \exp\left(\frac{\lambda^2\sigma^2 L_\infty^2}{2}\right)\right] \\
&\vdots \\
&\leq \exp\left(\frac{T_1\lambda^2\sigma^2 L_\infty^2}{2}\right).
\end{aligned}$$

Using the Chernoff bounding method, for any $\tau > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t > \tau\right) &= \mathbb{P}\left(\exp\left(\lambda \sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t\right) > \exp(\lambda\tau)\right) \\
&\leq \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t\right)\right] \exp(-\lambda\tau) \\
&\leq \exp\left(\frac{T_1\lambda^2\sigma^2 L_\infty^2}{2} - \lambda\tau\right).
\end{aligned}$$

By optimising λ, we obtain

$$\mathbb{P}\left(\sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t > \tau\right) \leq \exp\left(-\frac{\tau^2}{2T_1\sigma^2 L_\infty^2}\right).$$

Setting $\delta = \exp\left(-\frac{\tau^2}{2T_1\sigma^2 L_\infty^2}\right)$, we conclude that for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t \leq \sqrt{2T_1\sigma^2 L_\infty^2 \ln(1/\delta)}.$$

The same procedure can be used to derive the same upper bound on $-\sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t$. By a union bound argument, with probability at least $1 - \delta$, we have

$$\left|\sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t\right| \leq \sqrt{2T_1\sigma^2 L_\infty^2 \ln(2/\delta)}.$$

This holds for any $i \in \{1, \ldots, d\}$, using the union bound once more, with probability at least $1 - \delta$, we have

$$\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty = \max_{1 \leq i \leq d} \left|\sum_{t=1}^{T_1} \phi_i(a_t)\epsilon_t\right| \leq \sqrt{2T_1\sigma^2 L_\infty^2 \ln(2d/\delta)}.$$

□

C.2.2 Part 2: Population Covariance Satisfies Compatibility

For a distribution ρ on $\mathcal{A}$, let $\mathbf{\Sigma}_\rho = \mathbb{E}_{a\sim\rho}[\phi(a)\phi(a)^\top]$.

Lemma C.4. *Suppose we choose ρ such that $\nu_{\min}(\mathbf{\Sigma}_\rho) > 0$. Then for any set of indices $S \subseteq [d]$, we have*

$$\mathbf{\Sigma}_\rho \in C(S, \sqrt{\nu_{\min}(\mathbf{\Sigma}_\rho)})$$

Proof. If we choose $P_\mathcal{A}$ such that $\nu_{\min}(\mathbf{\Sigma}) > 0$, then for all vectors $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\|\boldsymbol{v}\|_2^2 \leq \frac{\boldsymbol{v}^\top \mathbf{\Sigma} \boldsymbol{v}}{\nu_{\min}(\mathbf{\Sigma}_\rho)}.$$

For any $S \subseteq \{1, \ldots, d\}$, we have

$$\|\boldsymbol{v}_S\|_1^2 \leq |S| \, \|\boldsymbol{v}_S\|_2^2 \leq |S| \, \|\boldsymbol{v}\|_2^2 \leq \frac{|S| \boldsymbol{v}^\top \mathbf{\Sigma} \boldsymbol{v}}{\nu_{\min}(\mathbf{\Sigma}_\rho)}.$$

Therefore, for any S, $\mathbf{\Sigma}_\rho \in C(S, \sqrt{\nu_{\min}(\mathbf{\Sigma}_\rho)})$. □

Note that it is possible to choose ρ such that $\nu_{\min}(\mathbf{\Sigma}_\rho) > 0$ if and only if the feature vectors $(\phi(a))_{a\in\mathcal{A}}$ span $\mathbb{R}^d$ (see Remark 3.2. of [76]). This is where the requirement for Assumption 5.9 comes from.

C.2.3 Part 3: Empirical Covariance Satisfies Compatibility

We use Lemma EC.6 of [26] to derive a lower bound on T_1, such that, with high probability, $\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\mathbf{\Sigma}_\rho)/2})$.

Lemma C.5 (Lemma EC.6 of [26])**.** *Suppose $a_1, \ldots, a_{T_1}$ are drawn i.i.d. from ρ. Let $\mathbf{\Sigma}_\rho = \mathbb{E}_{a\sim\rho}[\phi(a)\phi(a)^\top]$ and suppose that $\mathbf{\Sigma}$ satisfies $\nu_{\min}(\mathbf{\Sigma}_\rho) > 0$. For any $T_1 \geq \frac{3}{\xi^2}\ln(d)$,*

$$\mathbb{P}\left(\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C\left(J, \sqrt{\nu_{\min}(\mathbf{\Sigma}_\rho)/2}\right)\right) \geq 1 - \exp(-\xi^2 T_1),$$

where $\xi = \min(\frac{1}{2}, \frac{\nu_{\min}(\mathbf{\Sigma}_\rho)}{256 s L_\infty^2})$.

As done by [92], we can also derive statements that hold with probability at least $1-\delta$ for any $\delta \in (0, 1]$. This introduces another lower bound that T_1 must satisfy (which is obtained by rearranging $1 - \exp(-\xi^2 T_1) \geq 1 - \delta$).

Corollary C.6. *Suppose $a_1, \ldots, a_{T_1}$ are drawn i.i.d. from ρ. Let $\mathbf{\Sigma}_\rho = \mathbb{E}_{a\sim\rho}[\phi(a)\phi(a)^\top]$ and suppose that $\mathbf{\Sigma}$ satisfies $\nu_{\min}(\mathbf{\Sigma}_\rho) > 0$ For any $\delta \in (0, 1]$ and any $T_1 \geq \max(\frac{3}{\xi^2}\ln(d), \frac{1}{\xi^2}\ln(2/\delta))$,*

$$\mathbb{P}\left(\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C\left(J, \sqrt{\nu_{\min}(\mathbf{\Sigma}_\rho)/2}\right)\right) \geq 1 - \delta/2. \tag{C.17}$$

C.2.4 Part 4: Bounds on ℓ_1 Estimation Error and Squared Prediction Error

If $\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$ and $\eta > \frac{2}{T_1}\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\|_\infty$, we can prove bounds on the ℓ_1 estimation error $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1$ and the squared prediction error $\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2$ that depend only logarithmically on d. We use the following inequality for the regularised Lasso estimate $\widehat{\boldsymbol{\theta}}$ (see Equation (5.6)), which is stated in the proof of Theorem 11.1 of [77]. In particular, see Eq. (11.23) on page 298 of [77].

Lemma C.7 ([77])**.** *If $\eta > \frac{2}{T_1}\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\|_\infty$, and $\widehat{\boldsymbol{\theta}}$ is the Lasso estimate, then for any $T_1 \geq 1$ we have*

$$\frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1} \leq \frac{\eta}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + \eta(\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 - \|\widehat{\boldsymbol{\theta}}_{J^c} - \boldsymbol{\theta}^*_{J^c}\|_1) \tag{C.18}$$

We also use the following basic inequality.

Lemma C.8. *For any $x, y \in \mathbb{R}$, we have*

$$2\sqrt{2}xy \leq \frac{1}{2}x^2 + 4y^2.$$

Proof. We aim to find the smallest constant $c > 0$, such that

$$2\sqrt{2}xy \leq \frac{1}{2}x^2 + cy^2.$$

This is equivalent to

$$\frac{1}{2}x^2 + cy^2 - 2\sqrt{2}xy \geq 0. \tag{C.19}$$

We first minimise the LHS with respect to x. By setting the derivative of the LHS with respect to x equal to 0, we obtain

$$\min_{x\in\mathbb{R}} \left\{ \frac{1}{2}x^2 + cy^2 - 2\sqrt{2}xy \right\} = (c-4)y^2.$$

If we choose $c \geq 4$, then (C.19) is satisfied for all $x, y \in \mathbb{R}$. □

We can now prove bounds on the estimation and prediction errors.

Lemma C.9. *If $\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$, $\eta \geq \frac{2}{T_1}\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty$, and $\widehat{\boldsymbol{\theta}}$ is the Lasso estimate, then*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{8s\eta}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}, \tag{C.20}$$

$$\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2 \leq \frac{18s\eta^2 T_1}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}. \tag{C.21}$$

Proof. We start from the inequality in Lemma C.7 and add $\eta\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1$ to both sides to obtain

$$\begin{aligned}\frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1} + \eta\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 &\leq \frac{\eta}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + \eta(\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 - \|\widehat{\boldsymbol{\theta}}_{J^c} - \boldsymbol{\theta}^*_{J^c}\|_1) + \eta\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \\ &= \frac{\eta}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + \eta(\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 - \|\widehat{\boldsymbol{\theta}}_{J^c} - \boldsymbol{\theta}^*_{J^c}\|_1) \\ &\quad + \eta(\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 + \|\widehat{\boldsymbol{\theta}}_{J^c} - \boldsymbol{\theta}^*_{J^c}\|_1) \\ &= \frac{\eta}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + 2\eta\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1.\end{aligned}$$

This can be rearranged to give

$$\frac{\eta}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq 2\eta\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 - \frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1}.$$

Since $\frac{1}{T_1}\Phi_{T_1}^\top\Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$, we can apply the compatibility condition to obtain

$$\frac{\eta}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq 2\sqrt{2}\frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2}{\sqrt{T_1}}\frac{\eta\sqrt{s}}{\sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}} - \frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1}.$$

We then use the inequality in Lemma C.8, to obtain

$$\frac{\eta}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{4s\eta^2}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)} + \frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1} - \frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1} = \frac{4s\eta^2}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}.$$

Multiplying both sides by $2/\eta$, yields the ℓ_1 error bound in the statement of the lemma. To prove the prediction error bound, we start from inequality in Lemma C.7, and obtain

$$\begin{aligned}\frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1} &\leq \frac{\eta}{2}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 + \eta(\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 - \|\widehat{\boldsymbol{\theta}}_{J^c} - \boldsymbol{\theta}^*_{J^c}\|_1) \\ &= \frac{\eta}{2}(\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 + \|\widehat{\boldsymbol{\theta}}_{J^c} - \boldsymbol{\theta}^*_{J^c}\|_1) + \eta(\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 - \|\widehat{\boldsymbol{\theta}}_{J^c} - \boldsymbol{\theta}^*_{J^c}\|_1) \\ &\leq \frac{3\eta}{2}\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1. \qquad \text{(C.22)}\end{aligned}$$

Since $\frac{1}{T_1}\Phi_{T_1}^\top\Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$, we have

$$\frac{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{4s}\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1^2 \leq \frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1} \leq \frac{3\eta}{2}\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1.$$

This can be rearranged into a quadratic inequality for $\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1$.

$$\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1\left(\frac{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{4s}\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 - \frac{3\eta}{2}\right) \leq 0.$$

The solution of this inequality is

$$0 \leq \|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1 \leq \frac{6s\eta}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}.$$

Substituting the upper bound on $\|\widehat{\boldsymbol{\theta}}_J - \boldsymbol{\theta}^*_J\|_1$ into (C.22), we obtain

$$\frac{\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2}{2T_1} \leq \frac{18s\eta^2}{2\nu_{\min}(\boldsymbol{\Sigma}_\rho)}.$$

Multiplying both sides by $2T_1$ yields the prediction error bound in the statement of the lemma. □

C.2.5 Part 5: Support Recovery Under The Minimum Signal Condition

Next, we prove that for suitable values of η (which will later require T_1 to be sufficiently large) and suitable values of the threshold level τ, $\text{supp}(\boldsymbol{\theta}^*) \subseteq \text{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}) \subseteq \text{supp}(\widehat{\boldsymbol{\theta}})$.

Lemma C.10. *If* $0 \leq \tau < m$, $\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$, $\eta \geq \frac{2}{T_1}\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty$, $\eta < \frac{(m-\tau)\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{8s}$, $\boldsymbol{\theta}^*$ *satisfies the minimum signal assumption (Assumption 5.6) and* $\widehat{\boldsymbol{\theta}}$ *is the Lasso estimate in (5.6), then*

$$\text{supp}(\boldsymbol{\theta}^*) \subseteq \text{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}) \subseteq \text{supp}(\widehat{\boldsymbol{\theta}}).$$

Proof. Since $\eta < \frac{(m-\tau)\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{8s}$, we have

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{8s\eta}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)} < m - \tau.$$

Therefore, for all $i \in [d]$, we have

$$\tau - m < \widehat{\theta}_i - \theta^*_i < m - \tau.$$

For all $i \in \text{supp}(\boldsymbol{\theta}^*)$, we have $|\theta^*_i| \geq m$. For each $i \in \text{supp}(\boldsymbol{\theta}^*)$, if $\theta^*_i \geq m$, then

$$\widehat{\theta}_i = \widehat{\theta}_i - \theta^*_i + \theta^*_i \geq \widehat{\theta}_i - \theta^*_i + m > \tau - m + m = \tau.$$

Otherwise, if $\theta^*_i \leq -m$, then

$$\widehat{\theta}_i = \widehat{\theta}_i - \theta^*_i + \theta^*_i \leq \widehat{\theta}_i - \theta^*_i - m < m - \tau - m = -\tau.$$

Therefore, we conclude that for all $i \in \text{supp}(\boldsymbol{\theta}^*)$, we have $|\widehat{\theta}_i| > \tau$. This means that $\text{supp}(\boldsymbol{\theta}^*) \subseteq \text{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau})$ must be satisfied. □

Lemma C.11. *If* $\tau = m/2$, $\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$, $\eta \geq \frac{2}{T_1}\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty$, $\eta < \frac{m\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{16s}$, $\boldsymbol{\theta}^*$ *satisfies the minimum signal assumption (Assumption 5.6) and* $\widehat{\boldsymbol{\theta}}$ *is the Lasso estimate in (5.6), then*

$$\operatorname{supp}(\boldsymbol{\theta}^*) = \operatorname{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}).$$

Proof. Since $\eta < \frac{m\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{16s}$, we have

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_1 \leq \frac{8s\eta}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)} < m/2 = \tau.$$

Therefore, for all $i \in [d]$, we have

$$-m/2 < \widehat{\theta}_i - \theta_i^* < m/2.$$

For all $i \in \operatorname{supp}(\boldsymbol{\theta}^*)$, we have $|\theta_i^*| \geq m$. For each $i \in \operatorname{supp}(\boldsymbol{\theta}^*)$, if $\theta_i^* \geq m$, then

$$\widehat{\theta}_i = \widehat{\theta}_i - \theta_i^* + \theta_i^* \geq \widehat{\theta}_i - \theta_i^* + m > -m/2 + m = \tau.$$

Otherwise, if $\theta_i^* \leq -m$, then

$$\widehat{\theta}_i = \widehat{\theta}_i - \theta_i^* + \theta_i^* \leq \widehat{\theta}_i - \theta_i^* - m < m/2 - m = -\tau.$$

Therefore, for all $i \in \operatorname{supp}(\boldsymbol{\theta}^*)$, we have $|\widehat{\theta}_i| > \tau$, which means $i \in \operatorname{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau})$.

For all $i \notin \operatorname{supp}(\boldsymbol{\theta}^*)$, we have $\theta_i^* = 0$, which means that

$$|\widehat{\theta}_i| = |\widehat{\theta}_i - \theta_i^*| < m/2 = \tau.$$

Therefore, for all $i \notin \operatorname{supp}(\boldsymbol{\theta}^*)$, we have $|\widehat{\theta}_i| < \tau$, which means $i \notin \operatorname{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau})$. We conclude that

$$\operatorname{supp}(\boldsymbol{\theta}^*) = \operatorname{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}).$$

□

C.2.6 Part 6: Sparsity of Lasso Estimate

In the proof their Theorem 5.2, [76] prove an order $\mathcal{O}(s)$ upper bound on the number of features selected by Lasso for sufficiently large T_1 and η. This part of the proof closely follows the proof of Eq. (7.9) in Theorem 7.2 of [29], which states a similar upper bound. We closely follow these proofs to prove another $\mathcal{O}(s)$ upper bound on the number of features selected by Lasso. Since thresholding can only reduce the number of selected features, this is also an upper bound on the number of features selected by the thresholded Lasso estimate. The main difference between the proof of [76] and our own proof is that [76] uses a bound on $\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2$ that holds under a restricted eigenvalue condition (see Condition A.1 of [76]), whereas we use our prediction error bound from Lemma C.9, which requires that $\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1}$ satisfies the compatibility condition.

Lemma C.12. *If* $\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$, $\eta \geq \frac{2}{T_1}\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty$, *and* $\widehat{\boldsymbol{\theta}}$ *is the Lasso estimate in (5.6), then*

$$|\mathrm{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{J}_\tau})| \leq |\mathrm{supp}(\widehat{\boldsymbol{\theta}})| \leq \frac{72\nu_{\max}(\frac{1}{T_1}\Phi_{T_1}\Phi_{T_1}^\top)s}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}.$$

Proof. The requirement

$$\eta \geq \frac{2}{T_1}\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty = \frac{2}{T_1}\max_{1\leq i\leq d}\left|\sum_{t=1}^{T_1}\phi_i(a_t)\epsilon_t\right|,$$

ensures that

$$\forall i \in \{1,\ldots,d\}, \quad -\frac{\eta}{2} \leq \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)\epsilon_t \leq \frac{\eta}{2}.$$

From the Karush-Kuhn-Tucker (KKT) condition, the solution $\widehat{\boldsymbol{\theta}}$ of the Lasso optimisation problem in (5.6) satisfies

$$\forall \widehat{\theta}_i \neq 0, \quad \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)(r_t - \phi(a_t)^\top\widehat{\boldsymbol{\theta}}) = \mathrm{sign}(\widehat{\theta}_i)\eta.$$

Therefore, for all i where $\widehat{\theta}_i \neq 0$, we have

$$\begin{aligned}\frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)(\phi(a_t)^\top\boldsymbol{\theta}^* - \phi(a_t)^\top\widehat{\boldsymbol{\theta}}) &= \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)(r_t - \phi(a_t)^\top\widehat{\boldsymbol{\theta}}) - \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)\epsilon_t \\ &= \mathrm{sign}(\widehat{\theta}_i)\eta - \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)\epsilon_t.\end{aligned}$$

If $\mathrm{sign}(\widehat{\theta}_i) = 1$, then

$$\mathrm{sign}(\widehat{\theta}_i)\eta - \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)\epsilon_t = \eta - \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)\epsilon_t \in [\eta/2, 3\eta/2].$$

Otherwise, if $\mathrm{sign}(\widehat{\theta}_i) = -1$, then

$$\mathrm{sign}(\widehat{\theta}_i)\eta - \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)\epsilon_t = -\eta - \frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)\epsilon_t \in [-3\eta/2, -\eta/2].$$

Combining everything so far, we have established that for all $i \in \{1,\ldots,d\}$ where $\widehat{\theta}_i \neq 0$, we have

$$\left|\frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)(\phi(a_t)^\top\boldsymbol{\theta}^* - \phi(a_t)^\top\widehat{\boldsymbol{\theta}})\right| \geq \eta/2.$$

Therefore, we have

$$\frac{1}{T_1^2}\sum_{i=1}^{d}\left(\sum_{t=1}^{T_1}\phi_i(a_t)(\phi(a_t)^\top\boldsymbol{\theta}^* - \phi(a_t)^\top\widehat{\boldsymbol{\theta}})\right)^2 \geq \sum_{i:\widehat{\theta}_i\neq 0}\left(\frac{1}{T_1}\sum_{t=1}^{T_1}\phi_i(a_t)(\phi(a_t)^\top\boldsymbol{\theta}^* - \phi(a_t)^\top\widehat{\boldsymbol{\theta}})\right)^2$$
$$\geq |\mathrm{supp}(\widehat{\boldsymbol{\theta}})|\eta^2/4.$$

We also have

$$\frac{1}{T_1^2}\sum_{i=1}^{d}\left(\sum_{t=1}^{T_1}\phi_i(a_t)(\phi(a_t)^\top\boldsymbol{\theta}^* - \phi(a_t)^\top\widehat{\boldsymbol{\theta}})\right)^2 = \frac{1}{T_1^2}(\Phi_{T_1}\boldsymbol{\theta}^* - \Phi_{T_1}\widehat{\boldsymbol{\theta}})^\top\Phi_{T_1}\Phi_{T_1}^\top(\Phi_{T_1}\boldsymbol{\theta}^* - \Phi_{T_1}\widehat{\boldsymbol{\theta}})$$
$$\leq \frac{\nu_{\max}(\frac{1}{T_1}\Phi_{T_1}\Phi_{T_1}^\top)}{T_1}\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2.$$

Using our prediction error bound in (C.21), we obtain

$$|\mathrm{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau})| \leq |\mathrm{supp}(\widehat{\boldsymbol{\theta}})| \leq \frac{4\xi_{\max}(\frac{1}{T_1}\Phi_{T_1}\Phi_{T_1}^\top)}{\eta^2 T_1}\|\Phi_{T_1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2 \leq \frac{72\nu_{\max}(\frac{1}{T_1}\Phi_{T_1}\Phi_{T_1}^\top)s}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}.$$

□

C.2.7 Proof of Theorem 5.10

Proof. We first verify that, with probability at least $1-\delta$, the conditions of Lemma C.11 are satisfied.

Using Lemma C.3, for any $\delta \in (0,1]$ and any $T_1 \geq 1$, with probability at least $1-\delta/2$

$$\eta = \frac{2\sigma L_\infty\sqrt{2\ln(4d/\delta)}}{\sqrt{T_1}} = \frac{2}{T_1}\sqrt{2T_1\sigma^2L_\infty^2\ln(4d/\delta)} \geq \frac{2}{T_1}\left\|\Phi_{T_1}^\top\boldsymbol{\epsilon}_{T_1}\right\|_\infty.$$

Since we have chosen T_1 such that

$$T_1 > \max\left(\frac{3}{\xi^2}\ln(d), \frac{1}{\xi^2}\ln(1/\delta), \frac{2048s^2\sigma^2L_\infty^2\ln(4d/\delta)}{m^2\nu_{\min}(\boldsymbol{\Sigma}_\rho)^2}\right), \quad \text{where} \quad \xi = \min\left(\frac{1}{2}, \frac{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{256sL_\infty^2}\right),$$

The conditions of Corollary C.6 are satisfied. Therefore, with probability at $1-\delta/2$, we have $\frac{1}{T_1}\Phi_{T_1}^\top\Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$. Finally, since our choice of T_1 satisfies

$$\sqrt{T_1} > \sqrt{\frac{2048s^2\sigma^2L_\infty^2\ln(4d/\delta)}{m^2\nu_{\min}(\boldsymbol{\Sigma}_\rho)^2}} = \frac{32s\sigma L_\infty\sqrt{2\ln(4d/\delta)}}{m\nu_{\min}(\boldsymbol{\Sigma}_\rho)},$$

We have

$$\eta = \frac{2\sigma L_\infty\sqrt{2\ln(4d/\delta)}}{\sqrt{T_1}} < 2\sigma L_\infty\sqrt{2\ln(4d/\delta)}\frac{m\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{32s\sigma L_\infty\sqrt{2\ln(4d/\delta)}} = \frac{m\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{16s}.$$

Combining everything so far, and using the union bound, with probability at least $1-\delta$, the conditions of Lemma C.11 all hold. Therefore, using Lemma C.11, we have

$$\mathrm{supp}(\boldsymbol{\theta}^*) = \mathrm{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}).$$

□

C.2.8 Proof of Theorem C.1

Proof. We first verify that, with probability at least $1-\delta$, the conditions of Lemma C.12 and Lemma C.10 are satisfied.

Using Lemma C.3, for any $\delta \in (0,1]$ and any $T_1 \geq 1$, with probability at least $1-\delta/2$

$$\eta = \frac{2\sigma L_\infty \sqrt{2\ln(4d/\delta)}}{\sqrt{T_1}} = \frac{2}{T_1}\sqrt{2T_1\sigma^2 L_\infty^2 \ln(4d/\delta)} \geq \frac{2}{T_1}\left\|\Phi_{T_1}^\top \boldsymbol{\epsilon}_{T_1}\right\|_\infty.$$

Since we have chosen T_1 such that

$$T_1 > \max\left(\frac{3}{\xi^2}\ln(d), \frac{1}{\xi^2}\ln(1/\delta), \frac{512 s^2\sigma^2 L_\infty^2 \ln(4d/\delta)}{(m-\tau)^2 \nu_{\min}(\boldsymbol{\Sigma}_\rho)^2}\right), \quad \text{where} \quad \xi = \min\left(\frac{1}{2}, \frac{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{256 s L_\infty^2}\right),$$

The conditions of Corollary C.6 are satisfied. Therefore, with probability at $1-\delta/2$, we have $\frac{1}{T_1}\Phi_{T_1}^\top \Phi_{T_1} \in C(J, \sqrt{\nu_{\min}(\boldsymbol{\Sigma}_\rho)/2})$. Finally, since our choice of T_1 satisfies

$$\sqrt{T_1} > \sqrt{\frac{512 s^2\sigma^2 L_\infty^2 \ln(4d/\delta)}{(m-\tau)^2\nu_{\min}(\boldsymbol{\Sigma}_\rho)^2}} = \frac{16 s\sigma L_\infty \sqrt{2\ln(4d/\delta)}}{(m-\tau)\nu_{\min}(\boldsymbol{\Sigma}_\rho)},$$

We have

$$\eta = \frac{2\sigma L_\infty\sqrt{2\ln(4d/\delta)}}{\sqrt{T_1}} < 2\sigma L_\infty \sqrt{2\ln(4d/\delta)} \frac{(m-\tau)\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{16 s\sigma L_\infty \sqrt{2\ln(4d/\delta)}} = \frac{(m-\tau)\nu_{\min}(\boldsymbol{\Sigma}_\rho)}{8s}.$$

Combining everything so far, and using the union bound, with probability at least $1-\delta$, the conditions of Lemma C.12 and Lemma C.10 all hold. Using Lemma C.12, we have

$$|\mathrm{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{J}_\tau})| \leq |\mathrm{supp}(\widehat{\boldsymbol{\theta}})| \leq \frac{72\xi_{\max}(\frac{1}{T_1}\Phi_{T_1}\Phi_{T_1}^\top)s}{\nu_{\min}(\boldsymbol{\Sigma}_\rho)}. \tag{C.23}$$

Using Lemma C.10, we have

$$\mathrm{supp}(\boldsymbol{\theta}^*) \subseteq \mathrm{supp}(\widehat{\boldsymbol{\theta}}_{\widehat{S}_\tau}) \subseteq \mathrm{supp}(\widehat{\boldsymbol{\theta}}).$$

□

Glossary

Notation	Description
$\mathbb{I}\{A\}$	Indicator for an event A
π	Policy
b	Behaviour policy
t	Time
s_t	State
a_t	Action
r_t	Reward
Π	Policy class
$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$\mathcal{R}$	Reward space
D	Data set
$\mathcal{D}$	σ-algebra (generated by a data set)
$\boldsymbol{\theta}$	Parameter vector
ϕ	Feature map
Φ	Design matrix
Q	"Posterior" distribution
P	"Prior" or mixture distribution

Acronym	Description
i.i.d.	Independently and identically distributed
PAC	Probably Approximately Correct
MAB	Multi-armed bandit
CB	Contextual bandit
UCB	Upper Confidence Bound
LCB	Lower Confidence Bound

List of Figures

List of Tables

Curriculum Vitae

Education

Technische Universität Darmstadt (Darmstadt, Germany) 2020 - 2023
PhD (Dr. rer. nat.)

University of Southampton (Southampton, UK) 2018 - 2019
MSc in Artificial Intelligence

University of Bristol (Bristol, UK) 2014 - 2017
BSc in Mathematics

Employment

PhD Student, Bosch Center for AI, Renningen, Germany, 2020 - 2023

Data Analyst, The Stars Group, Isle of Man, 2018

Honours and Awards

NeurIPS 2023 Oral (September 2023)
Our paper "Improved Algorithms for Stochastic Linear Bandits Using Tail Bounds for Martingale Mixtures" was selected for an oral presentation at the Conference on Neural Information Processing Systems (NeurIPS) (top 0.5% of over 12000 submissions).

Richard Newitt Bursary (February 2019)
Awarded by the Faculty of Engineering and Physical Sciences at the University of Southampton for achieving the best marks in the first semester.

Workshops and Events

PAC-Bayes Meets Interactive Learning (July 2023)
Workshop at at the International Conference on Machine Learning (ICML).

Publication List

Journal Papers

1. **Flynn, H.**, Reeb, D., Kandemir, M. and Peters, J., (2023). PAC-Bayes Bounds for Bandit Problems: A Survey and Experimental Comparison, *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

2. **Flynn, H.**, Reeb, D., Kandemir, M. and Peters, J., (2022). PAC-Bayesian Lifelong Learning for Multi-armed Bandits, *Data Mining and Knowledge Discovery*, Springer

Conference Papers

1. **Flynn, H.**, Reeb, D., Kandemir, M. and Peters, J., (2023). Improved Algorithms for Stochastic Linear Bandits Using Tail Bounds for Martingale Mixtures, *Conference on Neural Information Processing Systems (NeurIPS)*.