

Design and Analysis of Optimal and Minimax Robust Sequential Hypothesis Tests

Vom Fachbereich 18
Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von
Dipl.-Ing. Michael Fauß
geboren am 04.02.1986 in Seeheim-Jugenheim

Referent:	Prof. Dr. Abdelhak M. Zoubir
Korreferent:	Prof. Dr. H. Vincent Poor
Tag der Einreichung:	08.02.2016
Tag der mündlichen Prüfung:	01.03.2016

D 17
Darmstadt, 2016

Acknowledgments

I would like to thank all the people that supported me during my doctoral study and contributed to this thesis in various ways.

I sincerely thank Prof. Abdelhak Zoubir for his sustained support, trust and academic guidance and for giving me the time and freedom to pursue my research interests. I thank Prof. Vincent Poor for being my co-supervisor and providing valuable feedback on my work. I further thank Prof. Heinz Koepll as well as Prof. Oliver Boine-Frankenheim for being on of my PhD committee.

A big thank you goes out to all current and former members and visitors of the Signal Processing Group: Sara Al-Sayed, Mouhammad Alhumaidi, Mark Ryan Balthasar, Patricia Binder, Nevine Demitri, Raquel Fandos, Hauke Fath, Tai Fei, Gökhan Güll, Jürgen Hahn, Khadidja Hamaidi, Phillip Heidenreich, Di Jin, Sahar Khawatmi, Renate Koschella, Michael Lang, Stefan Leier, Michael Leisnner, Zhihua Lu, Toufik Mouchini, Ahmed Moustafa, Michael Muma, Ivana Perna, Dominik Reinhard, Simon Rosenkranz, Tim Schäck, Fiky Suratman, Ann-Kathrin Seifert, Waqas Sharif, Adrian Šošić, Wassim Suleiman, Gebremichael Teame, Freweyni Teklehaymanot, Christian Weiβ, Feng Yin and everyone else whose company I enjoyed over the last years.

I thank Prof. Wolfgang Utschick for sparking my interest in signal processing and Prof. Roy Howard for many inspiring conversations.

I thank my family, Peter, Monika and Tobias, for always being there and for supporting me.

Thank you, Sara, for being who you are, for your charming company and for making me enjoy the time of my PhD candidature inside and outside the office.

Darmstadt, 24.05.2016

Kurzfassung

In dieser Dissertation wird ein Verfahren zum Entwurf und zur Analyse von optimalen und minimax robusten sequentiellen Hypothesentests entwickelt. Die Arbeit umfasst sowohl einen umfassenden Theorieteil als auch Algorithmen zur praktischen Implementierung robuster sequentieller Tests.

Nach einer Einführung in die grundlegenden Konzepte der sequentiellen Analyse und der Stoppzeit-Theorie, wird der optimale sequentielle Test für stochastische Prozesse mit Markov-Darstellung hergeleitet. Zu diesem Zweck wird das Problem des Entwurfs sequentieller Tests in ein Problem der Ermittlung optimaler Stoppzeiten überführt, dessen Kostenfunktion sich aus der zu erwartenden Laufzeit und den gewichteten Fehlerwahrscheinlichkeiten des Tests zusammensetzt. Aufbauend auf dieser Formulierung kann die Strategie eines mit minimalen Kosten verbundenen sequentiellen Tests durch Lösung einer nichtlinearen Integralgleichung bestimmt werden. In der Dissertation wird nachgewiesen, dass die partiellen Ableitungen der Kostenfunktion des optimalen Tests bis auf eine konstante Skalierung mit den Fehlerwahrscheinlichkeiten des zugrunde liegenden sequentiellen Tests übereinstimmen. Mit Hilfe dieses Zusammenhangs lässt sich der Entwurf optimaler sequentieller Test mit beschränkten Fehlerwahrscheinlichkeiten auf die Lösung einer Integralgleichung zurückführen, deren Lösungsfunktion zusätzliche Bedingungen an ihre partiellen Ableitungen erfüllen muss. Es wird weiterhin gezeigt, dass die gesuchte Lösungsfunktion sich mittels etablierter Methoden der linearen Optimierung bestimmen lässt, ohne die partiellen Ableitungen explizit zu berechnen. Das Verfahren wird anhand mehrerer numerischer Beispiele veranschaulicht.

In der zweiten Hälfte der Dissertation wird der Entwurf minimax robuster sequentieller Tests behandelt. Zunächst werden das Minimax-Prinzip und ein allgemeines Modell für Unsicherheiten in den Verteilungen eingeführt und erläutert. Danach werden hinreichende Bedingungen dafür hergeleitet, dass gegebene Wahrscheinlichkeitsverteilungen am ungünstigsten sind, das heißt, zu der größten mittleren Laufzeit und den größten Fehlerwahrscheinlichkeiten eines sequentiellen Tests führen. Durch Zusammenführung der Ergebnisse zu optimalen sequentiellen Tests und ungünstigsten Verteilungen ergeben sich hinreichende Bedingungen für die minimax Optimalität sequentieller Tests unter allgemeinen Verteilungsunsicherheiten. Des Weiteren wird die Kostenfunktion des minimax optimalen Tests als ein konkav-exponentielles statistisches Ähnlichkeitsmaß identifiziert und die ungünstigsten Verteilungen als diejenigen, welche die größte Ähnlichkeit bezüglich dieses Maßes aufweisen. Als konkretes Beispiel für nichtparametrische Verteilungsunsicherheiten wird das Dichte-Band-Modell (density band model) eingeführt. Die sich aus diesem Modell ergebenden ungünstigsten Verteilungen

werden zunächst in impliziter Form hergeleitet. Basierend auf der implizierten Darstellung wird ein Algorithmus zu ihrer numerische Berechnung entwickelt. Schließlich wird der minimax robuste sequentielle Test unter Unsicherheiten des Dichte-Band-Typs hergeleitet, welcher die für minimax Verfahren charakteristische Eigenschaft eines maximal flachen Profils der Laufzeiten und Fehlerwahrscheinlichkeiten über dem Zustandsraum aufweist. Ein numerisches Beispiel für einen minimax optimalen sequentiellen Test schließt die Dissertation ab.

Abstract

In this dissertation, a framework for the design and analysis of optimal and minimax robust sequential hypothesis tests is developed. It provides a coherent theory as well as algorithms for the implementation of optimal and minimax robust sequential tests in practice.

After introducing some fundamental concepts of sequential analysis and optimal stopping theory, the optimal sequential test for stochastic processes with Markovian representations is derived. This is done by formulating the sequential testing problem as an optimal stopping problem whose cost function is given by a weighted sum of the expected run-length and the error probabilities of the test. Based on this formulation, a cost minimizing testing policy can be obtained by solving a nonlinear integral equation. It is then shown that the partial generalized derivatives of the optimal cost function are, up to a constant scaling factor, identical to the error probabilities of the cost minimizing test. This relation is used to formulate the problem of designing optimal sequential tests under constraints on the error probabilities as a problem of solving an integral equation under constraints on the partial derivatives of its solution function. Finally, it is shown that the latter problem can be solved by means of standard linear programming techniques without the need to calculate the partial derivatives explicitly. Numerical examples are given to illustrate this procedure.

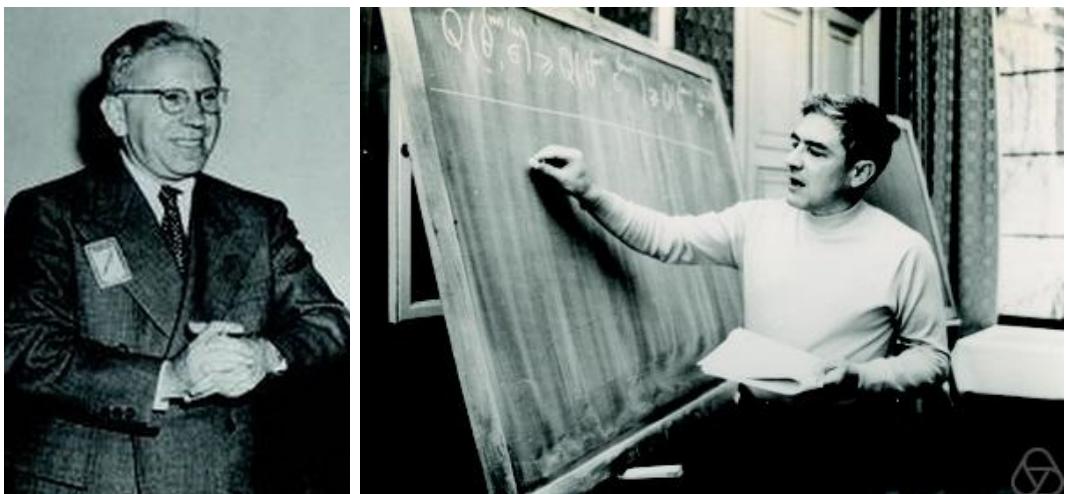
The second half of the dissertation is concerned with the design of minimax robust sequential hypothesis tests. First, the minimax principle and a general model for distributional uncertainties is introduced. Subsequently, sufficient conditions are derived for distributions to be least favorable with respect to the expected run-length and error probabilities of a sequential test. Combining the results on optimal sequential tests and least favorable distributions yields a sufficient condition for a sequential test to be minimax optimal under general distributional uncertainties. The cost function of the minimax optimal test is further identified as a convex statistical similarity measure and the least favorable distributions as the distributions that are most similar with respect to this measure. In order to obtain more specific results, the density band model is introduced as an example for a nonparametric uncertainty model. The corresponding least favorable distributions are stated in an implicit form, based on which a simple algorithm for their numerical calculation is derived. Finally, the minimax robust sequential test under density band uncertainties is discussed and shown to admit the characteristic minimax property of a maximally flat performance profile over its state space. A numerical example for a minimax optimal sequential test completes the dissertation.

Contents

1	Introduction	1
1.1	Motivation	2
1.2	State-of-the-Art	3
1.3	Aims, Contributions and Overview	3
1.4	Publications	5
2	Fundamentals I	7
2.1	Statistical Hypothesis Tests and Decision Rules	7
2.2	Discrete-Time Stochastic Processes	10
2.3	Discrete-Time Optimal Stopping	12
2.3.1	The Finite-Horizon Case	14
2.3.2	The Infinite-Horizon Case	16
2.4	Sequential Hypothesis Tests	16
2.4.1	The Sequential Probability Ratio Test	17
2.4.2	General Sequential Hypothesis Tests	18
2.4.3	Performance Analysis	21
3	Optimal Sequential Hypothesis Tests	27
3.1	Problem Formulation	27
3.2	Solution of the Optimal Stopping Problem	29
3.3	Properties of the Cost Function ρ_λ	33
3.4	Design of Optimal Sequential Tests	38
3.4.1	Newton-like Methods	39
3.4.2	Linear Programming	42
3.4.3	Remarks	44
3.5	Examples and Numerical Results	45
3.5.1	Mean Shifted Gaussian Distributions	46
3.5.2	Bernoulli Distributions	53
3.5.3	Observable Markov Chain	56
3.5.4	Gaussian AR(1) Process	57
3.6	Summary	61
4	Fundamentals II	63
4.1	Statistical Robustness and the Minimax Principle	63
4.2	Saddle Points and Minimax Optimality	66
4.3	Uncertainty Sets	67
4.4	Distributions and Uncertainty Sets Depending on the Test Statistic	69

4.5	Aspects of Convex Optimization in Banach Spaces	72
4.5.1	The Dual Space	72
4.5.2	Fréchet Differentials and Subdifferentials	73
4.5.3	First Order Optimality Conditions	74
5	Minimax Robust Sequential Hypothesis Tests	77
5.1	Problem Formulation	77
5.2	Least Favorable Distributions	79
5.3	Minimax Optimal Sequential Tests	82
5.4	Statistical Similarity Measures	85
5.5	The Density Band Uncertainty Model	89
5.5.1	Most Similar Distributions	91
5.5.2	Examples	93
5.5.3	Calculation of the Most Similar Distributions	96
5.6	Minimax Robust Sequential Tests under Density Band Uncertainties	99
5.7	Test Design	101
5.8	Examples and Numerical Results	103
5.9	Summary	109
6	Conclusions and Outlook	111
6.1	Tests for Multiple Hypotheses	111
6.2	Asymptotic Results	112
6.3	Investigation of Special Cases	113
6.4	Existence of Minimax Optimal Tests	113
6.5	Comparison to Alternative Procedures	113
Appendix		115
A.1	Recursive Definition of Performance Measures	115
A.2	Proof of Theorem 3	116
A.3	Proof of Theorem 5	116
A.4	Proof of Corollary 3	118
A.5	Proof of Lemma 1	119
A.6	Proof of Theorem 6	120
A.7	Proof of Theorem 7	121
A.8	Proof of Theorem 8	122
A.9	Proof of Theorem 9	125
A.10	Proof of Corollary 5	126
A.11	Proof of Theorem 10	126
A.12	Enforcing Equality in the Constraint of the Relaxed Linear Program	127
A.13	Proof of Theorem 11	128

A.14 Proof of Theorem 12	128
A.15 Proof of Theorem 13	129
A.16 Proof of Theorem 14	130
A.17 Proof that ρ_π induces an f -similarity	135
A.18 Proof of Theorem 15	135
A.19 Proof of Corollary 9	137
List of Symbols	139
References	141
Curriculum Vitae	153



Abraham Wald (left) and Peter J. Huber (right)¹

Chapter 1

Introduction

The dissertation at hand is based on two mainstays, namely, *sequential* and *robust* statistical hypothesis testing. Interestingly, the birth of each field can be traced back to a particular time and person. Abraham Wald [Wol52] developed the sequential probability ratio test in the 1940s and published the results, which had been classified during the war, in his seminal book on sequential analysis in 1947 [Wal47]. Peter J. Huber [BK08], who contributed invaluable to transforming the concept of robustness from a vague idea into a recognized branch of mathematical statistics, presented his robust version of the probability ratio test in 1965 [Hub65].

The goal of this work is to combine both approaches to statistical hypothesis testing, which can be viewed as having somewhat complementary objectives, into a consistent framework of minimax optimal sequential hypothesis testing. This includes, on the one hand, a thorough mathematical theory that is well embedded in existing results and concepts, and, on the other hand, efficient algorithms and numerical implementations of optimal and minimax robust sequential tests in practice.

The advantages and challenges of bringing together sequential and robust methods are briefly discussed in the next section. It is followed by an overview of existing work on optimal and robust sequential tests as well as a more detailed statement of the aim and the contributions of this dissertation.

¹Both images are covered by the Creative Commons license.

1.1 Motivation

Sequential methods are typically used to increase the efficiency of a statistical inference procedure. Wald showed that, under ideal conditions, a sequential hypothesis test can achieve the same reliability as its fixed sample size counterpart with roughly half the number of samples. In general, the ability to decide about the necessary number of samples based on the data collected so far makes sequential procedures more flexible than standard procedures, which are designed under the assumption that the number of observations is given and cannot be increased. Naturally, this additional degree of freedom can only be leveraged if the observations are indeed generated sequentially. A comprehensive overview of applications for which this is the case can be found in the first Chapters of [GS91] and [PH09]. The list includes medical diagnosis [BLS13], environment monitoring [Pet98], quality control [YZ99], hazard detection [BMG⁺10], image processing [BEG81], spectrum sensing [SJ15] and many more. In general, sequential methods offer an increased efficiency if obtaining new observations is possible, but costly in some sense, and if the data indeed follows the assumed model. If the latter is not the case, sequential tests can end up using significantly more samples than regular tests [Wal47].

This effect motivates the use of robust methods. Instead of increasing the efficiency, robust inference procedures sacrifice some efficiency under ideal conditions in order to be less sensitive to deviations from the ideal case. This approach is based on the idea that an inference method should continue to work reasonably well, even if the underlying model holds only approximately or only for a subset of the data. This is not the case for strictly model based methods, which are tuned to perform optimally under ideal conditions, but may suffer severe performance degradation if the data is even slightly corrupted [Hub81]. Thus, robust methods form the middle ground between parametric and nonparametric approaches. The list of applications is long and includes voice activity detection [KMZ15] and seismic data analysis [Cla00] to name just two. See [KP85, ZKCM12, AMR15] for many more examples.

The idea of robust sequential hypothesis testing is to combine the benefits of both approaches. Ideally, a robust sequential test is supposed to be *fast and reliable*, i.e., it requires fewer observations than a fixed sample size test and at the same time works reliably under moderate model mismatches. In this sense, both concepts complement each other in two ways: By sequentially performing a robust test, the loss in nominal efficiency can be compensated. By robustly performing a sequential test, its sensitivity to model mismatches can be reduced. Exploring and exploiting this synergy is the main motivation for the work that constitutes this dissertation.

1.2 State-of-the-Art

The amount of existing literature on both sequential and robust hypothesis testing is vast and grows steadily. General overviews of sequential hypothesis testing and related topics can be found in [Wal47, Sie85, GS91, PH09, TNB14]. A summary of results in robust statistics is given in, for example, [Hub81, MMY06].

On the other hand, the literature on minimax optimal sequential hypothesis testing is rather scarce. To the best of the author's knowledge, the general design of strictly minimax optimal sequential hypothesis tests has not been treated in the literature yet. Some of the earliest results in the field can be found in [Mau57], where a very specific minimax problem concerning a test for the mean of two normal distributions is solved, and [DeG60], where a minimax optimal procedure for the detection of a mean shift in Brownian motion is derived. The latter has been further generalized in [Sch87]. The vast majority of the literature that has been published over the last decades deals with the design of *asymptotically* minimax sequential tests. Asymptotic results exist, for example, for tests of distributions of the exponential family [Hol75], the presence of a signal in additive noise [ESV79], multiple distributions with unknown parameters [BD08b] and discrete distributions [FT12a]. An approach to robust sequential testing based on adaptive nonlinearities is suggested in [VK88], some application specific procedures are given in [CCF95, Vos01]. Closely related to the problem of minimax sequential tests is the problem of minimax quickest change detection, which was studied in [BD08a, UVM11, FT12b, BV15].

A noteworthy, but lesser-known exception from the asymptotic approach to minimax sequential hypothesis testing is the work presented in [Kha02]. It is concerned with the design of strictly minimax optimal tests for discrete distributions and is probably closest in spirit to this work.

1.3 Aims, Contributions and Overview

The aim of this dissertation is to develop a coherent framework for the design and analysis of optimal and minimax robust sequential hypothesis tests. In this sense, the results presented in the upcoming chapters are intended to provide a self-contained theory that does not require the reader to have in-depth prior knowledge of sequential analysis or particular methods for the design of sequential tests.

This approach makes it necessary to occasionally recast existing results in order to better fit the context of this work. Some well known results of optimal stopping and sequential testing theory are restated, but in general the overlap with existing work was kept at a minimum. Known results are clearly marked as such.

An important aspect of this dissertation is that the presented results are supposed to build on each other and form a coherent chain of arguments. Most mathematical concepts used in the derivations, such as Fredholm integral equations, generalized derivatives and convex similarity measures, emerge naturally in the process of solving the sequential testing problem. The aim is not to give an example for the application of a particular method to an open problem in sequential analysis, but to arrive at an appropriate procedure of solution by close inspection of the problem at hand.

The presented framework is supposed to offer both a thorough theory of optimal and minimax sequential hypothesis tests as well as practical algorithms for their implementation. In contrast to most existing algorithms, the ones presented in this work allow for a numerical design of *strictly* (minimax) optimal tests. They offer a generic implementation approach with very mild assumptions on the underpinning stochastic process. Although the implementation of strictly optimal sequential tests will be shown to be prohibitively complex for most real-world applications, the presented algorithms offer a systematic approach based on standard numerical techniques that can already be used to solve small to medium size problems and is bound to become more useful with growing computational powers.

The detailed overview of the dissertation and its contributions is as follows.

In **Chapter 2**, fixed sample size as well as sequential hypothesis tests are introduced and some fundamental concepts of time-discrete stochastic processes and optimal stopping theory are reviewed.

In **Chapter 3**, the design of optimal sequential tests for stochastic processes with Markovian representations is detailed. It builds on and extends the work in [FZ15a]. While the formulation of the sequential testing problem in an optimal stopping context in Section 3.2 is rather standard, the results in Section 3.3 that connect the optimal cost function and its partial derivatives to the performance measures of the sequential test are novel and the main contribution of the chapter. Based on these results, two numerical approaches to the design of optimal sequential tests are discussed in Section 3.4. In particular, linear programming is shown to be a simple, generic and highly efficient technique suitable for the task. The main contribution in this context is a unified problem formulation that allows sequential tests with pre-specified error

probabilities to be designed by solving a single linear program. Previous approaches assumed the cost coefficients of the optimal stopping problem to be known so that additionally an outer optimization problem had to be solved to determine the optimal cost coefficients. The results are illustrated with numerical examples in Section 3.5.

In **Chapter 4**, the minimax principle and the concept of uncertainty sets is introduced. It further covers sequential tests for distributions that depend on the test statistic and reviews some aspects of convex optimization in Banach spaces.

In **Chapter 5**, the minimax robust sequential hypothesis test is derived. After stating the problem in a formal manner in Section 5.1, the characterization of least favorable distributions under general convex uncertainty sets is discussed in Section 5.2. A sufficient condition for a sequential test for two Markov processes to be minimax optimal is given in Section 5.3. It constitutes the main contribution of the chapter. In order to clarify the results and put them into context, statistical similarity measures are introduced in Section 5.4 and it is shown that the least favorable distributions are most similar with respect to a particular similarity measure that is induced by the cost function of the sequential test. In order to obtain more specific results, the density band uncertainty model is introduced and discussed in Section 5.5. It serves as an example for a typical uncertainty model in robust statistics and includes the ε -contamination model as a special case. The minimax robust test under density band uncertainties is detailed in Section 5.6. A numerical example for a minimax robust sequential test is given in Section 5.8.

1.4 Publications

The following publications have been produced during the period of doctoral candidacy.

Internationally Refereed Journal Articles

- M. Fauß and A. M. Zoubir, “A Linear Programming Approach to Sequential Hypothesis Testing”, *Sequential Analysis*, Vol 34, No 2, pp. 235–263, 2015.
- M. Fauß and A. M. Zoubir, “Two Distributions Designed to Minimize the Expected Delay in CSMA Networks”, *IEEE Signal Processing Letters*, Vol 23, No 2, pp. 267–271, 2016.

- M. Fauß and A. M. Zoubir, “Old Bands, New Tracks—Revisiting the Band Model for Robust Hypothesis Testing.”, accepted for publication in the *IEEE Transactions on Signal Processing*.
- M. Fauß and A. M. Zoubir, “On the Minimization of Convex Functionals of Probability Distributions under Band Constraints.”, under revision in the *IEEE Transactions on Signal Processing*.

Internationally Refereed Conference Papers

- M. Fauß and A. M. Zoubir, “Designing Discrete Sequential Tests via Mixed Integer Programming”, In the Proc. of the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014*, in Florence, Italy.
- M. Fauß and A. M. Zoubir, “Energy Efficient Sequential Detection Using Feedback Aided Censoring”, In the Proc. of the *IEEE International Workshop on Signal Processing Advances for Wireless Communications (SPAWC) 2013*, in Darmstadt, Germany.
- M. Fauß and A. M. Zoubir, “Performance Analysis of Sequential Detection for Collision Avoidance in Sensor Networks”, In the Proc. of the *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013*, in Vancouver, Canada.

Other Contributions

- M. Fauß and A. M. Zoubir, “Towards Minimax-Optimal Sequential Tests.”, invited talk at the *Fifth International Workshop in Sequential Methodologies 2015*, in New York City, USA.

Chapter 2

Fundamentals I

The purpose of this chapter is to introduce some basic concepts and notations that are used throughout the dissertation. While the section on statistical hypothesis tests focuses on ideas and concepts, the sections on stochastic processes and optimal stopping put additional emphasis on a mathematically rigorous formulation. The latter is essential for the analysis of sequential hypothesis tests in later chapters, which heavily build on the results presented here. In particular the recursive approach to the performance analysis of sequential tests, which is reviewed in Section 2.4.3, provides the basis for the derivation of optimal and minimax robust sequential tests in Chapter 3 and Chapter 5.

2.1 Statistical Hypothesis Tests and Decision Rules

The fundamental problem in statistical hypothesis testing is to infer from some given observations, say (x_1, \dots, x_N) , which probabilistic law the system¹ that generated the observations follows. The final goal is to be able to reliably associate all possible observations with a certain state of the system. In this work, the focus is on the *binary* hypothesis test, meaning that the number of possible states or alternatives is limited to two. This uncertainty corresponds to two *hypotheses* about the true, unknown distribution P_X , namely

$$\begin{aligned}\mathcal{H}_0: P_X &= P_0, \\ \mathcal{H}_1: P_X &= P_1,\end{aligned}\tag{2.1}$$

which are referred to as the null hypothesis and the alternative hypothesis, respectively. Designing a statistical hypothesis test for \mathcal{H}_0 and \mathcal{H}_1 is to define a function that maps every possible vector of observations to a decision for one of the two hypotheses, i.e.,

$$\delta_N: \Omega_X^N \rightarrow \{0, 1\},\tag{2.2}$$

where $\Omega_X^N = \Omega_X \times \dots \times \Omega_X$ (N times) denotes the sample space from which (x_1, \dots, x_N) is taken. The function δ is called a *decision function* or *decision rule*. It separates the

¹The term “system” is used in very wide sense here and can refer to an RF transmitter as well as to a stock market or a group of patients.

sample space Ω_X^N into two regions, one where \mathcal{H}_0 is accepted and one where \mathcal{H}_0 is rejected. For historical reasons, the latter is called the *critical region*.² It is defined by

$$\mathcal{C}_N := \{ (x_1, \dots, x_N) : \delta_N(x_1, \dots, x_N) = 1 \}.$$

Its complement is accordingly given by

$$\bar{\mathcal{C}}_N := \{ (x_1, \dots, x_N) : \delta_N(x_1, \dots, x_N) = 0 \}.$$

Whichever decision rule is used for a statistical test, it involves the risk of taking a wrong decision. More precisely, \mathcal{H}_0 can be rejected, even though it is true, or \mathcal{H}_1 can be rejected, even though it is true. The two possible errors are known as type I and type II errors. Their probabilities are given by

$$\begin{aligned} \text{Type I: } & P_0(\delta(X_1, \dots, X_N) = 1), \\ \text{Type II: } & P_1(\delta(X_1, \dots, X_N) = 0), \end{aligned}$$

where (X_1, \dots, X_N) denote the random variables that generate the observations (x_1, \dots, x_N) . Intuitively, a decision rule is “good”, if it results in small error probabilities and “bad”, if it results in large error probabilities. The approach Neyman and Pearson, [NP33], suggested to quantify this notion is to design the decision rule such that it minimizes the type II error probability and at the same time guarantees that the type I error probability does not exceed a certain bound. The corresponding optimization problem is given by

$$\min_{\delta \in \Delta} P_1(\delta(X_1, \dots, X_N) = 0) \quad \text{s.t.} \quad P_0(\delta(X_1, \dots, X_N) = 1) \leq \alpha. \quad (2.3)$$

Here Δ denotes the set of all decision rules, i.e., the set of all functions of the form (2.2). Unfortunately, from an optimization point of view, this set is nonconvex, meaning that the convex combination of two decision functions is not necessarily a valid decision function itself. This issue can be overcome by considering the class of *randomized* decision rules instead, i.e., functions of the form

$$\delta: \Omega_X^N \rightarrow [0, 1]. \quad (2.4)$$

Randomized decision rules do not directly map the observations to a decision, but to a *probability* with which a certain decision should be taken. More precisely, in a randomized test $\delta(x_1, \dots, x_N)$ denotes the probability to decide for \mathcal{H}_1 given the observations

²The approach to statistical hypothesis tests adopted in this dissertation goes back to Neyman and Pearson [NP33], who pioneered the work on tests of statistical hypotheses in the 1930s. Slightly earlier, a theory of *statistical significance tests* had been developed by Fisher [Pea00, Fis66, Box80], who proposed to base the test on a single hypothesis that is either accepted or rejected, depending on how significantly the observations support the hypothesis. The term “critical region” has its origins in the latter approach, where the formulated hypothesis is rejected, if the observation is an element of the critical region. In the Neyman and Pearson framework, however, rejecting one hypothesis implies accepting the other one so that the definition of a critical region is somewhat arbitrary.

(x_1, \dots, x_N) . Accordingly, $1 - \delta(x_1, \dots, x_N)$ denotes the conditional probability to decide for \mathcal{H}_0 . In practice, randomized decision rules are implemented by first evaluating the decision rule and subsequently performing a Bernoulli experiment with the corresponding success probability. Using this generalization, the convex sum of two decision rules is guaranteed to be a valid decision rule itself so that Δ is a convex set. Wald further showed that under mild assumptions Δ is also compact [Wal50a, Wal50b].

In practice, randomized decision rules are useful if a certain event is equally likely to occur under both hypothesis. In such cases, associating the ambiguous outcome with a deterministic decision for either hypothesis biases the test towards this hypothesis. Using randomization, however, it is possible to incorporate the ambiguity in the decision rule. Intuitively speaking, if there is no significant evidence, guessing at random is better than systematically reading a non-existing preference for one of the hypotheses into the data. Consequently, if the probability of ambiguous outcomes is negligibly small, there is no need for randomized decision rules. However, if the number of samples is small or the distributions under both hypotheses are exceedingly similar, randomization cannot be ignored; see [GZa] for an example in the context of robust detection.

The testing problem (2.3) can be formulated in terms of randomized decision rules by writing the error probabilities as

$$\begin{aligned} \text{Type I: } & E_{P_0}[\delta(X_1, \dots, X_N)], \\ \text{Type II: } & E_{P_1}[1 - \delta(X_1, \dots, X_N)], \end{aligned}$$

where E_P denotes the expected value with respect to the probability measure P . The randomized version of (2.3) reads as

$$\min_{\delta \in \Delta} E_{P_1}[1 - \delta(X_1, \dots, X_N)] \quad \text{s.t.} \quad E_{P_0}[\delta(X_1, \dots, X_N)] \leq \alpha. \quad (2.5)$$

Neyman and Pearson showed [NP33] that the optimal stopping rule δ^* that solves (2.5) is of the form

$$\delta^*(x_1, \dots, x_N) = \begin{cases} 0, & z^n(x_1, \dots, x_n) < c, \\ \kappa, & z^n(x_1, \dots, x_n) = c, \\ 1, & z^n(x_1, \dots, x_n) > c, \end{cases} \quad (2.6)$$

where

$$z^n(x_1, \dots, x_n) := \frac{p_1(x_1, \dots, x_N)}{p_0(x_1, \dots, x_N)} \quad (2.7)$$

and p_0 and p_1 denote the density functions corresponding to P_0 and P_1 , respectively. It is further assumed that $p_0 > 0$ so that the fractions in (2.6) are well defined. The free parameters $\kappa \in [0, 1]$ and $c > 0$ need to be chosen such that the constraint on the

type I error probability is fulfilled with equality. Tests with decision rules of the form (2.6) are called *probability ratio* or *likelihood ratio* tests.³

Although the likelihood ratio test is provably optimal in the sense of (2.3), Wald showed in a series of seminal publications [Wal47, WW48] that a statistical hypothesis test can be designed that achieves the same error probabilities, but requires fewer samples *on average*. The idea behind this seemingly contradictory result is to take samples *sequentially* and to stop the test as soon as the observations allow for a reliable decision. By introducing this additional degree of freedom, a sequential hypothesis test is able to adapt to the significance of the observations and adjust the required number accordingly. This also implies that for sequential hypothesis tests randomization is usually not as critical since in case of ambiguous observations, a sequential test has the option to continue collecting observations until the ambiguity is resolved. Nevertheless, for the sake of general and consistent treatment, all decision rules in this work are assumed to be randomized. Moreover, it is assumed that the samples are taken at discrete time instants. For an up-to-date overview of continuous-time sequential tests [DKW53] see [PH09] and the references therein.

The sequential hypothesis test is introduced in a more formal manner in Section 2.4. Before going into details, however, some fundamentals of stochastic processes are discussed in Section 2.2 and a short introduction to the theory of optimal stopping is given in Section 2.3.

2.2 Discrete-Time Stochastic Processes

Sequential hypothesis testing is based on the assumption that the observations are not given in advance, but become available sequentially over time. Underlying every discrete-time sequential test is, hence, a discrete-time stochastic process that generates observations according to some probabilistic law. Let (Ω, \mathcal{F}, P) be a probability space and let $(\Omega_X, \mathcal{F}_X)$ be some measurable space. A discrete-time stochastic process X is an $(\mathcal{F}, \mathcal{F}_X)$ -measurable mapping of the form

$$X: \mathbb{N} \times \Omega \rightarrow \Omega_X, \tag{2.8}$$

where \mathbb{N} denotes the natural numbers (excluding zero) and Ω_X is called the *state space* of the stochastic process. In general, the stochastic process X maps a time instant

³Throughout the dissertation, the latter term is used, which originates from a Bayesian context, but is now more widespread and used almost universally.

$n \in \mathbb{N}$ and an outcome $\omega \in \Omega$ of some underlying random experiment to a state $x \in \Omega_X$. Keeping the time instant fixed, X reduces to a regular random variable X_n with

$$X_n: \Omega \rightarrow \Omega_X.$$

It is hence possible to write a stochastic process as a sequence of random variables

$$X = (X_n)_{n \in \mathbb{N}}$$

defined on the product space $(\Omega_X^{\mathbb{N}}, \mathcal{F}_X^{\mathbb{N}})$, where

$$\Omega_X^{\mathbb{N}} := \bigtimes_{n \in \mathbb{N}} \Omega_X \quad \text{and} \quad \mathcal{F}_X^{\mathbb{N}} := \bigotimes_{n \in \mathbb{N}} \mathcal{F}_X.$$

The joint distribution P_X of $(X_n)_{n \in \mathbb{N}}$ is completely specified by the mapping (2.8) and the probability measure P via

$$P_X(\mathcal{E}) = P(\{\omega \in \Omega : (X(n, \omega))_{n \in \mathbb{N}} \in \mathcal{E}\}) \quad \forall \mathcal{E} \in \mathcal{F}_X^{\mathbb{N}}$$

so that $(\Omega_X^{\mathbb{N}}, \mathcal{F}_X^{\mathbb{N}}, P_X)$ is a well-defined probability space. See [Sae96] or [Dud02, Chapter 8] for a detailed proof. In the following, only the probability space $(\Omega_X^{\mathbb{N}}, \mathcal{F}_X^{\mathbb{N}}, P_X)$ is of interest. The existence of some random experiment corresponding to (Ω, \mathcal{F}, P) is implicitly assumed.

In order to balance generality and tractability, the analysis in this work is limited to stochastic processes that satisfy the following three assumptions.

1. X admits a time-homogeneous Markovian representation, meaning that a sequence of sufficient statistics $(\Theta_n)_{n \in \mathbb{N}_0}$ in a state space $(\Omega_\theta, \mathcal{F}_\theta)$ exists such that

$$P_X(X_{n+1} \in \mathcal{E} \mid X_1 = x_1, \dots, X_n = x_n) = P_X(X_{n+1} \in \mathcal{E} \mid \Theta_n = \theta_n) \quad (2.9)$$

for all $n \in \mathbb{N}_0$ and all $\mathcal{E} \in \mathcal{F}_X$. Here \mathbb{N}_0 denotes the natural numbers including zero. Condition (2.9) guarantees that knowledge of θ_n is sufficient to determine the distribution of X_{n+1} conditioned on all past observations. Let this conditional distribution be denoted by $P_{X_{n+1}|x_1, \dots, x_n}$, i.e.,

$$P_{X_{n+1}|x_1, \dots, x_n}(\mathcal{E}) := P_X(X_{n+1} \in \mathcal{E} \mid X_1 = x_1, \dots, X_n = x_n).$$

By (2.9), it holds that for every n and every sequence of past observations

$$P_{X_{n+1}|x_1, \dots, x_n} = P_{\theta_n} \quad (2.10)$$

so that the conditional distribution of X_{n+1} is a member of a family of distributions indexed by θ ,

$$P_{X_{n+1}|x_1, \dots, x_n} \in \{P_\theta : \theta \in \Omega_\theta\},$$

which is independent of n .

2. A function $\xi: \Omega_X \times \Omega_\theta \rightarrow \Omega_\theta$ exists that is measurable with respect to all $P \in \mathcal{P}_\theta$ and that satisfies

$$\xi(x_n, \theta_{n-1}) =: \xi_{\theta_{n-1}}(x_n) = \theta_n$$

for all $n \in \mathbb{N}$. In words, the conditional distribution of X_{n+1} is completely specified by θ_n , which can in turn be calculated recursively from θ_{n-1} and the observation x_n . The initial value θ_0 is assumed to be deterministic and given *a priori*.

3. The probability measure P_X admits a density p_X with respect to some σ -finite product measure $\mu = \mu(x_1) \otimes \mu(x_2) \otimes \dots$ so that, according to the previous assumptions, it holds that

$$p_X(x_1, \dots, x_N) = \prod_{n=1}^N p_X(x_n | \theta_{n-1}) =: \prod_{n=1}^N p_{\theta_{n-1}}(x_n)$$

for all $N \in \mathbb{N}$.

Under the hypotheses \mathcal{H}_0 and \mathcal{H}_1 , the conditional distribution P_θ and its density p_θ are denoted by P_θ^0, P_θ^1 and p_θ^0, p_θ^1 , respectively. The reasons for this particular choice of assumptions will become apparent in the course of the dissertation. In general, the assumptions have been made in order to simplify analysis and notation, while still allowing for rather general distributions and dependence structures. Random sequences satisfying Assumptions 1 to 4 cover a wide range of commonly used models, such as ARMA and ARCH models.

2.3 Discrete-Time Optimal Stopping

The theory of optimal stopping deals with the question when to stop a sequential procedure in order to maximize the expected reward or minimize the expected cost. Intuitive examples for optimal stopping problems are games of chance: When should a blackjack player take one more card? [Zir01] When should a Roulette player leave the table? [Tij12, Chapter 3] These are questions that motivated early results in optimal stopping theory. Today, examples for stopping problems can be found in a variety of fields, including medicine (“How long should a drug be taken?”), finance (“When should a stock be sold?”) and quality control (“How frequently should a machine part be replaced?”).

In the context of sequential hypothesis testing, the stopping problem is whether to continue or stop the test given the observations collected so far. A “good” stopping

procedure should balance the two goals of a sequential tests: to be reliable on the one hand and to use as few samples as possible on the other hand. In order to formulate the sequential testing problem in an optimal stopping framework, some basic results in this field need to be introduced.

Similar to the decision making problem discussed in Section 2.1, the solution of an optimal stopping problem is a *stopping rule*, which is a (possibly infinite) sequence of decision rules that map the currently available observations to a decision to either stop or continue the procedure. More formally, given the discrete-time stochastic process X , a stopping rule ψ is defined as the sequence $\psi = (\psi_n)_{n \in \mathbb{N}_0}$, where each ψ_n is a randomized decision rule defined on Ω_X^n , i.e.,

$$\psi_n : \Omega_X^n \rightarrow [0, 1], \quad n \in \mathbb{N}_0.$$

The final decision corresponding to the randomized decision rule ψ_n is in the following denoted by S_n and is a Bernoulli random variable with success probability $\psi_n(x_1, \dots, x_n)$. The decision rule ψ_0 determines whether the sequential procedure is started in the first place. Although it is a rather pathological corner case, not starting a sequential procedure can indeed be optimal if the *a priori* knowledge is already sufficient.

The time instant at which the sequential procedure is stopped is called the *stopping time* and defined as

$$\tau = \min\{n \in \mathbb{N}_0 : S_n = 1\}. \quad (2.11)$$

The stopping time τ is a random variable that takes on values in \mathbb{N}_0 . In some stopping problems *infinite* stopping times may be encountered, meaning that the procedure might not stop at all under a given stopping rule. In sequential hypothesis testing, however, it has been shown that as long as the distributions under both hypotheses are not identical a stopping rule can be found that results in a stopping time that is almost surely finite, i.e., the probability that the test eventually stops is one. More details on this aspect can be found in [GS91] and [Nov09].

Given a stopping rule ψ , the probability that a sequential procedure does continue beyond the N th time instant calculates to

$$\begin{aligned} P_X(\tau > N) &= P_X(S_1 = \dots = S_N = 0) = E_{P_X} \left[\prod_{n=1}^N (1 - S_n) \right] \\ &= E_{P_X} \left[\prod_{n=1}^N \left(1 - (1\psi_n + 0(1 - \psi_n)) \right) \right] \\ &= E_{P_X} \left[\prod_{n=1}^N (1 - \psi_n) \right] = E_{P_X} [\phi^N] \end{aligned}$$

where

$$\phi^N := \prod_{n=1}^N (1 - \psi_n) \quad (2.12)$$

is introduced for the sake of a more compact notation. The probability that the procedure is stopped at time instant N is, therefore, given by

$$\begin{aligned} P_X(\tau = N) &= P_X(\tau > N - 1) - P_X(\tau > N) \\ &= E_{P_X}[\phi^{N-1}] - E_{P_X}[\phi^N] \\ &= E_{P_X}[\phi^{N-1}\psi_N]. \end{aligned}$$

The type of stopping problem relevant to this work is as follows: Let $J_n: \Omega_X^n \rightarrow \mathbb{R}$ denote the cost of stopping a sequential procedure at the n th time instant as function of the observations (x_1, \dots, x_n) . The aim is to minimize the expected cost of the stopped procedure, i.e.,

$$\min_{\psi \in \Delta^{\mathbb{N}_0}} E_{P_X}[J_\tau(X_1, \dots, X_\tau)]. \quad (2.13)$$

The solution of (2.13) for finite and infinite time horizons is summarized in the next sections.

2.3.1 The Finite-Horizon Case

A stopping problem is said to have a finite horizon if some $N \in \mathbb{N}$ exists such that

$$\psi_N = 1 \quad P_X - \text{almost surely.}$$

In general, every stopping problem can be converted into a finite horizon problem by simply forcing the sequential procedure to stop at time instant N . The corresponding problem is then referred to as a *truncated* stopping problem. The truncated version of problem (2.13) is given by

$$\min_{\psi \in \Delta^{N+1}} E_{P_X}[J_\tau(X_1, \dots, X_\tau)] \quad \text{s.t.} \quad \psi_N = 1 \quad (2.14)$$

The solution of (2.14) is stated below.

Theorem 1 (Finite-Horizon Optimal Stopping) *For a truncated optimal stopping problem of the form (2.14), it holds that*

$$\min_{\psi \in \Delta^{N+1}} E_{P_X}[J_\tau(X_1, \dots, X_\tau)] = J_{0,N}^*, \quad (2.15)$$

where $J_{n,N} : \Omega_X^n \rightarrow \mathbb{R}$ is given by

$$(2.16) \quad J_{n,N}^*(x_1, \dots, x_n) = \begin{cases} \min\{J_n(x_1, \dots, x_n), V_{n,N}(x_1, \dots, x_n)\}, & n = 0, \dots, N-1 \\ J_N(x_1, \dots, x_N), & n = N \end{cases}$$

and $V_{n,N} : \Omega_X^n \rightarrow \mathbb{R}$ is defined via the backward recursion

$$(2.17) \quad V_{n,N}(x_1, \dots, x_n) = E_{P_X}[J_{n+1,N}^*(X_1, \dots, X_n, X_{n+1}) \mid (X_1, \dots, X_n) = (x_1, \dots, x_n)].$$

The stopping rule ψ^* solves (2.14) if and only if for all $n \in \{0, \dots, N-1\}$

$$\psi_n^* = \begin{cases} 1, & J_{n,N}(x_1, \dots, x_n) < V_{n,N}(x_1, \dots, x_n), \\ \kappa, & J_{n,N}(x_1, \dots, x_n) = V_{n,N}(x_1, \dots, x_n), \\ 0, & J_{n,N}(x_1, \dots, x_n) > V_{n,N}(x_1, \dots, x_n), \end{cases}$$

where $\kappa \in [0, 1]$ can be chosen arbitrarily.

Theorem 1 is one of the most fundamental results in optimal stopping theory. In slightly varying forms, it can be found in every standard textbook on the subject, like [CRS71, PS06] or [PH09] to name just a few.

The intuitive interpretation of Theorem 1 is that a sequential procedure should be continued as long as the cost for stopping, which is known exactly, is higher than the *expected* cost for continuing, given that the *optimal* stopping rule is applied in all subsequent time instants. This seemingly circular definition is reflected in the recursive nature of the optimal stopping rule. In (2.17), the expected cost for continuing is calculated as a function of the observations that have been collected up to time instant n , while the minimum in (2.16) ensures that this expectation is taken with respect to the optimal stopping rule. The general procedure for solving finite horizon stopping problems is hence to recursively calculate the functions $(V_n)_{0 \leq n \leq N}$ according to (2.17) and (2.16) starting with the last time instant N .

While this procedure works well in theory, the calculation of the sequence $(V_n)_{0 \leq n \leq N}$, where V_n is a function of n observations, can often be too complex to allow for an analytic or even numeric solution. As a consequence, optimal stopping theory has been applied most successfully to problems where a low-dimensional sufficient statistic exists such that J_n and V_n can be written as functions of this statistic. A more detailed discussion of this case is deferred to Chapter 3, where the optimal stopping rule of the binary sequential hypothesis test is derived. Before formally introducing the latter, the solution of infinite-horizon stopping problems is briefly revised.

2.3.2 The Infinite-Horizon Case

The solution to infinite-horizon stopping problems is defined in a rather straightforward manner, namely, as the limiting case of the finite horizon problem.

Theorem 2 (Infinite-Horizon Optimal Stopping) *Let $J_{n,N}^*$ and $V_{n,N}$ be as in (2.16) and (2.17) in Theorem 1. If the limits*

$$J_{n,\infty}^* := \lim_{N \rightarrow \infty} J_{n,N}^*$$

and

$$V_{n,\infty} := \lim_{N \rightarrow \infty} V_{n,N}$$

exist for all $n \in \mathbb{N}_0$, it holds that

$$\min_{\psi \in \Delta^{\mathbb{N}_0}} E_X[J_\tau(X_1, \dots, X_\tau)] = J_{0,\infty}^*$$

and the optimal stopping rule ψ^* is of the form

$$\psi_n^* = \begin{cases} 1, & J_{n,\infty}(x_1, \dots, x_n) < V_{n,\infty}(x_1, \dots, x_n), \\ \kappa, & J_{n,\infty}(x_1, \dots, x_n) = V_{n,\infty}(x_1, \dots, x_n), \\ 0, & J_{n,\infty}(x_1, \dots, x_n) > V_{n,\infty}(x_1, \dots, x_n), \end{cases}$$

where $\kappa \in [0, 1]$ can be chosen arbitrarily.

The question under which conditions the limits in Theorem 2 exist is intricate and beyond the scope of this work. In Chapter 3, it is shown that the sequential testing problem can be formulated as an infinite-horizon optimal stopping problem that has a well defined solution in the sense of Theorem 2.

2.4 Sequential Hypothesis Tests

In this section, the sequential hypothesis test as well as a framework for its performance analysis is introduced. Most of the presented results can be found in the literature, but often in a different form or under stricter assumptions. The purpose of this chapter is hence to give a summary of important results and to present them in a form that facilitates the derivations in the subsequent sections. Especially the section on performance analysis is central to this work. First, the sequential probability ratio test, which started the field of sequential analysis, is reviewed.

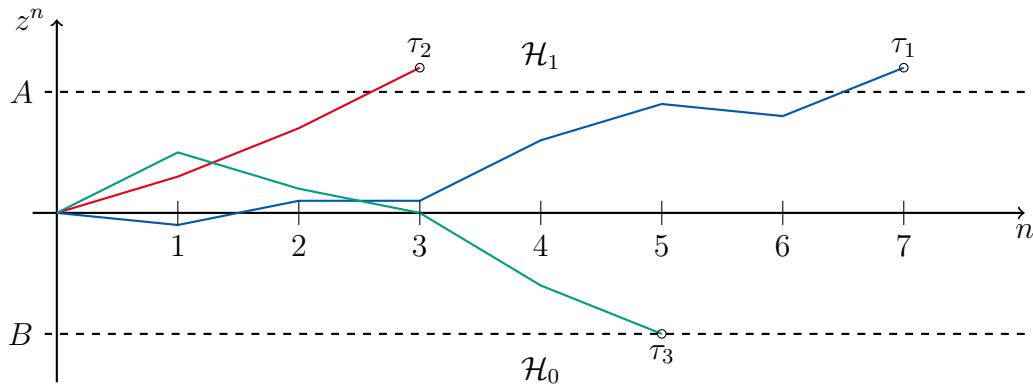


Figure 2.1: Example for realizations of a binary sequential probability ratio tests with likelihood ratio thresholds A and B .

2.4.1 The Sequential Probability Ratio Test

Sequential hypothesis testing is best introduced by means of an example, namely, the *sequential probability ratio test* as introduced by Wald [Wal47]. Down to the present day, it is most widely used in practice and serves as a benchmark for more elaborate testing procedures. Its underlying idea is to perform a likelihood ratio test of the form (2.6), but to update the likelihood ratio sequentially, instead of evaluating it for all N observations at once.

The exact procedure Wald suggested is to define two thresholds $A > B > 0$ such that the test is stopped with a decision for \mathcal{H}_1 when the likelihood ratio crosses the upper threshold A and is stopped with a decision for \mathcal{H}_0 when the likelihood ratio crosses the lower threshold B . Hence, at time instant n the following stopping and decision rule is applied:

$$\begin{aligned} z^n(x_1, \dots, x_n) \geq A &\Rightarrow \text{accept } \mathcal{H}_1, \\ z^n(x_1, \dots, x_n) \in (B, A) &\Rightarrow \text{continue testing with } x_{n+1}, \\ z^n(x_1, \dots, x_n) \leq B &\Rightarrow \text{accept } \mathcal{H}_0. \end{aligned} \quad (2.18)$$

The sequential probability ratio test is illustrated in Figure 2.1.

One of the reasons for the popularity of the sequential probability ratio test is that the thresholds A and B can be approximated by the simple expressions [Wal47]

$$A \leq \frac{1 - \beta}{\alpha} \quad \text{and} \quad B \geq \frac{\beta}{1 - \alpha}, \quad (2.19)$$

where α and β denote the desired bounds on the type I and type II error probabilities, respectively. If A and B are chosen according to (2.19) it is the case that the sum of the true error probabilities is upper bounded by the sum of the bounds, i.e.,

$$P_0(\delta_\tau = 1) + P_1(\delta_\tau = 0) \leq \alpha + \beta.$$

In practice, the true error probabilities are often significantly smaller than the bounds. This effect is due to the so-called *overshoot*, i.e., the amount by which the likelihood ratio z^τ exceeds either A or B . If the thresholds are guaranteed to be hit exactly, Wald's approximations (2.19) hold with equality. It is further shown in [WW48] that the approximations (2.19) are asymptotically optimal as $\max\{\alpha, \beta\} \rightarrow 0$. Additional correction terms to account for the overshoot have been suggested throughout the decades [Pag54, TV65, SZ13], but often complicate the design of the thresholds significantly.

It is worth noting that the approximations for A and B are *independent* of P_0 and P_1 and that both bounds on the error probabilities can be chosen freely. For fixed sample size tests, in contrast, the type II error probability, β , is completely determined by P_0 , P_1 , α and the number of samples. In this sense, the additional degree of freedom of sequential tests not only results in more efficient tests, but also allows for a more flexible test design.

Many more results on the error probabilities of the sequential probability ratio test and its expected run-length under P_0 , P_1 and general distributions P can be found in the literature. However, since the corresponding derivations contribute little to the understanding of the approach that is taken in this work, they are omitted at this point. The design and performance analysis of sequential tests is general is detailed in the next sections.

2.4.2 General Sequential Hypothesis Tests

In general, a sequential test is specified via a stopping rule $\psi = (\psi_n)_{n \in \mathbb{N}_0}$ and a sequence of decision rules $\delta = (\delta_n)_{n \in \mathbb{N}_0}$, each depending on past samples only, i.e.,

$$\psi_n = \psi_n(x_1, \dots, x_n) \quad \text{and} \quad \delta_n = \delta_n(x_1, \dots, x_n). \quad (2.20)$$

Here ψ_n denotes the probability to stop the test and δ_n denotes the probability to decide for the alternative hypothesis \mathcal{H}_1 , given that the test has stopped. Both probabilities are conditioned on the observations (x_1, \dots, x_n) . The decision tree of a sequential

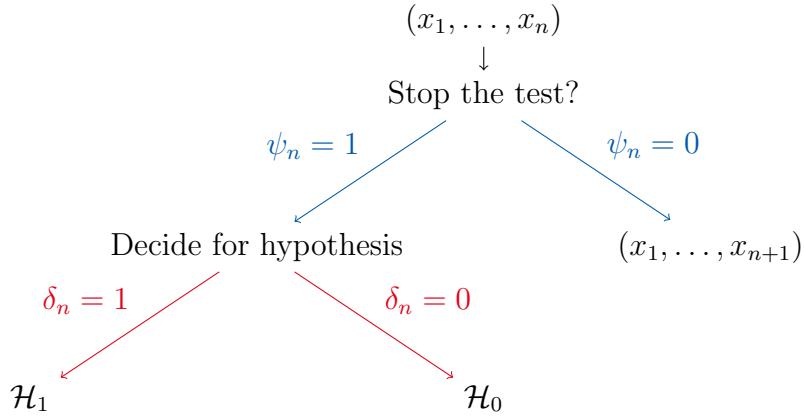


Figure 2.2: Decision tree of a sequential test between two hypotheses.

test is depicted in Figure 2.2. At each time instant, the stopping rule ψ_n is evaluated at (x_1, \dots, x_n) . If the stopping decision is negative, the test continues and the next observation is acquired. If the stopping decision is positive, the decision rule δ_n is evaluated at (x_1, \dots, x_n) and the corresponding decision for either \mathcal{H}_0 or \mathcal{H}_1 is taken. Note that this implies that the stopping rule is a function of the decision rule since the cost for stopping depends on which hypothesis is accepted. The stopping rule of the threshold test (2.18), for example, is given by

$$\psi_n(x_1, \dots, x_n) = \mathbf{1}_{(B,A)}(z^n),$$

where $\mathbf{1}_{\mathcal{A}}$ denotes the indicator function of the set \mathcal{A} . The corresponding decision rule is given by

$$\delta_n(x_1, \dots, x_n) = \mathbf{1}_{[A,\infty)}(z^n).$$

For the sake of a more compact notation, a pair of stopping and decision rules $(\psi, \delta) \in \Delta^{\mathbb{N}_0} \times \Delta^{\mathbb{N}_0}$ is, in the following, referred to as a *policy* and denoted by

$$\pi := (\psi, \delta). \quad (2.21)$$

The set of all feasible policies is denoted by

$$\Pi := \Delta^{\mathbb{N}_0} \times \Delta^{\mathbb{N}_0}. \quad (2.22)$$

Apart from the condensed notation, the notion of policies has no conceptual advantages or disadvantages and in general π can be understood as a mere placeholder for the pair (ψ, δ) .

While policies that map the collected observations to stopping and decision probabilities are the most universal way of specifying sequential tests, a typical testing procedure involves an intermediate step, namely the calculation of a *test statistic*. The test

statistic of the sequential probability ratio test discussed in the previous section is the likelihood ratio z^n , which will later be shown to be indeed the optimal test statistic. However, many different statistics can and have been used in sequential tests, including quantized [FZ14] and multi-dimensional [Lor76] test statistics. In particular the latter are of importance in the upcoming chapters.

In general, a test statistic $T^n: \Omega_X^n \rightarrow \Omega_T$ maps the observations (x_1, \dots, x_n) to a point in a measurable space $(\Omega_T, \mathcal{F}_T)$. An important advantage of using test statistics is that the space $(\Omega_T, \mathcal{F}_T)$ is independent of the time instant n so that the sequential test can be performed based on a single, preferably low-dimensional quantity instead of an ever-growing sequence of observations. For this to hold, the decision rules ψ_n and δ_n need to be defined as functions of the test statistic, i.e.,

$$\psi_n, \delta_n: \Omega_T \rightarrow [0, 1] \quad \forall n \in \mathbb{N}_0. \quad (2.23)$$

At every time instant n , a randomized stopping rule partitions the space Ω_T into a region where the test is continued, a region where the test is continued with a certain probability (excluding zero and one) and a region where the test is stopped. The latter is usually referred to as the *stopping region* and is denoted by \mathcal{S} in this dissertation. It is defined as

$$\begin{aligned} \mathcal{S}_n &= \{ t \in \Omega_T : \psi_n(t) = 1 \}, \\ \partial\mathcal{S}_n &= \{ t \in \Omega_T : \psi_n(t) \in (0, 1) \}, \\ \overline{\mathcal{S}}_n &= \{ t \in \Omega_T : \psi_n(t) = 0 \}. \end{aligned}$$

Analogously, the critical region of a sequential test at time instant n is given by

$$\begin{aligned} \mathcal{C}_n &= \{ t \in \Omega_T : \delta_n(t) = 1 \}, \\ \partial\mathcal{C}_n &= \{ t \in \Omega_T : \delta_n(t) \in (0, 1) \}, \\ \overline{\mathcal{C}}_n &= \{ t \in \Omega_T : \delta_n(t) = 0 \}. \end{aligned}$$

The initial test statistic T^0 is assumed to be deterministic and known, i.e., $T^0 = t_0$. There typically exists a natural choice for t_0 , such as $t_0 = 1$ for the sequential probability ratio test, or $t_0 = 0$ in cases where the sign of the test statistic indicates the preferred hypothesis [FZ14].

The formulation of sequential tests in terms of test statistics is particularly helpful for their performance analysis, which is the subject of the next section.

2.4.3 Performance Analysis

Before dealing with the question how to *design* sequential hypothesis tests, it is instructive to investigate how to *analyze* a sequential test whose test statistic and policy are given. The reduction in the average number of samples that is achieved by sequential tests is well known and documented and relevant examples can be found in every standard textbook [Sie85, GS91, TLY03]. As a rule of thumb, a well designed sequential test reduces the average number of samples by roughly 50%, compared to a fixed sample size test with the same error probabilities. How to obtain exact expressions for the error probabilities and expected run-length of general sequential tests is detailed in this section.

Let T^n , with $T^n = T^n(x_1, \dots, x_n)$ and initial value $T^0 = t_0$, denote a given test statistic. Further, assume that the sequence $(T^n)_{n \in \mathbb{N}}$ fulfills

$$P_X(T^{n+1} \in \mathcal{E} \mid (T^0, \dots, T^n) = (t_0, \dots, t_n), \tau \geq n) = P_X(T^{n+1} \in \mathcal{E} \mid T^n = t_n) \quad (2.24)$$

for all $\mathcal{E} \in \mathcal{F}_T$, $n \in \mathbb{N}_0$ and $(t_0, \dots, t_n) \in \Omega_T^{n+1}$. Property (2.24) has two implications on a sequential test. First, at every time instant n , the current test statistic T^n is a sufficient representation of the sequence of past test statistics $(T^m)_{m \leq n}$. Second, the update of the test statistic is independent of whether the test has already stopped or not. In order to understand the last implication, it is helpful to think of a sequential test being performed by first generating the sequence of test statistics $(T^n)_{n \leq N}$ for some sufficiently large N and then determining the stopping time according to the definition in (2.11). If (2.24) holds, the sequence of test statistics can be generated by only keeping track of its current value and without knowing whether or not the test has actually stopped already. Assuming that the sequence of test statistics satisfies property (2.24) simplifies the analysis of sequential tests significantly. In Section 2.3, it is shown that for *optimal* test statistics this assumption always holds.

For the analysis of the expected run-length and the two error probabilities, three sequences of functions are introduced, namely $(\alpha_\pi^n)_{n \in \mathbb{N}_0}$, $(\beta_\pi^n)_{n \in \mathbb{N}_0}$ and $(\gamma_\pi^n)_{n \in \mathbb{N}_0}$, with $\alpha_\pi^n, \beta_\pi^n: \Omega_T \rightarrow [0, 1]$ and $\gamma_\pi^n: \Omega_T \rightarrow \mathbb{R}_+$ defined as

$$\alpha_\pi^n(t) := E_{P_0}[\delta_\tau \mid T^n = t, \tau \geq n], \quad (2.25)$$

$$\beta_\pi^n(t) := E_{P_1}[1 - \delta_\tau \mid T^n = t, \tau \geq n], \quad (2.26)$$

$$\gamma_\pi^n(t) := E_P[\tau - n \mid T^n = t, \tau \geq n]. \quad (2.27)$$

Since the expected run-length can be defined under arbitrary distributions, the measure P is left unspecified. This notation is used in later chapters as well to indicate that a

distribution is not determined by the hypotheses, but can be chosen by the test designer. The interpretation of α_π^n , β_π^n and γ_π^n is as follows: $\alpha_\pi^n(t)$ denotes the probability that a test using policy π finishes with an erroneous decision for \mathcal{H}_1 given that it is in state $T^n = t$ at time instant n . Analogously, $\beta_\pi^n(t)$ denotes the probability that a test using policy π finishes with an erroneous decision for \mathcal{H}_0 given that it is in state $T^n = t$ at time instant n . The third function, $\gamma_\pi^n(t)$, denotes the expected *remaining* run-length of a test using policy π given that it is in state $T^n = t$ at time instant n .

The connection between the three sequences of functions and the performance measures of the corresponding sequential test is given by conditioning on the certain events $T^0 = t_0$ and $\tau \geq 0$ so that

$$\alpha_\pi^0(t_0) = E_{P_0}[\delta_\tau \mid T^0 = t_0, \tau \geq 0] = E_{P_0}[\delta_\tau], \quad (2.28)$$

$$\beta_\pi^0(t_0) = E_{P_1}[1 - d_\tau \mid T^0 = t_0, \tau \geq 0] = E_{P_1}[1 - \delta_\tau], \quad (2.29)$$

$$\gamma_\pi^0(t_0) = E_P[\tau \mid T^0 = t_0, \tau \geq 0] = E_P[\tau(\psi)]. \quad (2.30)$$

The advantage of expressing the error probabilities and expected run-length of sequential tests in terms of sequences of functions of the test statistic is that this approach allows for a recursive calculation of these quantities. More precisely, using property (2.24), it is shown in Appendix A.1 that α_π^n , β_π^n and γ_π^n satisfy the backward recursions

$$\alpha_\pi^n(t) = \psi_n(t)\delta_n(t) + (1 - \psi_n(t))E_{P_0}[\alpha_\pi^{n+1}(T^{n+1}) \mid T^n = t], \quad (2.31)$$

$$\beta_\pi^n(t) = \psi_n(t)(1 - \delta_n(t)) + (1 - \psi_n(t))E_{P_1}[\beta_\pi^{n+1}(T^{n+1}) \mid T^n = t], \quad (2.32)$$

$$\gamma_\pi^n(t) = (1 - \psi_n(t))(1 + E_P[\gamma_\pi(T^{n+1}) \mid T^n = t]). \quad (2.33)$$

The validity of these recursions is most obvious if a deterministic policy is used, meaning that the stopping and decision rules are non-randomized. In this case, (2.31)–(2.33) simplify to

$$\alpha_\pi^n(t) = \begin{cases} 1, & t \in \mathcal{S} \cap \mathcal{C}, \\ 0, & t \in \mathcal{S} \cap \bar{\mathcal{C}}, \\ E_{P_0}[\alpha_\pi^{n+1}(T^{n+1}) \mid T^n = t], & t \in \bar{\mathcal{S}}, \end{cases} \quad (2.34)$$

$$\beta_\pi^n(t) = \begin{cases} 0, & t \in \mathcal{S} \cap \mathcal{C}, \\ 1, & t \in \mathcal{S} \cap \bar{\mathcal{C}}, \\ E_{P_1}[\beta_\pi^{n+1}(T^{n+1}) \mid T^n = t], & t \in \bar{\mathcal{S}}, \end{cases} \quad (2.35)$$

$$\gamma_\pi^n(t) = \begin{cases} 0, & t \in \mathcal{S}, \\ 1 + E_P[\gamma_\pi^{n+1}(T^{n+1}) \mid T^n = t], & t \in \bar{\mathcal{S}}. \end{cases} \quad (2.36)$$

The intuitive interpretation of (2.34) and (2.35) is that the error probabilities at time instant n , given the test statistic T^n , are either zero or one if the test stops, or are given by the expected value of the error probabilities at time instant $n + 1$, which are determined by the updated test statistic T^{n+1} . Analogously, in (2.36) the expected remaining run-length is either zero if the test stops at the current time instant, or one larger than the expected remaining run-length at the next time instant. The general case in (2.31)–(2.33) can be interpreted analogously, when taking the randomization into account.

Without making further assumptions, it is not possible to calculate the error probabilities and expected run-length of a test via the α, β and γ -functions since the infinite backward recursions (2.31)–(2.33) cannot be resolved. However, for two important special cases an explicit evaluation is possible.

The first special case is that of *truncated* sequential tests, i.e., tests for which some $N \in \mathbb{N}$ exists such that $\psi_N = 1$. For truncated sequential tests it holds that

$$\begin{aligned}\alpha_\pi^N(t) &= \delta_N(t), \\ \beta_\pi^N(t) &= 1 - \delta_N(t), \\ \gamma_\pi^N(t) &= 0,\end{aligned}\tag{2.37}$$

so that the error probabilities and the expected run-length can be calculated by evaluating the backward recursions (2.31)–(2.33) N times, starting with $\alpha_\pi^N, \beta_\pi^N$ and γ_π^N .

The second type of tests for which (2.31)–(2.33) can be evaluated explicitly are *time-invariant* tests, i.e., tests whose stopping and decision rules do not depend on the current time instant, but only on the value of the test statistic. The sequential probability ratio test with constant thresholds is a typical example for a time-invariant test. A more formal definition is stated below.

Definition 1 *A sequential test is referred to as time-invariant if the policy π and the sequence of test statistics $(T^n)_{n \in \mathbb{N}_0}$ satisfy the following conditions:*

1. *For all $m, n \in \mathbb{N}_0$, it is the case that*

$$\psi_n = \psi_m \quad \text{and} \quad \delta_n = \delta_m.\tag{2.38}$$

2. *For all $t \in \Omega_t$ and all $\mathcal{E} \in \mathcal{F}_T$, it holds that*

$$P_X(T^{n+1} \in \mathcal{E} \mid T^n = t) = P_X(T^{m+1} \in \mathcal{E} \mid T^m = t).\tag{2.39}$$

If a sequential test is time-invariant in the sense of Definition 1, it holds that for all $m, n \in \mathbb{N}_0$

$$\alpha_\pi^n = \alpha_\pi^m =: \alpha_\pi, \quad \beta_\pi^n = \beta_\pi^m =: \beta_\pi \quad \text{and} \quad \gamma_\pi^n = \gamma_\pi^m =: \gamma_\pi, \quad (2.40)$$

where $\alpha_\pi, \beta_\pi: \Omega_T \rightarrow [0, 1]$ and $\gamma_\pi: \Omega_T \rightarrow \mathbb{R}_+$ solve the following second-type Fredholm integral equations:

$$\alpha_\pi(t) = \psi(t)\delta(t) + (1 - \psi(t))E_{P_0}[\alpha_\pi(T^1) | T^0 = t], \quad (2.41)$$

$$\beta_\pi(t) = \psi(t)(1 - \delta(t)) + (1 - \psi(t))E_{P_1}[\beta_\pi(T^1) | T^0 = t], \quad (2.42)$$

$$\gamma_\pi(t) = (1 - \psi(t))(1 + E_P[\gamma_\pi(T^1) | T^0 = t]). \quad (2.43)$$

The pair of test statistics T^{n+1}, T^n , here set to T^1 and T^0 , can be chosen arbitrarily due to property (2.39). For the same reason, the time index of the stopping and decision rules has been omitted in the notation.⁴

The integral equations (2.41)–(2.43) are a fundamental result of sequential analysis and have been derived in many works. For the sequential probability ratio test, (2.41)–(2.43) are proven in detail in [Kem58, CM65, Fel66, TNB14] to name just a few references. More general results based on random walks in multidimensional spaces can be found in [CM65, GS06]. An important result on the existence and uniqueness of the solutions of (2.41)–(2.43) is obtained as a special case of [Oga06, Theorem 3.1]. In essence, the theorem states that under mild boundedness conditions the integral equations (2.41)–(2.43) have a *unique* solution, if the homogeneous integral equations

$$\alpha_\pi(t) = E_{P_0}[\alpha_\pi(T^1) | T^0 = t], \quad (2.44)$$

$$\beta_\pi(t) = E_{P_1}[\beta_\pi(T^1) | T^0 = t], \quad (2.45)$$

$$\gamma_\pi(t) = E_P[\gamma_\pi(T^1) | T^0 = t], \quad (2.46)$$

only have trivial (constant) solutions. In the context of sequential hypothesis tests, this result has a very intuitive interpretation. By inspection, the homogeneous integral equations (2.44) and (2.45) can be obtained from (2.41) and (2.43) by choosing $\psi(t) = 0$, meaning that the test never stops. As a consequence, the definition of state-dependent error probabilities becomes meaningless since all states are equivalent. Therefore, every constant function still technically solves (2.44) and (2.45). Meaningful solutions, however, with truly state-dependent error probabilities, can no longer be found. Similar contradictions arise for the expected run-length. The homogeneous

⁴This notation is slightly inaccurate since ψ and δ are defined as *sequences* of functions. However, for time invariant tests, these functions are identical so that ψ and δ can be associated unambiguously with a single function.

equation (2.46) can be obtained from (2.43) by setting $\psi(t) = 0$ and additionally assigning a cost of zero to each additional sample taken by the sequential test. In (2.43), this cost is set to one so that it coincides with the number of samples. Again, such a scenario allows one to assign the same arbitrary expected cost to all states, but makes tests with finite expected run-length, in the sense of (2.27), impossible.

In a nutshell, [Oga06, Theorem 3.1] guarantees that the error probabilities and run-length of every non-trivial time-invariant sequential test are unique and can be calculated by solving (2.41)–(2.43). Since the discussion of trivial sequential tests⁵ is of little theoretical or practical interest, it is in the following assumed that every time-invariant test is non-trivial.

In the larger context of the dissertation, the results in this section serve two purposes. First, they provide a means to numerically analyze and compare a broad class of sequential tests. Second, they are used in the upcoming chapters to relate integral equations that arise in the design of optimal sequential tests to the performance measures of these tests. The latter method is a core contribution of this work and is used for the derivation of both optimal and minimax robust sequential tests.

⁵An example for a trivial time-invariant sequential test is a test whose test statistic is constant so that it either stops at $n = 0$ or keeps running indefinitely.

Chapter 3

Optimal Sequential Hypothesis Tests

The problem of sequentially testing for two simple hypotheses, under different restrictions and assumptions, has been treated extensively in the literature; see, for example, [TNB14] and references therein. The approach used in this thesis differs from most works in that it aims for a *strictly optimal*, yet *implementable* solution. In the existing literature, the two objectives are typically handled separately, in the sense that the derivation of optimal solutions is treated as a purely theoretical exercise, whereas the actual design of sequential tests is based on asymptotic results or approximations [BM55, GL11, LZL12]. In fact, asymptotic results are prevailing in sequential analysis to an extent that *optimal* and *asymptotically optimal* are used interchangeably in the titles of many works [Nik94, CVMM02, NWJ08, LZL12].

In addition, the presented results shed new light on the relation between the sequential testing problem in its classic, constrained formulation and its formulation as an optimal stopping problem. While both problem formulations have been studied independently, the connection between them has not received comparable attention. In particular, the question of how the cost function of the optimal stopping problem relates to the three individual performance measures of the corresponding sequential test has not been covered in detail in the literature before.

The chapter is mostly based on [FZ15a], where the linear programming approach to the design of sequential hypothesis tests has first been suggested. This approach is discussed in detail in Section 3.4.2. In comparison to [FZ15a], a slightly different line of arguments is used in this chapter and the results are extended to randomized stopping and decision rules.

3.1 Problem Formulation

In analogy to the fixed sample size test in Section 2.1, the two hypotheses of a binary sequential hypothesis test are given by

$$\begin{aligned}\mathcal{H}_0: P_X &= P_0, \\ \mathcal{H}_1: P_X &= P_1,\end{aligned}\tag{3.1}$$

where both P_0 and P_1 are assumed to fulfill the assumptions in Section 2.2.

The sequential testing problem investigated in this section is known as the *modified Kiefer–Weiss problem*. It was suggested in [KW57] and consists in designing a sequential test that guarantees certain error probabilities under P_0 and P_1 while minimizing the expected run-length of the test under some third measure $P_X = P$. The measure P needs to fulfill the assumptions in Section 2.2, but can otherwise be chosen arbitrarily. It is further assumed that P_0 , P_1 and P share the same support so that their mutual likelihood ratios are well defined.

In this formulation, the problem of minimizing the expected run-length under one of the hypotheses, i.e., $P = P_0$ or $P = P_1$, is included as a special case. For the design of *robust* sequential tests in Chapter 5, however, the case where $P \neq P_0, P_1$, i.e., a *mismatch* between the true and the assumed distributions occurs, is particularly important.

Let the stopping rule ψ , the decision rule δ and the stopping time τ be as defined in Section 2.3 and Section 2.4. The modified Kiefer–Weiss problem, which can be seen as the sequential equivalent of the Neyman–Pearson problem (2.5), reads as

$$\min_{(\psi, \delta) \in \Pi} E_P[\tau(\psi)] \quad \text{s.t.} \quad E_{P_0}[\delta_\tau] \leq \alpha, \quad E_{P_1}[1 - \delta_\tau] \leq \beta, \quad (3.2)$$

where $\alpha, \beta \in [0, 1]$ denote the bounds on the error probabilities, or target error probabilities, and Π is the set of all feasible policies defined in (2.22). Throughout this work, α and β without subscript refer to given target error probabilities, α_π and β_π , as introduced in Section 2.4.3, refer to the error probabilities of a test using policy π . The set of policies that solve (3.2) is denoted by

$$\Pi_{\alpha, \beta}^* \subset \Pi. \quad (3.3)$$

The sets of stopping and decision rules that can be used to compose an optimal policy are denoted by $\Sigma_{\alpha, \beta}^*$ and $\Delta_{\alpha, \beta}^*$, i.e.,

$$\begin{aligned} \Sigma_{\alpha, \beta}^* &:= \{ \psi \in \Delta^{\mathbb{N}_0} : \exists \delta \in \Delta^{\mathbb{N}_0} : (\delta, \psi) \in \Pi_{\alpha, \beta}^* \}, \\ \Delta_{\alpha, \beta}^* &:= \{ \delta \in \Delta^{\mathbb{N}_0} : \exists \psi \in \Delta^{\mathbb{N}_0} : (\delta, \psi) \in \Pi_{\alpha, \beta}^* \}. \end{aligned}$$

A commonly used procedure to solve (3.2), in an optimal stopping framework, is to reformulate the constrained problem as an unconstrained minimization problem. This is done by replacing the explicit constraints on the error probabilities with two weighted penalty terms in the cost function so that (3.2) becomes

$$\min_{(\psi, \delta) \in \Pi} E_P[\tau(\delta)] + \lambda_0 E_{P_0}[\delta_\tau] + \lambda_1 E_{P_1}[1 - \delta_\tau], \quad (3.4)$$

where $\lambda = (\lambda_0, \lambda_1) \in \mathbb{R}_+$ are two positive weights. In analogy to (3.3), let $\Pi_\lambda^* \subset \Pi$ denote the set of all policies that solve (3.4) and Σ_λ^* and Δ_λ^* the corresponding sets of optimal stopping and decision rules. The next theorem states a well known relation between Problem (3.2) and Problem (3.4).

Theorem 3 *If a policy $\pi \in \Pi$ solves (3.4) and it holds that*

$$E_{P_0}[\delta_\tau] = \alpha \quad \text{and} \quad E_{P_1}[1 - \delta_\tau] = \beta,$$

then π also solves (3.2), i.e., $\pi \in \Pi_\lambda^ \cap \Pi_{\alpha, \beta}^*$.*

A proof is detailed in Appendix A.2. Theorem 3 gives a sufficient condition for a sequential test to be optimal in the sense of (3.2). In the course of the thesis, it will become clear that this condition can always be fulfilled by some, possibly randomized, testing policy. Theorem 3 has in similar form been given in other works on sequential testing, compare, for example, [Nov09, Theorem 2.1]. However, the question how to choose the weights λ_0, λ_1 in order to meet the target error probabilities has received little to no attention in the literature. A systematic approach to this question is one of the main contributions of this chapter. The relation between λ and the error probabilities of tests with policies $\pi \in \Pi_\lambda^*$ is discussed in detail in Section 3.3. For now, the focus is on Problem (3.4) with λ given and fixed.

3.2 Solution of the Optimal Stopping Problem

In order to formulate (3.4) as an optimal stopping problem of the form (2.13), first the expected cost for stopping the sequential test at an arbitrary, but fixed, time instant $\tau = n$ needs to be derived. It is given by

$$\begin{aligned} E_P[n] + \lambda_0 E_{P_0}[\delta_n] + \lambda_1 E_{P_1}[1 - \delta_n] &= n + \int \delta_n \lambda_0 p_0 d\mu^n + \int (1 - \delta_n) \lambda_1 p_1 d\mu^n \\ &= \int \left(n + \delta_n \lambda_0 \frac{p_0}{p} + (1 - \delta_n) \lambda_1 \frac{p_1}{p} \right) p d\mu^n \quad (3.5) \\ &= E_P[n + \delta_n \lambda_0 z_0^n + (1 - \delta_n) \lambda_1 z_1^n], \end{aligned}$$

where

$$z_i^n = z_i^n(x_1, \dots, x_n) := \frac{dP_i}{dP}(x_1, \dots, x_n) = \frac{p_i(x_1, \dots, x_n)}{p(x_1, \dots, x_n)}, \quad i = 0, 1 \quad (3.6)$$

denotes the Radon-Nikodym derivative (likelihood ratio) of P_i with respect to P , evaluated at (x_1, \dots, x_n) . Moreover, since for all (λ_0, λ_1) and (z_0^n, z_1^n) it holds that

$$\delta_n \lambda_0 z_0^n + (1 - \delta_n) \lambda_1 z_1^n \geq \min\{\lambda_0 z_0^n, \lambda_1 z_1^n\},$$

the optimal decision rule δ_n^* needs to be chosen such that

$$\delta_n^* = \begin{cases} 0, & \lambda_0 z_0^n > \lambda_1 z_1^n, \\ 1, & \lambda_0 z_0^n < \lambda_1 z_1^n. \end{cases} \quad (3.7)$$

For $\lambda_0 z_0^n = \lambda_1 z_1^n$, δ_n can be chosen arbitrarily since both decisions are equally costly. This result is fixed in the next theorem.

Theorem 4 *The set of decision rules that solve Problem (3.4) is given by*

$$\Delta_\lambda^* = \{ \delta \in \Delta^{\mathbb{N}_0} : \mathbf{1}_{\mathcal{C}_\lambda}(z) \leq \delta_n(z) \leq \mathbf{1}_{\mathcal{C}_\lambda \cup \partial \mathcal{C}_\lambda}(z) \quad \forall n \in \mathbb{N}_0 \}, \quad (3.8)$$

where $z = (z_0, z_1)$ and the critical region \mathcal{C}_λ is given by

$$\begin{aligned} \mathcal{C}_\lambda &= \{ z \in \mathbb{R}_+^2 : \lambda_0 z_0 < \lambda_1 z_1 \}, \\ \partial \mathcal{C}_\lambda &= \{ z \in \mathbb{R}_+^2 : \lambda_0 z_0 = \lambda_1 z_1 \}, \\ \bar{\mathcal{C}}_\lambda &= \{ z \in \mathbb{R}_+^2 : \lambda_0 z_0 > \lambda_1 z_1 \}. \end{aligned}$$

Theorem 4 implies that once the sequential test has stopped, a regular likelihood ratio test with threshold λ_0/λ_1 is performed to decide for a hypothesis. This is in agreement with the well established optimality of the likelihood ratio test for fixed sample sizes [NP33, LR05]. Moreover, it follows from Theorem 4 that the deterministic decision rules $\delta_n = \mathbf{1}_{\mathcal{C}_\lambda}$ and $\delta_n = \mathbf{1}_{\mathcal{C}_\lambda \cup \partial \mathcal{C}_\lambda}$ both solve Problem (3.4). The question whether optimal deterministic decision rules also exist for the constrained Problem (3.2) is addressed in the next section.

Given the set of optimal decision rules Δ_λ^* , Problem (3.4) reduces to an optimal stopping problem of the form (2.13) with instantaneous cost functions

$$J_n(x_1, \dots, x_n) = J_n(z^n) = n + g_\lambda(z^n), \quad n \in \mathbb{N}_0, \quad (3.9)$$

where $z = (z_0, z_1)$ and the function

$$g_\lambda(z) := \min\{\lambda_0 z_0, \lambda_1 z_1\} \quad (3.10)$$

has been introduced for the sake of a more compact notation. Problem (3.4) can now be formulated as the infinite-horizon stopping problem

$$\min_{\psi \in \Delta^{\mathbb{N}_0}} E_P[\tau + g_\lambda(z^\tau)]. \quad (3.11)$$

The solution of (3.11) is stated in the following theorem.

Theorem 5 Let $\lambda > 0$ be given and let P, P_0 and P_1 be such that they fulfill the assumptions in Section 2.2. The functional equation

$$\rho_\lambda(z, \theta) = \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta(x) \right\}, \quad (3.12)$$

with g_λ defined in (3.10), has a unique solution $\rho_\lambda \geq 0$ on

$$(\Omega_\rho, \mathcal{F}_\rho) := (\mathbb{R}_+^2 \times \Omega_\theta, \mathcal{B}_+^2 \otimes \mathcal{F}_\theta), \quad (3.13)$$

where \mathcal{B}_+^2 denotes the Borel σ -algebra on \mathbb{R}_+^2 . It further holds that

$$\min_{\psi \in \Delta^{\mathbb{N}_0}} E_P [\tau + g_\lambda(z^\tau)] = \rho_\lambda(1, 1, \theta_0). \quad (3.14)$$

A proof of Theorem 5 is given in Appendix A.5.

One of the consequences of Theorem 5 is that the tuple (z^n, θ_n) is a sufficient test statistic for an optimal sequential test, if the underlying random process satisfies the assumptions in Section 2.2 under P_0, P_1 and P . The two components of the test statistic correspond to the two types of information that are necessary to follow the optimal stopping rule. The likelihood ratios z are needed to evaluate the cost for stopping, while the state of the Markov process θ is needed to evaluate the conditional expectation that determines the cost for continuing the test. This means that in general more information is required to perform an optimal *sequential* test than is required to perform an optimal *fixed sample size* test. For the latter, the likelihood ratio is always a sufficient test statistic, irrespective of the process that generates the observations. The need for additional knowledge in the sequential case arises, on the one hand, from the objective to minimize the expected run-length under a distribution that is different from P_0 and P_1 and, on the other hand, from the need to make predictions about the behavior of the underlying stochastic process. Both are properties of sequential tests that do not arise in the fixed sample size case. However, if $P = P_0$ or $P = P_1$ it holds that either $z_0^n = 1$ or $z_1^n = 1$ for all n so that the z -component of the test statistic becomes one-dimensional. This is illustrated with an example in Section 3.5.

The optimal test statistic and testing policies that solve Problem (3.11) follow directly from Theorem 2 and Theorem 5. For the sake of completeness and later reference, they are fixed in the following corollaries.

Corollary 1 The optimal test statistic of a test solving (3.4) is given by

$$T^n(x_1, \dots, x_n) = (z_0^n, z_1^n, \theta_n), \quad (3.15)$$

where z_0 and z_1 are the likelihood ratios defined in (3.6) and θ_n is a sufficient statistic for (x_1, \dots, x_n) in the sense of (2.9).

Corollary 2 *The set of testing policies Π_λ^* that solve Problem (3.4) is given by*

$$\Pi_\lambda^* = \{ (\delta, \psi) \in \Pi : \delta \in \Delta_\lambda^*, \psi \in \Sigma_\lambda^* \},$$

where Δ_λ^* is defined in Theorem 4 and

$$\Sigma_{\lambda, \psi}^* = \{ \psi \in \Delta^{\mathbb{N}_0} : \mathbf{1}_{\mathcal{S}_\lambda}(z, \theta) \leq \psi_n(z, \theta) \leq \mathbf{1}_{\mathcal{S}_\lambda \cup \partial \mathcal{S}_\lambda}(z, \theta) \quad \forall n \in \mathbb{N}_0 \} \quad (3.16)$$

where the stopping region \mathcal{S}_λ is given by

$$\begin{aligned} \mathcal{S}_\lambda &= \{ (z, \theta) \in \Omega_\rho : g_\lambda(z) < d_\lambda(z, \theta) \}, \\ \partial \mathcal{S}_\lambda &= \{ (z, \theta) \in \Omega_\rho : g_\lambda(z) = d_\lambda(z, \theta) \}, \\ \bar{\mathcal{S}}_\lambda &= \{ (z, \theta) \in \Omega_\rho : g_\lambda(z) > d_\lambda(z, \theta) \}, \end{aligned}$$

g_λ is defined in (3.10) and

$$d_\lambda(z, \theta) := 1 + \int \rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta(x), \quad (3.17)$$

with ρ_λ defined in Theorem 5.

Corollary 3 *A sequential test that uses a test statistic chosen according to Corollary 1 and a policy $\pi \in \Pi_\lambda^*$ is time invariant, i.e., π satisfies (2.38) and the sequence of test statistics $T^n = (z^n, \theta_n)$ satisfies (2.24) and (2.39).*

Corollary 1 and Corollary 2 follow immediately from Theorem 2 and Theorem 5. Corollary 3 is shown in Appendix A.4.

The results in this section, in particular the implicit definition of the cost function ρ_λ in (3.12), provide the basis for the analysis and the design of optimal and minimax robust sequential tests. Before proceeding with a more detailed investigation of the properties of ρ_λ and its connection to the constrained problem (3.2), it is helpful to note that the integral equation (3.12) can alternatively be written as

$$\rho_\lambda(z, \theta) = \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda dH_{z, \theta} \right\}, \quad (3.18)$$

where $\{ H_{z, \theta} : (z, \theta) \in \Omega_\rho \}$ is a family of probability measures on $(\Omega_\rho, \mathcal{F}_\rho)$ satisfying

$$H_{z, \theta}(\mathcal{E}_z \times \mathcal{E}_\theta) = P_\theta \left(\left\{ x \in \Omega_X : \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)} \right) \in \mathcal{E}_z, \xi_\theta(x) \in \mathcal{E}_\theta \right\} \right) \quad (3.19)$$

for all $\mathcal{E}_z \in \mathcal{B}_+^2$ and $\mathcal{E}_\theta \in \mathcal{F}_\theta$. The formulations (3.12) or (3.18) are both used in the upcoming sections. The notations $H_{z, \theta}^0$ and $H_{z, \theta}^1$ are used to refer to the families of distributions where P_θ in (3.19) is replaced by P_θ^0 or P_θ^1 , respectively.

3.3 Properties of the Cost Function ρ_λ

The solution of the infinite horizon optimal stopping problem (3.4) was derived in the previous section. The corresponding optimal testing policy was shown to be determined by the function ρ_λ , which is defined as the solution of a nonlinear integral equation. This result is remarkable in the sense that an optimization over stopping and decision rules is transformed into a problem of pure functional analysis that can be solved without explicitly introducing the functions ψ and δ in the first place. However, being able to solve the optimal stopping problem (3.4) does not mean that one is able to solve the constrained sequential testing problem (3.2) as well. The two questions that still remain open are, first, how to choose the coefficients λ and, second, how to guarantee that a policy $\pi \in \Pi_\lambda^*$ indeed satisfies the constraints on the error probabilities in Theorem 3. The connection between the function ρ_λ and the properties of sequential tests using policies $\pi \in \Pi_\lambda^*$ is the subject of this section. This analysis is a core contribution of the dissertation and key to solving the sequential testing problem in a systematic manner.

The connection between (3.2) and (3.4) is established by showing that the error probabilities and the expected run-length of a sequential test that follows a policy $\pi \in \Pi_\lambda^*$ can be expressed in terms of ρ_λ and its partial derivatives with respect to z . This approach is based on the observation that the integral equations (2.41) and (2.42) that describe the error probabilities of a sequential test are rederived when calculating the partial derivatives of ρ_λ with respect to z_0 and z_1 . Before going into the details, some technical issues need to be addressed.

Lemma 1 *The sequence $(\rho_\lambda^n)_{n \in \mathbb{N}_0}$ with $\rho_\lambda^0 = g_\lambda$ and*

$$\rho_\lambda^n(z, \theta) = \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda^{n-1} dH_{z, \theta} \right\}$$

converges uniformly to ρ_λ on Ω_ρ .

The proof of Lemma 1 is laid down in Appendix A.5. Even though it is not central to the understanding of sequential tests, Lemma 1 is needed to guarantee that certain properties of the elements ρ_λ^n also hold for the limit ρ_λ .

The first result connecting ρ_λ to the performance measures in Section 2.4.3 is that ρ_λ is a weighted sum of α_π , β_π and γ_π .

Theorem 6 For ρ_λ as in Theorem 5, test statistics T^n as in Corollary 1 and all policies $\pi \in \Pi_\lambda^*$ it holds that

$$\rho_\lambda(z, \theta) = \gamma_\pi(z, \theta) + \lambda_0 z_0 \alpha_\pi(z, \theta) + \lambda_1 z_1 \beta_\pi(z, \theta), \quad (3.20)$$

where α_π , β_π and γ_π are defined in (2.44)–(2.46).

A proof of Theorem 5 can be found in Appendix A.6. It follows almost immediately from the definition of the cost function of the optimal stopping problem (3.11). The additional scaling of the error terms by z_0 and z_1 is due to the change in measure that was introduced in (3.5) in order to subsume all three terms under the same expectation operator.

While Theorem 5 relates ρ_λ to the weighted sum of the expected run-length and the error probabilities, it does not make a statement about the individual terms. In what follows, it is shown that the latter can be obtained by evaluating the *partial differentials* of ρ_λ with respect to z_0 and z_1 . The next theorem guarantees that these differentials exist.

Theorem 7 For every $\theta \in \Omega_\theta$, the function $\rho_\lambda(z, \theta)$, as defined in Theorem 5, is non-decreasing and concave in z , i.e., for all $z'_0 \geq z_0$ and $z'_1 \geq z_1$ it is the case that

$$\rho_\lambda(z', \theta) \geq \rho_\lambda(z, \theta)$$

and for all $z', z \in \mathbb{R}^+$ and $\kappa \in [0, 1]$ it holds that

$$\rho_\lambda(\kappa z' + (1 - \kappa)z, \theta) \geq \kappa \rho_\lambda(z', \theta) + (1 - \kappa) \rho_\lambda(z, \theta).$$

See Appendix A.7 for a proof. The fact that ρ_λ is concave in z is significant for two reasons: first, because it ensures that ρ_λ admits a generalized differential, and second, because it qualifies ρ_λ as a statistical similarity measure. The latter property is used in the derivation of the least favorable distributions for minimax robust sequential tests and is discussed in detail in Section 5.4. The notion of generalized differentials is introduced in the next definition.

Definition 2 Let $f: \Omega_f \subset \mathbb{R}^K \rightarrow \mathbb{R}$ be a convex function of the K -dimensional vector $y = (y_1, \dots, y_K)$. The set

$$\partial_{y_k} f(y) := \{c \in \mathbb{R}: f(y') - f(y) \leq c(y'_k - y_k) \quad \forall y' \in \mathbb{R}^K\}, \quad k \in \{1, \dots, K\},$$

is called the partial subdifferential of f with respect to y_k at y . If a function f_{y_k} exists such that

$$f_{y_k}(y) \in \partial_{y_k} f(y) \quad \forall y \in \Omega_f$$

then f_{y_k} is called a partial subderivative of f with respect to y_k . The set of all partial subderivatives f_{y_k} is denoted by $\partial_{y_k} f$, i.e.,

$$\partial_{y_k} f := \{f_{y_k} : f_{y_k}(y) \in \partial_{y_k} f(y) \quad \forall y \in \Omega_f\}.$$

Analogously, for a concave function $\tilde{f} : \Omega_{\tilde{f}} \subset \mathbb{R}^K \rightarrow \mathbb{R}$, the set

$$\partial_{y_k} \tilde{f}(y) := \left\{ c \in \mathbb{R} : \tilde{f}(y') - \tilde{f}(y) \geq c(y'_k - y_k) \quad \forall y' \in \mathbb{R}^K \right\}, \quad k \in \{1, \dots, K\},$$

is called the partial superdifferential of \tilde{f} with respect to y_k at y and a function \tilde{f}_{y_k} that satisfies

$$\tilde{f}_{y_k}(y) \in \partial_{y_k} \tilde{f}(y) \quad \forall y \in \Omega_{\tilde{f}}$$

is called a partial superderivative of \tilde{f} with respect to y_k . The set of all superderivatives \tilde{f}_{y_k} is denoted by $\partial_{y_k} \tilde{f}$.

Sub- and superdifferentials are well known tools in convex analysis and detailed studies of their properties can be found in all standard textbooks [Roc70, Boy04]. Adopting a convention from nonsmooth analysis [Cla83], they are collectively referred to as *generalized differentials* in this work. It is important to note that if f is differentiable in y , the partial generalized differential $\partial_{y_k} f(y)$ reduces to the regular partial derivative, i.e., the singleton set $\{f_{y_k}(y)\}$. Before stating the main result of this section, the following lemma about the generalized differential of the minimum of two concave functions needs to be in place.

Lemma 2 Let $f, g : \Omega_f \subset \mathbb{R}^K \rightarrow \mathbb{R}$ be two concave functions. It holds that

$$\partial_{y_k} \min\{g(y), f(y)\} = \begin{cases} \partial_{y_k} g(y), & g(y) < f(y), \\ co\{\partial_{y_k} g(y) \cup \partial_{y_k} f(y)\}, & g(y) = f(y), \\ \partial_{y_k} f(y), & g(y) > f(y), \end{cases}$$

where $co\{\mathcal{A}\}$ denotes the convex hull of the set \mathcal{A} .

Lemma 2 is a consequence of a more general result on the generalized differentials of the minimum (maximum) of concave (convex) functions that is due to Dubovitskiy and Milyutin [DM65] and can be found in, for example, [MIT03]. In words, Lemma 2 states that the partial generalized differential of the minimum of two concave functions is either given by the partial generalized differential of one of the two functions or by the convex hull both differentials. Since here y_k is a scalar, the latter is simply a closed interval on the real axis. It is now possible to state the main theorem of this section.

Theorem 8 Let ρ_λ be as defined in Theorem 5 and let Π_λ^* be the set of all policies that solve the optimal stopping problem (3.4).

1. For all $\pi \in \Pi_\lambda^*$ it holds that

$$\alpha_\pi \in \partial_{z_0} \frac{\rho_\lambda}{\lambda_0},$$

$$\beta_\pi \in \partial_{z_1} \frac{\rho_\lambda}{\lambda_1},$$

where α_π and β_π are defined in (2.25) and (2.26).

2. For all $\pi \in \Pi_\lambda^*$ and all $(z, \theta) \in \Omega_\rho$ it holds that

$$\{ \alpha_\pi(z, \theta) : \pi \in \Pi_\lambda^* \} = \partial_{z_0} \frac{\rho_\lambda}{\lambda_0}(z, \theta),$$

$$\{ \beta_\pi(z, \theta) : \pi \in \Pi_\lambda^* \} = \partial_{z_1} \frac{\rho_\lambda}{\lambda_0}(z, \theta).$$

Theorem 8 is proven in Appendix A.8. Its two parts correspond to a *global* and a *local* statement about the generalized derivatives/differentials of ρ_λ . The first part states that for all $\pi \in \Pi_\lambda^*$ the functions α_π and β_π are valid generalized derivatives of ρ_λ . The second part states that at every point $(z, \theta) \in \Omega_\rho$ the generalized differential of ρ_λ coincides with the set of all error probabilities that can be realized by using policies $\pi \in \Pi_\lambda^*$. It is worth noting that the remaining degrees of freedom in terms of the error probabilities are exclusively due to the randomization on $\partial\mathcal{S}$ and $\partial\mathcal{C}$ since on $\mathcal{S}, \bar{\mathcal{S}}$ and $\mathcal{C}, \bar{\mathcal{C}}$ the optimal policy is already uniquely determined by Π_λ^* .

The local statement in Theorem 8 cannot be extended to a global statement. In general, the set of all functions α_π and β_π with $\pi \in \Pi_\lambda^*$ is *not* identical to the set of all generalized derivatives of ρ_λ , i.e.,

$$\{ \alpha_\pi : \pi \in \Pi_\lambda^* \} \neq \partial_{z_0} \frac{\rho_\lambda}{\lambda_0} \quad \text{and} \quad \{ \beta_\pi : \pi \in \Pi_\lambda^* \} \neq \partial_{z_1} \frac{\rho_\lambda}{\lambda_0}.$$

The reason for this is that the set of partial generalized derivatives that can be represented in terms of the integral equations (2.41) and (2.42) is only a subset of all feasible derivatives. More precisely, the integral equation establishes a coupling between the local differentials so that they cannot be chosen independently of each other. This coupling reflects the fact that changing the randomization on the boundaries $\partial\mathcal{S}$ and $\partial\mathcal{C}$ also affects the error probabilities in states (z, θ) that are not part of the boundary. In particular, changing the randomization on the boundary can change the error probabilities $\alpha_\pi(1, 1, \theta_0)$ and $\beta_\pi(1, 1, \theta_0)$ and hence the overall performance of the corresponding sequential test. This coupling does not hold for the generalized partial

derivatives of ρ_λ , whose mathematical definition and existence is entirely unrelated to the interpretation of ρ_λ as a cost function of a possibly randomized sequential hypothesis test. Accordingly, they can be chosen independently over the entire state space Ω_ρ without violating any feasibility constraints. On the other hand, it holds by the second statement in Theorem 8 that in every state (z, θ) the local generalized differential corresponds to the interval of error probabilities that can be realized by policies $\pi \in \Pi_\lambda^*$. Hence, there is a strong coupling in the local sense, but no coupling in a global sense. This subtle difference between the error functions α_π, β_π and the partial generalized derivatives of ρ_λ is due to the fact that the former are defined by regular integral equations, while the latter are defined by set integral equations, as shown in Appendix A.8. In other words, (2.41) and (2.42) filter out those generalized derivatives that correspond to possible error probabilities of the underlying sequential test.

It is now possible to state a sufficient condition for a policy to solve the constrained sequential testing problem (3.2). According to Theorem 3, a testing policy is optimal if it solves the optimal stopping problem (3.11) and realizes the error probabilities α and β . That is, π solves (3.2) if it holds that $\pi \in \Pi_\lambda^*$, $\alpha_\pi(1, 1, \theta_0) = \alpha$ and $\beta_\pi(1, 1, \theta_0) = \beta$. Using the results from Theorem 8, this condition can be stated directly in terms of ρ_λ .

Corollary 4 *Let ρ_λ be as defined in Theorem 5 and let Π_λ^* be as defined in (2.22).*

1. *If it holds that*

$$\alpha \in \partial_{z_0} \frac{\rho_\lambda}{\lambda_0}(1, 1, \theta_0),$$

$$\beta \in \partial_{z_1} \frac{\rho_\lambda}{\lambda_1}(1, 1, \theta_0),$$

then there exists at least one $\pi \in \Pi_\lambda^$ that solves Problem (3.2), i.e., $\Pi_{\alpha, \beta}^* \subset \Pi_\lambda^*$.*

2. *If ρ_λ is differentiable in $(1, 1, \theta_0)$ and it holds that*

$$\alpha = \partial_{z_0} \frac{\rho_\lambda}{\lambda_0}(1, 1, \theta_0),$$

$$\beta = \partial_{z_1} \frac{\rho_\lambda}{\lambda_1}(1, 1, \theta_0),$$

then every policy $\pi \in \Pi_\lambda^$ solves Problem (3.2), i.e., $\Pi_{\alpha, \beta}^* = \Pi_\lambda^*$.*

3. *For all $\pi \in \Pi_{\alpha, \beta}^*$ it holds that*

$$E_P[\tau(\delta, \psi)] = \rho_\lambda(1, 1, \theta_0) - \lambda_0\alpha - \lambda_1\beta.$$

Corollary 4 is a direct consequence of Theorem 6 and Theorem 8. It can be used to transform the design of an optimal sequential test into the search for a function that solves the integral equation (3.12) and whose partial generalized differentials at $(z, \theta) = (1, 1, \theta_0)$ contain α and β . This idea is explained in more detail in the next section, where an approach to the design of optimal sequential tests is proposed that exploits this relation between α_π, β_π and ρ_λ .

3.4 Design of Optimal Sequential Tests

In this section, an approach to the design of sequential hypothesis tests is proposed that makes use of the results derived in the previous section to express the sequential testing problem (3.2) in terms of a system of integral equations that can be solved using well-known numerical techniques. First, the equation system that characterizes the optimal policy is stated, then two possible techniques for its solution are discussed.

Let the distributions P_0, P_1 and P as well as the target error probabilities α, β be given. The design procedure proposed in this thesis consists of two steps:

1. Solve the integral equation

$$\begin{aligned} \rho_\lambda(z, \theta) &= \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda dH_{z,\theta} \right\} \\ \text{s.t. } \alpha_0 &\in \partial_{z_0} \frac{\rho_\lambda}{\lambda_0}(1, 1, \theta_0), \quad \beta \in \partial_{z_1} \frac{\rho_\lambda}{\lambda_1}(1, 1, \theta_0) \end{aligned} \tag{3.21}$$

for ρ_λ and λ . If ρ_λ is differentiable in $(1, 1, \theta_0)$, it holds that $\alpha_\pi(1, 1, \theta_0) = \alpha$ and $\beta_\pi(1, 1, \theta_0) = \beta$ for all $\pi \in \Pi_\lambda^*$ so that any randomized or non-randomized test using a stopping rule of the form (3.16) and a decision rule of the form (3.8) solves (3.2). If ρ_λ is not uniquely differentiable in $(1, 1, \theta_0)$, proceed with Step 2.

2. Solve the system of integral equations

$$\begin{aligned} \alpha_\pi(z, \theta) &= \psi(z, \theta)\delta(z) + (1 - \psi(z, \theta)) \int \alpha_\pi dH_{z,\theta}^0 \\ \beta_\pi(z, \theta) &= \psi(z, \theta)(1 - \delta(z)) + (1 - \psi(z, \theta)) \int \beta_\pi dH_{z,\theta}^1 \\ \text{s.t. } \pi &\in \Pi_\lambda^*, \quad \alpha_\pi(1, 1, \theta_0) = \alpha, \quad \beta_\pi(1, 1, \theta_0) = \beta \end{aligned} \tag{3.22}$$

for α_π, β_π and $\pi = (\delta, \psi)$. By construction, the solution satisfies $\pi \in \Pi_{\alpha, \beta}^*$.

Numerical approaches to the solution of both steps are presented in detail in the following sections. For the vast majority of cases, a sufficiently accurate testing policy can be found by performing only the first step, i.e., determining the function ρ_λ . As mentioned before, sequential tests inherently avoid making decisions based on insignificant observations so that randomized decision rules are rarely ever needed. More critical is the optimal stopping rule, which can indeed be randomized. An example for this case is given in Section 3.5. For small error probabilities and reasonably smooth distributions, however, the effect of randomization is typically negligible in the sense that the bounds provided by the generalized differentials of ρ_λ guarantee that the error probabilities are within an acceptable interval or even unique up to the precision of the numerical solution. However, for the design of minimax robust sequential tests, which are discussed in Chapter 5, randomized stopping rules are much more important since robust test statistics often are constant over large regions of the state space. This effect leads to point masses being placed on individual states, which is a precondition for randomization to have any noticeable effects on the test performance.

3.4.1 Newton-like Methods

Since the task in both steps outlined above is to solve a system of nonlinear integral equations, Newton-like methods for nonlinear functions are a natural choice to obtain numerical solutions. The generic form of a nonlinear equation in \mathbb{R}^M , $M \in \mathbb{N}$, is

$$F(x) = 0, \quad (3.23)$$

where $F: \mathbb{R}^M \rightarrow \mathbb{R}^M$. In order to formulate the problem in Step 1 in this form, first a finite dimensional representation of the continuous function ρ_λ has to be introduced. A variety of different methods to do this can be found in the literature [Lue69, Pol97, AS06] and discussing their advantages and disadvantages in detail is beyond the scope of this thesis. However, based on the knowledge that ρ_λ is

- nondecreasing in z ,
- concave and
- non-smooth

some methods seem more approximate than others. For the numerical examples presented in Section 3.5, a spline approximation was used and showed good results. In

general, any procedure that allows one to obtain an approximate value of ρ_λ at every point $(z, \theta) \in \Omega_\rho$ given only a finite number of sampling points $(z_m, \theta_m)_{1 \leq m \leq M}$ with

$$\rho_m = \rho_\lambda(z_m, \theta_m)$$

can be used to approximate ρ_λ . The notation

$$\rho_\lambda \approx \mathcal{R}_\rho^M$$

is used as a generic way to state that ρ_λ is approximated by an M -dimensional representation with sample points $(z_m, \theta_m)_{1 \leq m \leq M}$. Owing to the relaxation, the integral equation in Step 1 of the solution procedure can only be satisfied with equality at the M sample points and hence reduces to the equation system

$$\begin{aligned} \rho_m - \min \left\{ g_\lambda(z_m), 1 + \int \mathcal{R}_\rho^M dH_{z_m, \theta_m} \right\} &= 0, \\ \min \left\{ (c - \alpha)^2 : c \in \partial_{z_0} \mathcal{R}_\rho^M(1, 1, \theta_0) \right\} &= 0, \\ \min \left\{ (c - \beta)^1 : c \in \partial_{z_1} \mathcal{R}_\rho^M(1, 1, \theta_0) \right\} &= 0. \end{aligned} \quad (3.24)$$

The constraints on the partial differentials of \mathcal{R}_ρ^M do not need to be of this exact form. Every constraint that guarantees $\alpha \in \partial_{z_0} \mathcal{R}_\rho^M(1, 1, \theta_0)$ and $\beta \in \partial_{z_1} \mathcal{R}_\rho^M(1, 1, \theta_0)$ is feasible. From a numerical point of view, however, differentiable penalty functions are typically preferable.

The equation system (3.24) consists of $M + 2$ equations in $M + 2$ unknowns, namely $(\rho_m)_{1 \leq m \leq M}$ and (λ_0, λ_1) . In order to evaluate the left hand side, one needs to be able to integrate \mathcal{R}_ρ^M with respect to all H_{z_m, θ_m} and to determine the generalized differential of \mathcal{R}_ρ^M at $(1, 1, \theta_0)$. The latter is particularly simple if a bilinear approximation on a rectangular grid¹ is used. In this case it holds that

$$\partial_{z_i} \mathcal{R}_\rho^M(1, 1, \theta_0) = \text{co} \left\{ \frac{\Delta_i^+ \mathcal{R}_\rho^M(1, 1, \theta_0)}{\Delta z_i}, \frac{\Delta_i^- \mathcal{R}_\rho^M(1, 1, \theta_0)}{\Delta z_i} \right\}, \quad i = 0, 1,$$

where Δz_0 and Δz_1 are chosen equal to or smaller than the step sizes of the grid around $(1, 1, \theta_0)$ and

$$\begin{aligned} \Delta_0^- \mathcal{R}_\rho^M(z, \theta) &:= \mathcal{R}_\rho^M(z, \theta) - \mathcal{R}_\rho^M(z_0 - \Delta z_0, z_1, \theta), \\ \Delta_0^+ \mathcal{R}_\rho^M(z, \theta) &:= \mathcal{R}_\rho^M(z_0 + \Delta z_0, z_1, \theta) - \mathcal{R}_\rho^M(z, \theta), \\ \Delta_1^- \mathcal{R}_\rho^M(z, \theta) &:= \mathcal{R}_\rho^M(z, \theta) - \mathcal{R}_\rho^M(z_0, z_1 - \Delta z_1, \theta), \\ \Delta_1^+ \mathcal{R}_\rho^M(z, \theta) &:= \mathcal{R}_\rho^M(z_0, z_1 + \Delta z_1, \theta) - \mathcal{R}_\rho^M(z, \theta). \end{aligned}$$

¹Although it simplifies the calculation of the partial derivatives, sampling the likelihood ratios on a regular grid comes with its own problems; see the remark in Section 3.4.3.

If the grid can be chosen sufficiently fine, bilinear splines offer a good trade-off between accuracy and computational cost. In particular the fact that linear approximations do not impose smoothness is important for the design of randomized policies. In contrast, if it is known *a priori* that an optimal deterministic policy exists, this knowledge can be exploited by using, for example, cubic splines, which result in a more accurate approximation of smooth functions. In this case it is also advisable to approximate d_λ , as defined in (3.17), instead of ρ_λ . While the latter is nonsmooth on $\partial\mathcal{S}$ by definition, the function d_λ is differentiable everywhere in case no randomization is required. A formal proof of this property is omitted, but examples can be found in the next section.

For the evaluation of the integrals over \mathcal{R}_ρ^M either analytical or numerical methods can be used, depending on the distributions and the method used to construct \mathcal{R}_ρ^M . The most generic approach is to evaluate

$$\int \mathcal{R}_\rho^M \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) p_\theta(x) d\mu(x) \quad (3.25)$$

by means of some numerical integration technique. Note that the integral has to be calculated for all sampling points $(z_m, \theta_m)_{1 \leq m \leq M}$.

In case randomization turns out to be necessary after having solved Step 1, the procedure outlined above can also be used to solve the system of integral equations in Step 2. First, a finite dimensional representation for α_π and β_π on an appropriately chosen grid is introduced, then the resulting equation system is solved for the unknown variables. Depending on the shape of $\partial\mathcal{S}$ and $\partial\mathcal{C}$ it can be useful to introduce a new grid that samples these boundaries more densely.

The solution method used for some of the examples in Section 3.5 is the “bad” version of Broyden’s method [Bro65, Gri12]. It is a quasi Newton method that iteratively approximates the inverse of the Jacobian matrix of the function F in (3.23) and does neither require an explicit calculation of derivatives nor the solution of a large linear equation system. The most costly tasks in each iteration are the evaluation of the left hand side of (3.24), or the corresponding equations in Step 2, and a matrix multiplication involving the approximated inverse of the Jacobian matrix. Broyden’s bad method is locally convergent at a superlinear rate, but is not guaranteed to converge globally [Gri12]. Nevertheless, using g_λ as an initial guess for ρ_λ consistently resulted in fast convergence, typically after fewer than 100 iterations.

For large to moderate error probabilities the approach of solving (3.21) directly is reasonable and yields good results. For small error probabilities, however, the calculation of the (generalized) derivatives can become numerically unstable. In the next section,

a solution approach based on linear programming is proposed that does not require the calculation of derivatives, is guaranteed to converge and leverages the high efficiency of state-of-the-art linear programming solvers.

3.4.2 Linear Programming

The idea underlying the linear programming approach to sequential hypothesis testing, see [FZ15a], is to apply the machinery of Lagrangian duality to (3.2). While in Section 2.3 λ_0 and λ_1 were introduced as arbitrary cost coefficients, in this section they arise explicitly as Lagragian multipliers when formulating the dual problem of (3.2).

The Lagrangian dual problem of (3.2) is given by

$$\max_{\lambda \in \mathbb{R}_+^2} L_{\alpha, \beta}(\lambda), \quad (3.26)$$

where

$$\begin{aligned} rClL_{\alpha, \beta}(\lambda) &= \min_{(\delta, \psi) \in \Pi} \left\{ E_P[\tau(\psi)] + \lambda_0(\alpha_0(\delta, \psi) - \alpha) + \lambda_1(\beta(\delta, \psi) - \beta) \right\} \\ &= \min_{(\delta, \psi) \in \Pi} \left\{ E_P[\tau(\psi)] + \lambda_0\alpha(\delta, \psi) + \lambda_1\beta(\delta, \psi) \right\} - \lambda_0\alpha - \lambda_1\beta \\ &\stackrel{(3.11)}{=} \min_{\psi \in \Delta^{\mathbb{N}_0}} E_P[\tau + g_\lambda(z^\tau)] - \lambda_0\alpha - \lambda_1\beta \\ &\stackrel{(3.14)}{=} \rho_\lambda(1, 1, \theta_0) - \lambda_0\alpha - \lambda_1\beta. \end{aligned}$$

$L_{\alpha, \beta}$ is concave in λ by construction. However, the equivalence between (3.26) and (3.2), i.e., the absence of a duality gap, is not obvious. The following Theorem is useful to show that both problems are indeed equivalent.

Theorem 9 *Let ρ_λ be as defined in Theorem 5. For all $(z, \theta) \in \Omega_\rho$ it holds that*

$$\begin{aligned} \partial_{z_0} \frac{\rho_\lambda}{\lambda_0}(z, \theta) &= \partial_{\lambda_0} \frac{\rho_\lambda}{z_0}(z, \theta), \\ \partial_{z_1} \frac{\rho_\lambda}{\lambda_1}(z, \theta) &= \partial_{\lambda_1} \frac{\rho_\lambda}{z_1}(z, \theta), \end{aligned}$$

where ∂_{λ_i} denotes the partial generalized differential with respect to λ_i .

Lemma 9 is proven in Appendix A.9. It can be understood in the sense that ρ_λ is in fact a function of the element-wise product λz so that the derivatives with respect to λ

and z are identical up to a scaling factor. In this view, each λ_i is a multiplicative offset by which ρ_λ has been shifted to make sure that the initial state always corresponds to the point $(1, 1, \theta_0)$ instead of $(\lambda_0, \lambda_1, \theta_0)$.

The scaling of z by λ has another instructive interpretation. The examples in the upcoming section show that a sequential test is stopped once one of the likelihood ratios z_0, z_1 is small enough to reject the corresponding hypothesis. The bigger λ_i is, the smaller z_i has to become in order to compensate for the offset. In this sense, λ determines the initial distance of the test statistic from the stopping region.

By means of Lemma 9 it is straightforward to show that (3.26) and (3.21) are indeed equivalent.

Corollary 5 *If ρ_λ and λ solve*

$$\begin{aligned} & \max_{\lambda \in \mathbb{R}_+^2} \rho_\lambda(1, 1, \theta_0) - \lambda_0 \gamma_0 - \lambda_1 \gamma_1 \\ \text{s.t. } & \rho_\lambda(z, \theta) = \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda dH_{z, \theta} \right\}. \end{aligned} \quad (3.27)$$

they also solve (3.21).

Corollary 5 is a direct consequence of Theorem 9 and a proof is given in Appendix A.10. The simple trick at this point is to relax the equality constraint to an inequality and add ρ_λ to the set of free variables. It yields the main result of this section.

Theorem 10 *Let $\mathcal{L}_+^H = \mathcal{L}_+^H(\Omega_\rho)$ be the set of nonnegative $H_{z, \theta}$ -integrable functions on Ω_ρ . The problem*

$$\begin{aligned} & \max_{\lambda \in \mathbb{R}_+^2, \rho_\lambda \in \mathcal{L}_+^H} \rho_\lambda(1, 1, \theta_0) - \lambda_0 \alpha - \lambda_1 \beta \\ \text{s.t. } & \rho_\lambda(z, \theta) \leq \lambda_0 z_0, \\ & \rho_\lambda(z, \theta) \leq \lambda_1 z_1, \\ & \rho_\lambda(z, \theta) \leq 1 + \int \rho_\lambda dH_{z, \theta}. \end{aligned} \quad (3.28)$$

is equivalent to problem (3.21) in the sense that

$$P(\{(z^n, \theta_n) \in \Omega_\rho : \rho_\lambda^*(z^n, \theta_n) \neq \tilde{\rho}_\lambda(z^n, \theta_n)\}) = 0$$

for all $n \in \mathbb{N}_0$, where ρ_λ^ denotes the solution of (3.21) and $\tilde{\rho}$ the solution of (3.28).*

See Appendix A.11 for a proof. The optimization problem in Theorem 10 is linear in ρ_λ and λ and does not require the calculation of partial derivatives of ρ_λ . Again, introducing a grid $(z_m, \theta_m)_{1 \leq m \leq M}$ on Ω_ρ yields the discretized version of the linear problem in Theorem 10:

$$\max_{\lambda \in \mathbb{R}_+^2, \rho \in \mathbb{R}_+^M} \rho_{m^*} - \lambda_0 \alpha - \lambda_1 \beta \quad \text{s.t.} \quad \rho \leq \lambda_0 z_0, \quad \rho \leq \lambda_1 z_1, \quad \rho \leq 1 + \rho H, \quad (3.29)$$

where ρ, z_0 and z_1 are row vectors of size $1 \times M$ and H is a matrix of size $M \times M$ that corresponds to the family of measures $\{H_{z,\theta} : (z, \theta) \in \Omega_\rho\}$. The exact form of H depends on the representation of ρ_λ and, if no analytic solutions exists, the approximation used for the evaluation of the integral. The index m^* in the objective function needs to be chosen such that $(z_{m^*}, \theta_{m^*}) = (1, 1, \theta_0)$.

When solving problem (3.29) numerically, it can be the case that on some region $\mathcal{E} \subset \Omega_\rho$ the inequality constraint is not fulfilled with equality, even though $P((z^n, \theta^n) \in \mathcal{E})$ is not strictly zero for all $n \geq 0$. This effect is due to numerical inaccuracies and occurs when the coupling between $\rho_\lambda(1, 1, \theta_0)$ and $\rho_\lambda(\mathcal{E})$ is so weak that the contribution of \mathcal{E} to $\rho_\lambda(1, 1, \theta_0)$ is smaller than the precision of the solver. As a result, the stopping region can exhibit some areas, where the cost for continuing is erroneously declared to be smaller than that for stopping. However, given a reasonably precise solver, these artifacts occur only in regions of the state space that are highly unlikely to ever be reached during a test and usually are a purely cosmetic problem. In any case, the procedure given in the proof of Theorem 10 can be used to construct a valid solution from the inaccurate one. Alternatively, a regularization term can be added to the maximization that explicitly enforces equality—see Appendix A.12 for details.

After having determined ρ_λ via linear programming, the optional randomization step in the test design can be performed as outlined in the previous section. Due to the bilinear terms in (3.22), the problem of finding an optimal randomization cannot be cast as a linear program. However, since (3.22) does not require the calculation of derivatives and only needs to be solved accurately on $\partial\mathcal{S} \cup \partial\mathcal{C}$, it is typically a less challenging problem than (3.21).

3.4.3 Remarks

Before presenting the solutions of some example problems in the next section, some additional remarks concerning the numerical solution of (3.21) are in order.

First, the function ρ_λ should not be approximated on a regular likelihood-ratio grid. The reason for this is that the likelihood ratios z_0 and z_1 typically need to be sampled

with a much finer granularity on the interval $(0, 1]$ than on the interval $(1, \infty)$. In order to avoid irregular grids, some bijective function can be applied to z_0 and z_1 that warps the real line appropriately. An obvious choice is to use the log-likelihood ratios instead of the likelihood ratios, i.e., to define

$$s_i = \log z_i, \quad i = 0, 1,$$

and formulate (3.21) in terms of s . Another alternative is to concatenate the logarithm and a logistic function, which yields

$$s'_i = \frac{1}{1 + z_i^{-b}}, \quad i = 0, 1,$$

where $b > 0$ can be chosen freely. The advantage of the latter option is that it maps the real line onto the unit interval, which allows for a straightforward discretization irrespective of how P_0 , P_1 , P or α and β have been chosen.

For the numerical calculation of most of the results presented in the next section the log-likelihood ratio was sampled on a regular grid. Choosing the support of this grid appropriately is typically not a difficult task. Any available bound or approximation can be used to get an estimate for which region of Ω_ρ needs to be sampled in order to obtain sufficiently accurate results.

The appropriate method for solving the integrals in (3.21) and (3.22) depends on whether nonlinear equation solvers or the linear programming approach is used. Since the former allows to evaluate the current approximation for ρ_λ at arbitrary points on the chosen support, the integral (3.25) can be evaluated in every iteration by means of any numerical integration technique. The LP approach is more restricted in the sense that the matrix H in (3.29) needs to be calculated explicitly. A simple technique to discretize the integral in this case is to approximate

$$\rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right)$$

by a weighted average of ρ_λ at the K closest sample points. For the examples in the next section, the one- and two-dimensional trapezoidal rule [Atk89] was used.

3.5 Examples and Numerical Results

Several example problems are presented in this section to illustrate the proposed approach to the design of sequential hypothesis tests. The first, most basic example is

to detect a shift of known size in the mean of a normal distribution with known variance. For this problem, three optimal sequential tests are designed that minimize the expected run-length under different distributions P . The second example is a test for Bernoulli distributions and is included primarily to illustrate the effects of randomization. Finally, the optimal sequential tests for a Markov chain with observable states and the Gaussian AR(1) process are presented as examples of tests for data-dependent distributions.

As a reference for comparison with the optimal results, tests are used whose thresholds are calculated according to Wald's approximation (2.19). The thresholds and the corresponding policy are denoted by A_w , A_w and π_w , respectively. The functions α_π , β_π and γ_π are used as compact notation for the error probabilities and the expected run-length as a function of a certain policy. All quantities marked with a tilde have been obtained by means of Monte Carlo simulations with 10^5 runs.

The results are visualized either in the likelihood ratio domain, or the log-likelihood ratio domain, depending on which representation is more appropriate. Typically, the logarithmic domain is more convenient to present graphical results. However, some of the characteristic properties of ρ_λ are lost under the logarithmic transformation so that switching between both domains is necessary.

3.5.1 Mean Shifted Gaussian Distributions

The classic problem of testing for a unit shift in the mean of a Gaussian distribution is a good example to introduce the proposed approach to the design of sequential hypothesis tests. It corresponds to the two hypotheses

$$\begin{aligned}\mathcal{H}_0: \quad X_n &\sim \mathcal{N}(-0.5, \sigma), \\ \mathcal{H}_1: \quad X_n &\sim \mathcal{N}(0.5, \sigma),\end{aligned}\tag{3.30}$$

where $\mathcal{N}(\eta, \sigma)$ denotes the Gaussian probability measure with mean η and standard deviation σ . For this example it is assumed that $\sigma = 1$ and that all X_n are independent and identically distributed (i.i.d.) under both hypotheses. This means that no additional information about the previous samples is necessary for the look-ahead step so that $\Omega_\theta = \emptyset$ and ρ_λ becomes a function of z only.

Let the target error probabilities for now be chosen symmetrically as

$$\alpha = \beta = 0.05$$

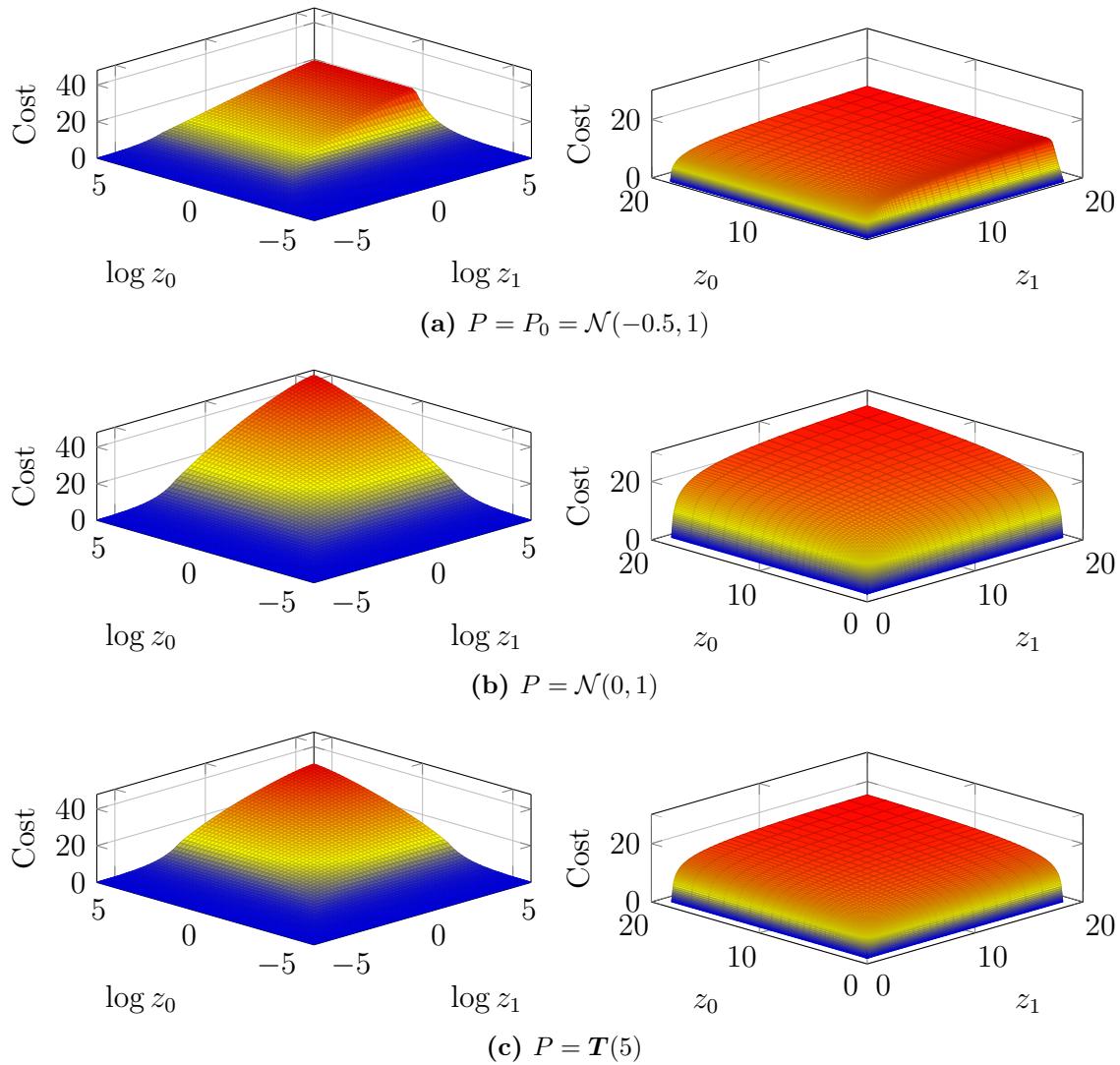


Figure 3.1: Mean Shifted Gaussians. Examples of cost functions ρ_λ of optimal sequential tests with error probabilities $\alpha = \beta = 0.05$ and minimum expected run-length under different choices for P .

and assume that an optimal test for \mathcal{H}_0 against \mathcal{H}_1 should be designed such that the expected run-length is minimized under

- 1) $P = P_0 = \mathcal{N}(-0.5, 1),$
 - 2) $P = \mathcal{N}(0, 1),$
 - 3) $P = \mathbf{T}(5),$
- (3.31)

where $\mathbf{T}(\nu)$ denotes Student's t-distribution with ν degrees of freedom. The cost functions ρ_λ corresponding to the three cases are depicted in Figure 3.1. While the subtle differences in shape are more obvious in the logarithmic domain, the concavity of ρ_λ can only be seen in the linear domain. The asymmetry in ρ_λ for $P = P_0$ is due to z_0^n being constant so that only the slice $z_0 = 1$ is relevant for the actual test.

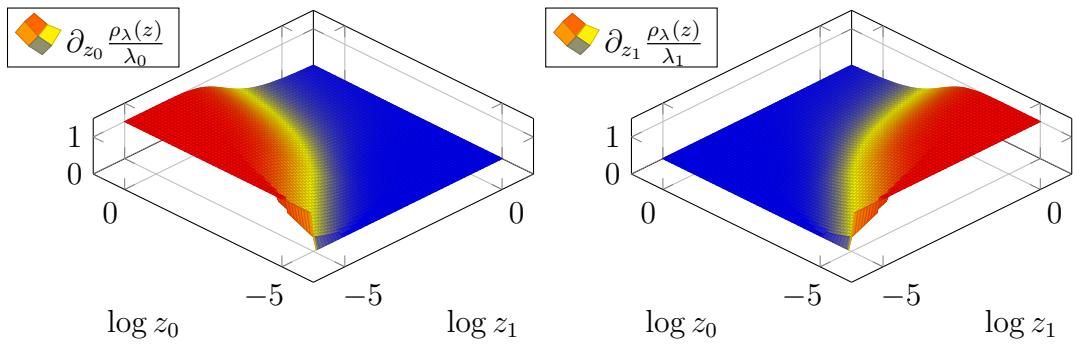


Figure 3.2: Mean Shifted Gaussians. Partial derivatives of ρ_λ with respect to z_0 (left) and z_1 (right) for $P = \mathcal{N}(0, 1)$.

For $P = \mathcal{N}(0, 1)$, the partial derivatives of ρ_λ with respect to z_0 and z_1 are depicted in Figure 3.2. Note that $\partial_{z_0}\rho_\lambda$ and $\partial_{z_1}\rho_\lambda$ were scaled by λ_0 and λ_1 , respectively, so that, by Theorem 8, they correspond to the error probabilities of a sequential test starting in the respective state. It can be seen how the latter are almost zero on regions where both likelihood ratios are relatively large, but rapidly increase towards the corners where one of the likelihood ratios is small and the other one is large. In fact, the areas of large error probabilities are blurred versions of the intersection of the stopping region with the critical region or its complement. The closer the test statistic approaches a region leading to an erroneous decision, the higher the corresponding error probability.

Figure 3.4 shows the stopping region, the critical region and their complements under the three distributions given in (3.31). More precisely, it shows the boundaries where d_λ and g_λ intersect, i.e., the costs for stopping and continuing the test are identical. The testing procedure is as follows. The initial state of the test statistic is $z = (1, 1)$, which corresponds to the origin in the logarithmic domain. It is marked with a black dot in Figure 3.4. With every new sample, the test statistic is updated and changes its state, that is, it performs a random walk in the z_0 - z_1 -plane. See Figure 3.3 for an illustrative example. As soon as the test statistic enters the stopping region, the value of $\lambda_0 z_0$ is compared to the value of $\lambda_1 z_1$ and

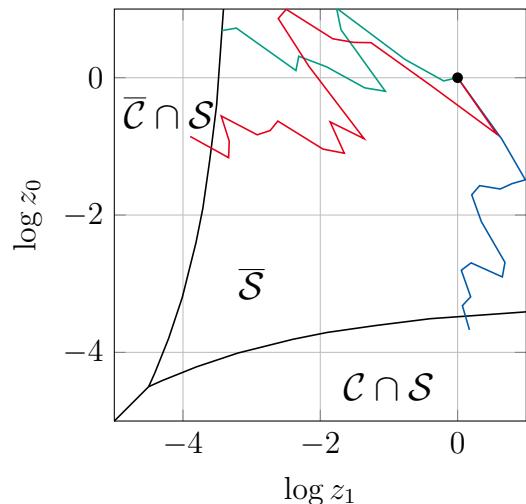


Figure 3.3: Mean Shifted Gaussians. Examples for random walks of a two-dimensional test statistic in the z -plane for $P = \mathcal{N}(0, 1)$ and $\alpha = \beta = 0.01$.

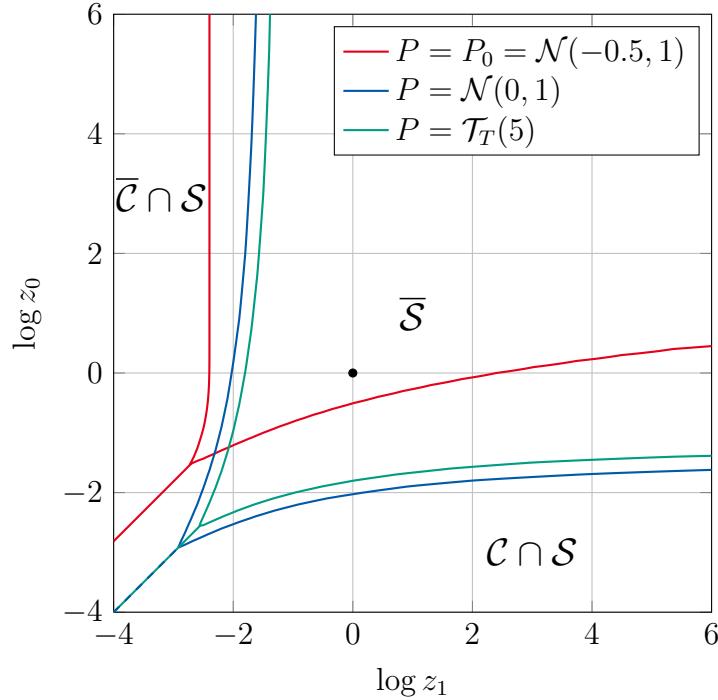


Figure 3.4: Mean Shifted Gaussians. Boundaries of the stopping regions of three optimal sequential test with error probabilities $\alpha = \beta = 0.05$ and minimum expected run-length under different choices for P .

the hypothesis corresponding to the smaller value is rejected. This last step is nothing but a likelihood ratio test of the form (2.6) with threshold λ_0/λ_1 . By tendency, the test statistic moves towards smaller likelihood ratios, i.e., towards the lower left corner of the coordinate system used for the plots. This is to be expected since the likelihood ratios are closely related to the error probabilities, which should decrease as the test progresses.

In the case $P = P_0$, the random walk is performed on the line defined by $\log z_0 = 0$. Therefore, the boundary of the stopping region reduces to two individual points, which are the thresholds A and B of the sequential probability ratio test. In order to inspect this relation between the cost function and the likelihood ratio thresholds in more detail, it is helpful not to consider the two-dimensional function ρ_λ , but a one-dimensional slice of it. The cost functions $g_\lambda(z_1) := g_\lambda(1, z_1)$ and $d_\lambda(z_1) := d_\lambda(1, z_1)$ are depicted in Figure 3.5; the points of intersection determine the optimal thresholds A^* and B^* .

The scaled derivative of $\rho_\lambda(z_1) := \rho_\lambda(1, z_1)$ is shown in Figure 3.6. By Theorem 8, it is identical to the probability of type II errors. Consequently, it is equal to zero on the interval $[\log A^*, \infty)$ and equal to one on the interval $(-\infty, \log B^*]$, where A^* and B^* are assumed to be included in the stopping region. In the initial state, $z_1 = 1$, the

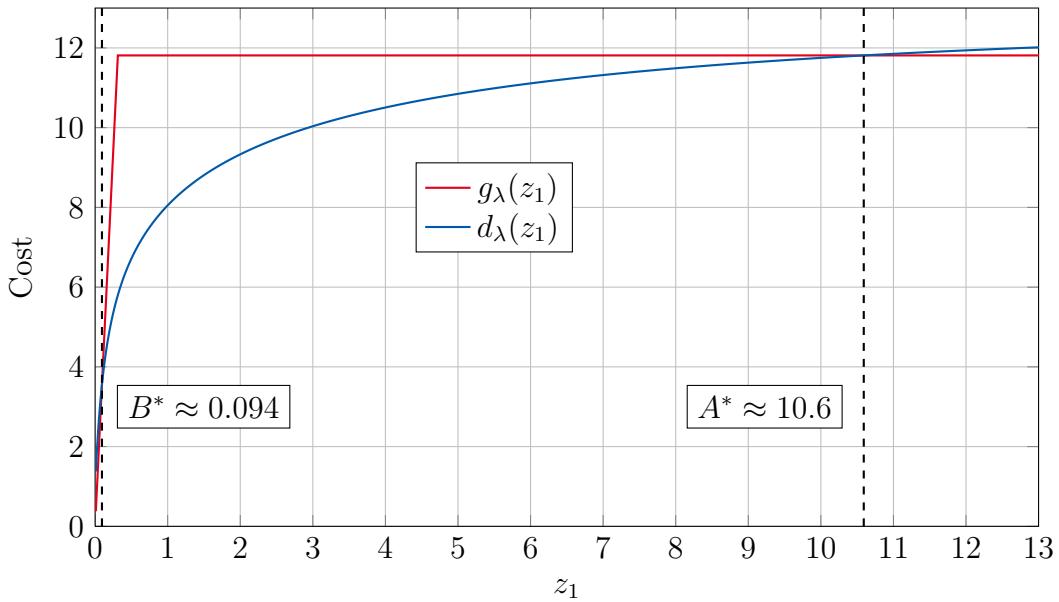


Figure 3.5: Mean Shifted Gaussians. Cost functions for stopping and continuing the optimal sequential test. Here $P = P_0$ and the error probabilities were chosen to be $\alpha = \beta = 0.05$. The optimal likelihood ratio thresholds correspond to the intersection points of the cost functions.

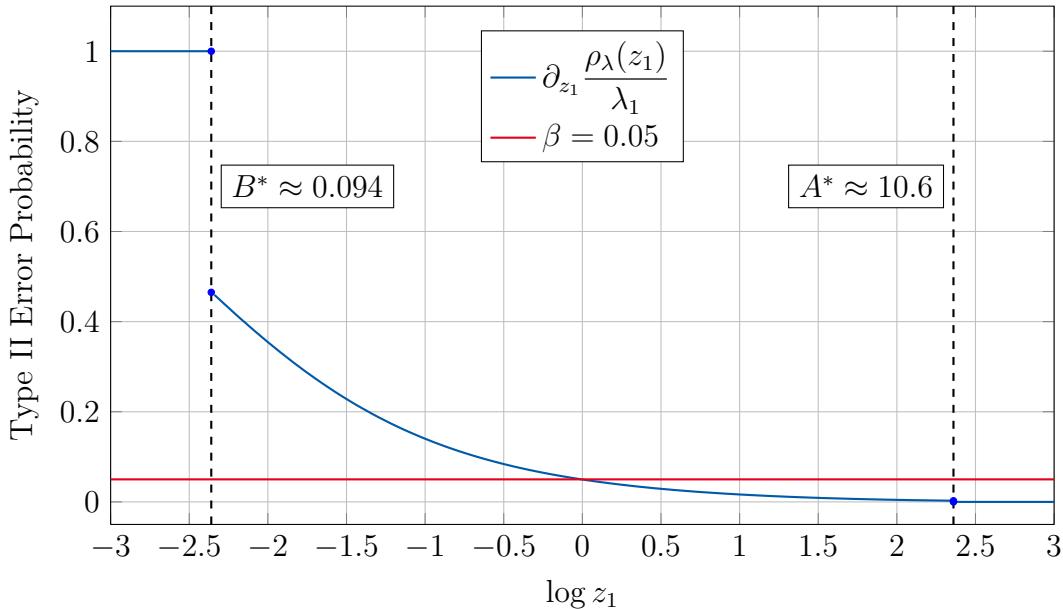


Figure 3.6: Mean Shifted Gaussians. Scaled generalized derivative of the cost functions ρ_λ depicted in Figure 3.5.

error probability of the test equals the target error probability of $\beta = 0.05$. The closer the test statistic approaches the lower threshold, the larger becomes the type II error probability. At $z_1 = B^*$, it holds that $g_\lambda(B^*) = d_\lambda(B^*)$ and an arbitrary randomization

Target		Thresholds		Error Prob.		Avg. and Exp. Run-Length		
α	β	A^*	B^*	$\tilde{\alpha}_{\pi^*}(1)$	$\tilde{\beta}_{\pi^*}(1)$	$\tilde{\gamma}_{\pi^*, P_0}(1)$	$\tilde{\gamma}_{\pi^*, P_1}(1)$	$\gamma_{\pi^*, P_0}(1)$
0.1		± 1.62		0.0995	0.0996	3.77	3.78	3.78
0.05		± 2.36		0.0505	0.0497	5.57	5.57	5.58
0.01		± 4.03		0.0097	0.0099	9.31	9.29	9.28
0.1	0.01	1.70	-3.93	0.1012	0.0097	7.93	4.69	7.91

Table 3.1: Mean Shifted Gaussians. Target error probabilities α, β , optimal log-likelihood ratio thresholds A^*, B^* , empirical error probabilities $\tilde{\alpha}_{\pi^*}(1), \tilde{\beta}_{\pi^*}(1)$, average run-length $\tilde{\gamma}_{\pi^*}(1)$ under both hypothesis and expected run-length $\gamma_{\pi^*}(1)$ under \mathcal{H}_0 .

can be applied, i.e., $\psi^*(B^*) \in [0, 1]$. For $\psi^*(B^*) = 1$, the test is surely stopped so that the error probability is equal to one. For $\psi^*(B^*) = 0$, the test is surely continued. As can be seen, this results in a type II error probability of roughly 46%. This relatively small number, given that the lower threshold has actually been reached, is due to the fact that under P_1 the log-likelihood ratio $\log z_1$ admits a strong drift towards the upper threshold. Therefore, the test still ends in a correct decision with a relatively high probability, even for likelihood ratio values close or equal to B^* . On the other hand, as the test statistic approaches the upper threshold, the probability of an erroneous decision becomes negligibly small, irrespective of the randomization.

In Table 3.1 and Table 3.2, the performance of the optimal test is compared to that of a test based on Wald's approximations (2.19). In anticipation of a notation used in Chapter 5, the expected run-length of a test using policy π under distribution P is denoted by $\gamma_{\pi, P}$. The comparison shows that the optimal test can perform significantly better. In particular in cases where large overshoots over the threshold can be expected, i.e., for medium to large error probabilities, the average run-length is reduced by up to 25%. For smaller error probabilities the improvement is less pronounced. This effect is expected, given the asymptotic optimality of Wald's approximations.

In Figure 3.7, the cost functions for optimal tests of P_0 against P_1 under $P = P_0$ are depicted for different error probabilities. Here the functions are given in the log-likelihood ratio domain, where it is easier to identify the thresholds. It is noteworthy that even if the target error probabilities are chosen to be identical, the values of the optimal cost coefficients λ differ significantly. This difference can be explained as follows: Since the expected run-length is minimized under the null hypothesis, the likelihood ratio admits a permanent drift towards the lower threshold. Choosing the latter closer to zero significantly reduces the expected run-length at the cost of an increased probability of

Target		Thresholds		Error Prob.		Avg. Run-Length		efficiency
α	β	A_W	B_W	$\tilde{\alpha}_{\pi_W}(1)$	$\tilde{\beta}_{\pi_W}(1)$	$\tilde{\gamma}_{\pi_W, P_0}(1)$	$\tilde{\gamma}_{\pi_W, P_1}(1)$	
0.1		± 2.20		0.0572	0.0578	5.19	5.18	0.73
0.05		± 2.94		0.0286	0.0284	6.94	6.93	0.80
0.01		± 4.60		0.0056	0.0057	10.51	10.50	0.89
0.1	0.01	2.29	-4.50	0.0562	0.0055	9.54	5.92	0.83

Table 3.2: Mean Shifted Gaussians. Target error probabilities α, β , log-likelihood ratio thresholds A_W, B_W obtained by Wald's approximation, empirical error probabilities $\tilde{\alpha}_{\pi_W}(1), \tilde{\beta}_{\pi_W}(1)$ and average run-length $\tilde{\gamma}_{\pi_W}(1)$ under both hypothesis. The last column gives the relative loss in the average run-length compared to the optimal test, i.e., $\tilde{\gamma}_{\pi^*}(1)/\tilde{\gamma}_{\pi_W}(1)$.

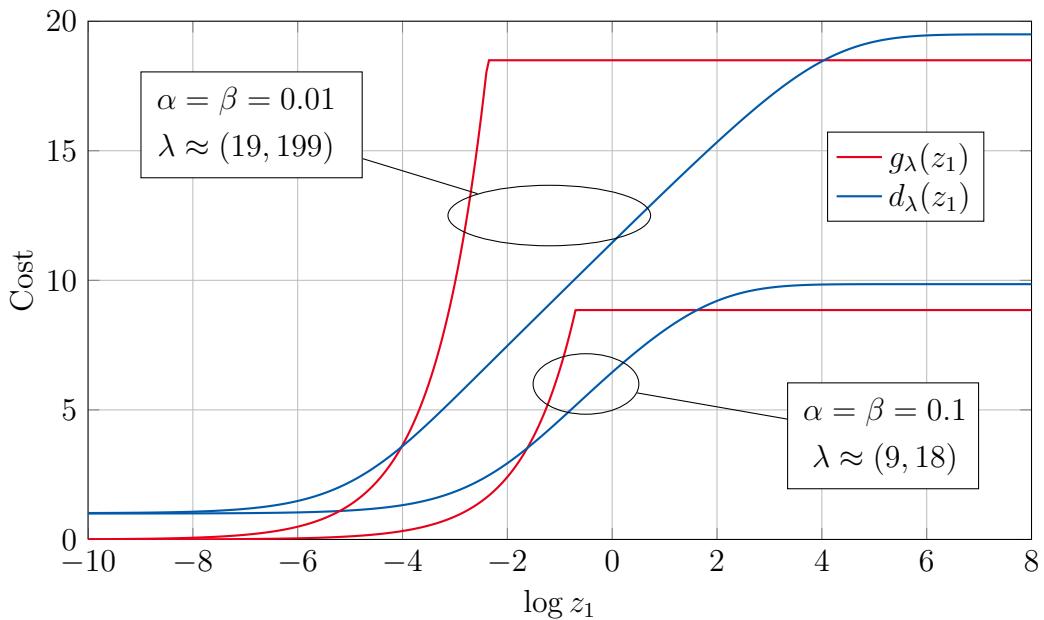


Figure 3.7: Mean Shifted Gaussians. Cost functions for stopping and continuing the optimal test under $P = P_0$ with error probabilities $\alpha = \beta = 0.1$ and $\alpha = \beta = 0.01$.

type II errors. The probability of type I errors, by contrast, is mainly determined by the upper threshold, which has very little influence on the expected run-length under P_0 . Consequently, type I errors have to be penalized much higher than type II errors if both are supposed to occur with the same probability. This asymmetry highlights problems with approaches that assume the cost coefficients to be given *a priori* or simply assume both error types to be equally costly.

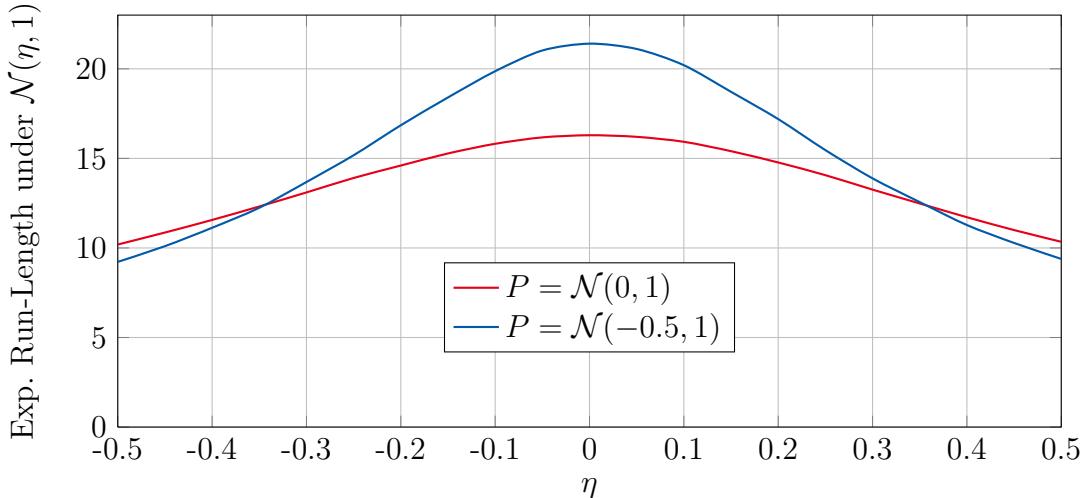


Figure 3.8: Mean Shifted Gaussians. Expected run-length of optimal tests with target error probabilities $\alpha = \beta = 0.01$ for \mathcal{H}_0 against \mathcal{H}_1 under $\mathcal{N}(\eta, 1)$ as a function of η .

If P is not chosen identical to either P_0 or P_1 , the drift towards one of the thresholds becomes smaller. For $P = \mathcal{N}(0, 1)$, it even holds that

$$E_P \left[\log \frac{p_1(x)}{p_0(x)} \right] = 0$$

so that $\log z_1^n$ is a zero-drift sequence under P . As a consequence, the expected run-length of a regular sequential probability ratio test increases significantly in this case. This is illustrated in Figure 3.8, which depicts the expected run-length of two optimal sequential tests with target error probabilities $\alpha = \beta = 0.01$ under $\mathcal{N}(\eta, 1)$ as a function of η . The blue curve shows the performance of a test that is designed to have minimum expected run-length under $P = P_0$. The red curve shows the performance of a test designed to have minimum expected run-length under $P = \mathcal{N}(0, 1)$, which is the zero-drift scenario. The latter test can be seen to require slightly more samples under P_0 and P_1 , but to have a much flatter performance profile over the interval $\eta \in [-0.5, 0.5]$. In other words, the test is more *robust* towards deviations of the true distribution from the two hypotheses. This observation motivates the design of robust sequential tests, which are introduced and discussed in Chapter 5.

3.5.2 Bernoulli Distributions

For this example it is assumed that

$$\begin{aligned} \mathcal{H}_0: \quad X_n &\sim \mathbf{B}(0.2), \\ \mathcal{H}_1: \quad X_n &\sim \mathbf{B}(0.8), \end{aligned} \tag{3.32}$$

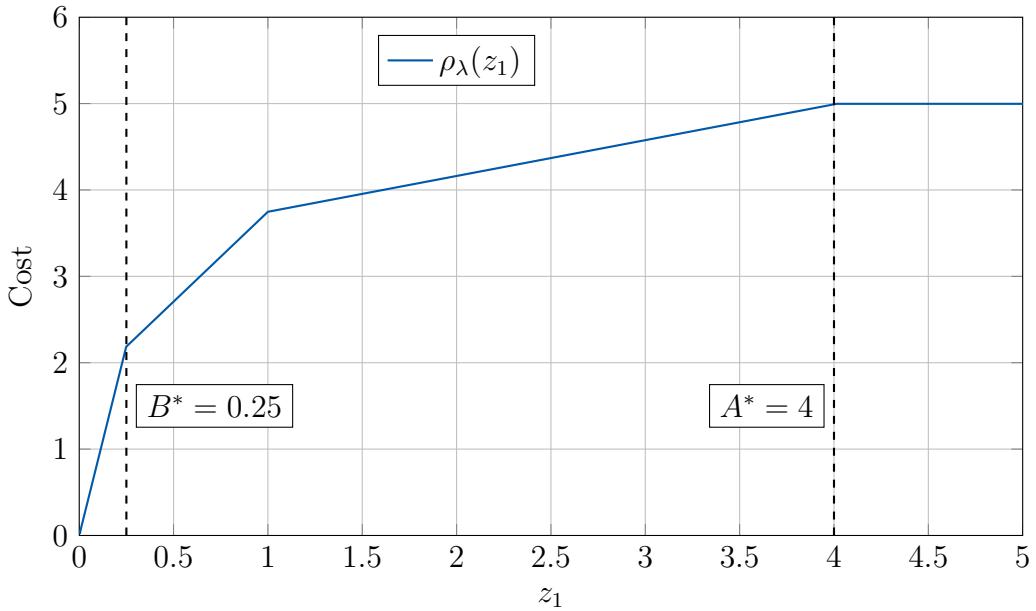


Figure 3.9: Bernoulli Distributions. Cost function of the optimal sequential test for the two Bernoulli distributions in (3.32) with target error probabilities $\alpha = \beta = 0.05$.

where $\mathbf{B}(\kappa)$ denotes a Bernoulli distribution with success probability κ and all X_n are assumed to be i.i.d. random variables. Due to the simplicity of the hypotheses, designing a sequential test for \mathcal{H}_0 against \mathcal{H}_1 is rather straightforward and does hardly warrant the application of optimal stopping theory in the first place. However, owing to this simplicity and the discrete nature of the the Bernoulli distribution, the example illustrates nicely the relation between the partial differentials of ρ_λ and the randomization rules of the underlying sequential test.

Assume that an optimal sequential hypothesis test for \mathcal{H}_0 against \mathcal{H}_1 needs to be designed with minimum expected run-length under P_0 and target error probabilities $\alpha = \beta = 0.1$. Since the likelihood ratio $p_1(x)/p_0(x)$ only takes on the values 4 and $0.25 = 4^{-1}$, the thresholds A and B need to be chosen as powers of 4 as well. It is not hard to show that $A = 1/B = 16$ are the smallest possible non-randomized thresholds that satisfy the given constraints on the error probabilities. However, this choice results in error probabilities of about 5.9% for both error types, which is well below the targeted 10%. The expected run-length calculates to 1.47 samples.

The cost function $\rho_\lambda(z_1)$ of the optimal sequential test for the Bernoulli distributions in (3.32) is depicted in Figure 3.9. The optimal thresholds are again obtained form the intersection points of $d_\lambda(z_1)$ and $g_\lambda(z_1)$ and are given by $A^* = 1/B^* = 4$, which implies that the optimal testing policy is indeed randomized. The derivative of $\rho_\lambda(z_1)$, which is a picewise constant function, is shown in Figure 3.10. The influence of randomization

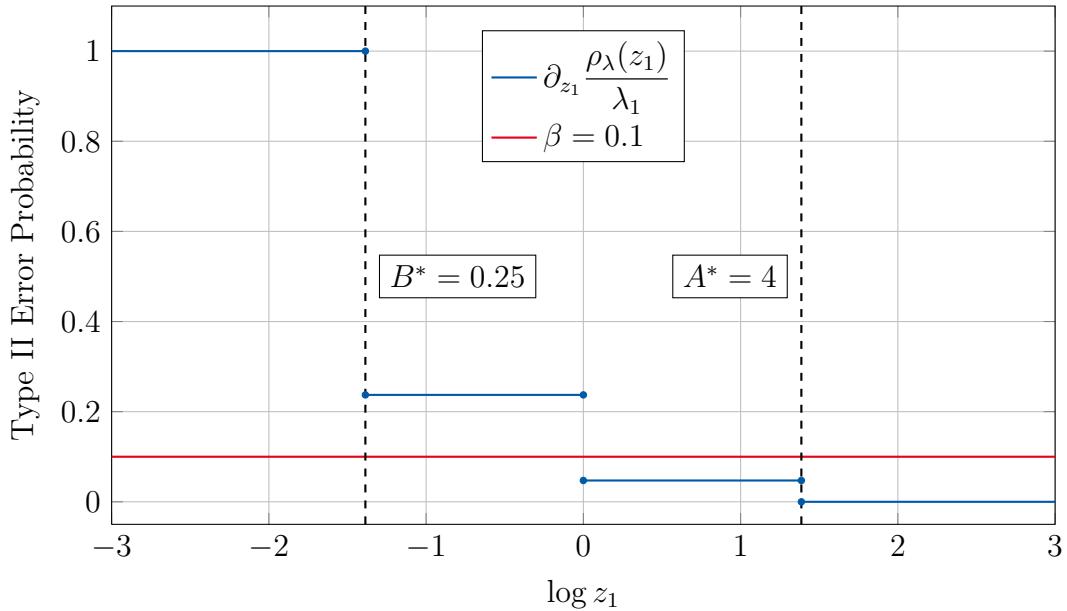


Figure 3.10: Bernoulli Distributions. Scaled generalized derivative of the cost function $\rho_\lambda(z_1)$ depicted in Figure 3.9.

on the error probabilities in the states $z_1 = A^*$ and $z_1 = B^*$ could already be seen in the previous example. Here, however, the randomization also effects the error probabilities in the initial state $z_1 = 1$. Without the need for additional analysis, it follows from Figure 3.10 and Theorem 8 that randomized stopping policies with thresholds $A^* = 1/B^* = 4$ can realize type II error probabilities between 0.047 and 0.237. The lower bound, which is even smaller than the type II error probability realized by the deterministic thresholds $A = 1/B = 16$, can be achieved by never stopping at the lower threshold, but always stopping at the upper threshold. The upper bound can accordingly be achieved by never stopping at the upper and always stopping at the lower threshold.

The optimal randomized stopping rule can be obtained by solving the integral equations (3.22). Because of the discrete state space, the latter reduce to a system of linear equations that can be solved analytically and yields

$$\psi^*(A^*) = \psi^*(B^*) = \frac{7}{32} \approx 0.2187.$$

The expected run-length calculates to

$$\gamma_{\pi^*, P_0}(1) = \frac{4}{3} \approx 1.3333,$$

which is a reduction by roughly 10%, compared to the best deterministic stopping rule. For smaller error probabilities this improvement becomes smaller as well.

In practice, the difference in expected run-length between randomized and non-randomized tests is often negligible. However, this example illustrates how the cost function of the optimal test, which is independent of the randomization, can be used to bound the error probabilities of both types of test. This method works reliably, even in more complex scenarios, and can help the test designer to decide whether randomization needs to be considered or not without going into the specifics of the given distributions.

3.5.3 Observable Markov Chain

The example in Section 3.5.1 can be complicated by assuming that the observed random sequence is governed by an observable Markov chain with state space $\Omega_\theta = \{1, 2\}$. More precisely,

$$\begin{aligned}\mathcal{H}_0: \quad X_n &:= (Y_n, \Theta_n) \sim p_{0,\theta}(\theta_n) \cdot \mathcal{N}(0, \sigma) \\ \mathcal{H}_1: \quad X_n &:= (Y_n, \Theta_n) \sim p_{1,\theta}(\theta_n | \theta_{n-1}) \cdot \mathcal{N}(\theta_n/2, \sigma),\end{aligned}$$

where $p_{0,\theta}$ and $p_{1,\theta}$ define the (transition) probabilities of the states. In this model, Y_n and Θ_n are independent Gaussian and Bernoulli random variables under the null hypothesis, while under the alternative hypothesis the distribution of Y_n depends on the current state θ_n . The initial state is assumed to be $\theta_0 = 1$. By inspection, θ_{n-1} is a sufficient statistic for the distribution of X_n , conditioned on the previous observations.

For the numerical results, $\sigma = 1$ is assumed and the transition probabilities under \mathcal{H}_1 are chosen symmetrically as $p(\theta' | \theta) = 0.8$ for $\theta' = \theta$ and $p(\theta' | \theta) = 0.2$ for $\theta' \neq \theta$. Under \mathcal{H}_0 , $p_0(1) = p_0(2) = 0.5$ is used.

The results of the optimal test are given in Table 3.3 and Figure 3.11. The expected run-length and error probabilities of a test using Wald's approximations are shown in Table 3.4. The results do not differ much from the i.i.d. scenario in terms of the efficiency of Wald's test. The reduction in expected samples by using the optimal strategy is still between 25% and 10%. However, to achieve this reduction, the likelihood ratio alone is no longer a sufficient test statistic since different thresholds have to be used in different states—see Figure 3.11. In line with the asymptotic optimality of Wald's approximations, the difference between the thresholds in the two states reduces with decreasing error probabilities.

Target		Thresholds		Error Prob.		Avg. and Exp. Run-Length		
α	β	$A^*(1)$	$B^*(1)$	$\tilde{\alpha}_{\pi^*}(1, 1)$	$\tilde{\beta}_{\pi^*}(1, 1)$	$\tilde{\gamma}_{\pi^*, P_0}(1, 1)$	$\tilde{\gamma}_{\pi^*, P_1}(1, 1)$	$\gamma_{\pi^*, P_0}(1, 1)$
		$A^*(2)$	$B^*(2)$					
0.1		1.76	-1.47	0.0996	0.1005	3.54	4.17	3.54
		1.63	-1.64					
0.05		2.48	-2.25	0.0496	0.0494	5.19	6.05	5.22
		2.35	-2.42					
0.01		4.13	-3.90	0.0097	0.0099	8.69	9.85	8.7
		4.00	-4.07					
0.1	0.01	1.85	-3.80	0.0988	0.0096	7.45	5.13	7.42
		1.71	-3.97					

Table 3.3: Observable Markov Chain. Target error probabilities α, β , optimal log-likelihood ratio thresholds A^*, B^* as functions of θ , empirical error probabilities $\tilde{\alpha}_{\pi^*}(1, 1), \tilde{\beta}_{\pi^*}(1, 1)$, average run-length $\tilde{\gamma}_{\pi^*}(1, 1)$ under both hypothesis and expected run-length $\gamma_{\pi^*}(1, 1)$ under \mathcal{H}_0 .

Target		Thresholds		Error Prob.		Avg. Run-Length		efficiency
α	β	A_W	B_W	$\tilde{\alpha}_{\pi_W}(1, 1)$	$\tilde{\beta}_{\pi_W}(1, 1)$	$\tilde{\gamma}_{\pi_W, P_0}(1, 1)$	$\tilde{\gamma}_{\pi_W, P_1}(1, 1)$	
0.1		± 2.20		0.0640	0.0580	4.84	5.51	0.73
0.05		± 2.94		0.0300	0.0269	6.94	7.35	0.80
0.01		± 4.60		0.0056	0.0055	9.90	11.00	0.88
0.1	0.01	2.29	-4.50	0.0596	0.0051	9.00	6.25	0.83

Table 3.4: Observable Markov Chain. Target error probabilities α, β , log-likelihood ratio thresholds A_W, B_W obtained by Wald's approximation, empirical error probabilities $\tilde{\alpha}_{\pi_W}(1, 1), \tilde{\beta}_{\pi_W}(1, 1)$ and average run-length $\tilde{\gamma}_{\pi_W}(1, 1)$ under both hypothesis. The last column gives the relative loss in the average run-length compared to the optimal test, i.e., $\tilde{\gamma}_{\pi^*}(1, 1)/\tilde{\gamma}_{\pi_W}(1, 1)$.

3.5.4 Gaussian AR(1) Process

The final example is the Gaussian AR(1) process. In [Nov09] it is shown that the optimal stopping strategy for this process is a function of the likelihood ratio and the current observation. The exact optimal testing policy was first presented and implemented in [FZ15a].

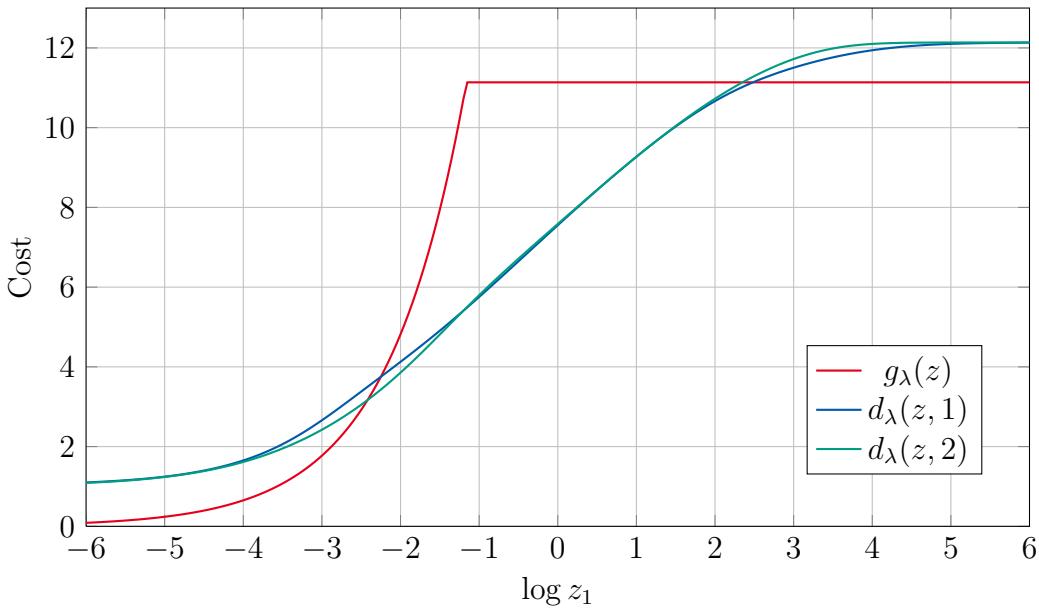


Figure 3.11: Observable Markov Chain: State dependent cost functions for an optimal test with error probabilities $\alpha = \beta = 0.05$.

The two hypotheses are given by

$$\begin{aligned}\mathcal{H}_0: \quad & X_n = a_0 X_{n-1} + \epsilon_n, \\ \mathcal{H}_1: \quad & X_n = a_1 X_{n-1} + \epsilon_n,\end{aligned}$$

where $(\epsilon_n)_{n \geq 1}$ is a sequence of i.i.d. zero mean Gaussian random variables with standard deviation σ . Since knowledge of x_{n-1} is sufficient to describe the conditional distribution of X_n , it holds that $\theta_{n-1} = x_{n-1}$. For the experiment, $\sigma = 1$, $a_0 = 0$ and $a_1 = 1$ are chosen, which corresponds to a test for an AR(1) process against Gaussian noise. It is further assumed that $P = P_0$ and that $x_0 = \theta_0 = 0$.

The average run-length and the error probabilities of the optimal test and the one using Wald's approximations are given in Tables 3.5 and 3.6.

A segment of the cost functions for $\alpha = \beta = 0.05$ is depicted in Figure 3.12. The intersection of the two surfaces corresponds to the thresholds of the test. In Figure 3.13, the latter is shown together with the approximated constant ones. Interestingly, the optimal thresholds are not uniformly tighter than the approximations. Instead, the additional degree of freedom is used to loosen the thresholds for observations that are very unlikely under P_0 and tighten them in the region around the origin. Evidently, this strategy is more efficient than uniformly tightening the thresholds. Another noteworthy fact is that in contrast to the lower threshold, the upper threshold is far from being constant. This does not contradict the asymptotic optimality of the constant

Target		Error Prob.		Avg. and Exp. Run-Length		
α	β	$\tilde{\alpha}_{\pi^*}(1, 0)$	$\tilde{\beta}_{\pi^*}(1, 0)$	$\tilde{\gamma}_{\pi^*, P_0}(1, 0)$	$\tilde{\gamma}_{\pi^*, P_1}(1, 0)$	$\gamma_{\pi^*, P_0}(1, 0)$
0.1		0.0980	0.1006	5.69	5.86	5.64
0.05		0.0509	0.0477	7.46	6.98	7.48
0.01		0.0095	0.0103	11.25	9.06	11.24
0.1	0.01	0.0986	0.0098	9.95	6.8278	9.93

Table 3.5: AR(1) process. Target error probabilities α, β , empirical error probabilities $\tilde{\alpha}_{\pi^*}(1, 0), \tilde{\beta}_{\pi^*}(1, 0)$, average run-length $\tilde{\gamma}_{\pi^*}(1, 0)$ under both hypothesis and expected run-length $\gamma_{\pi^*}(1, 0)$ under \mathcal{H}_0 .

Target		Thresholds		Error Prob.		Avg. Run-Length		efficiency
α	β	A_W	B_W	$\tilde{\alpha}_{\pi_W}(1, 0)$	$\tilde{\beta}_{\pi_W}(1, 0)$	$\tilde{\gamma}_{\pi_W, P_0}(1, 0)$	$\tilde{\gamma}_{\pi_W, P_1}(1, 0)$	
0.1		± 2.20		0.0410	0.0535	7.73	6.54	0.74
0.05		± 2.94		0.0171	0.0253	9.45	7.51	0.79
0.01		± 4.60		0.0027	0.0049	12.98	8.97	0.87
0.1	0.01	2.29	-4.50	0.0366	0.0050	12.33	7.05	0.81

Table 3.6: AR(1) process. Target error probabilities α, β , log-likelihood ratio thresholds A_W, B_W obtained by Wald's approximation, empirical error probabilities $\tilde{\alpha}_{\pi_W}(1, 0), \tilde{\beta}_{\pi_W}(1, 0)$ and average run-length $\tilde{\gamma}_{\pi_W}(1, 0)$ under both hypothesis. The last column gives the relative loss in the average run-length compared to the optimal test, i.e., $\tilde{\gamma}_{\pi^*}(1, 0)/\tilde{\gamma}_{\pi_W}(1, 0)$.

threshold test. It does, however, indicate that there is no longer a stopping strategy that concurrently minimizes the expected run-length under both hypotheses, as is the case for i.i.d. observations [WW48, Sie85]. Minimizing the run-length under \mathcal{H}_1 yields a mirrored version of the thresholds in Figure 3.13, with the lower threshold following the parabolic shape and vice versa.

A nice property of the optimal thresholds shown here is that they are relatively easy to approximate by polynomials or rational functions. In practice, a few coefficients can be sufficient to implement a nearly optimal strategy that combines the ease of the constant threshold test with the efficiency of the optimal one.

In addition to the results given in Table 3.5, an optimal test for the AR(1) model was designed with $(\alpha, \beta) = (0.0410, 0.0535)$, which are the (approximate) error probabili-

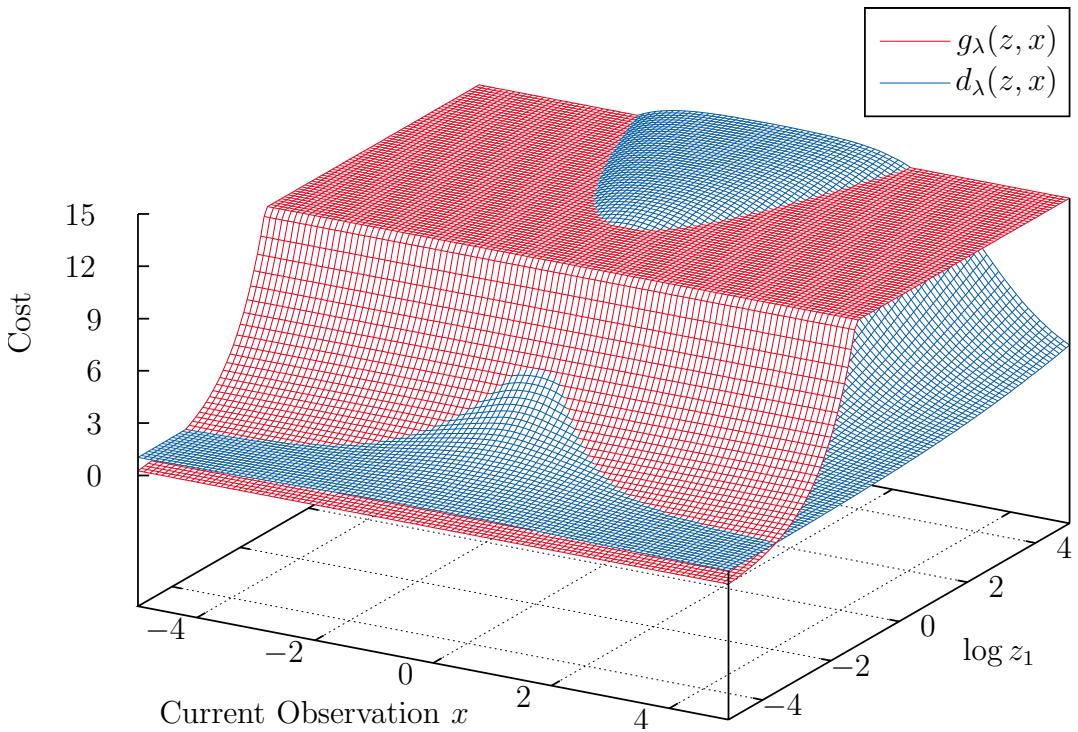


Figure 3.12: AR(1) process. Segment of the cost functions for an optimal test with error probabilities $\alpha = \beta = 0.05$ and $P = P_0$.

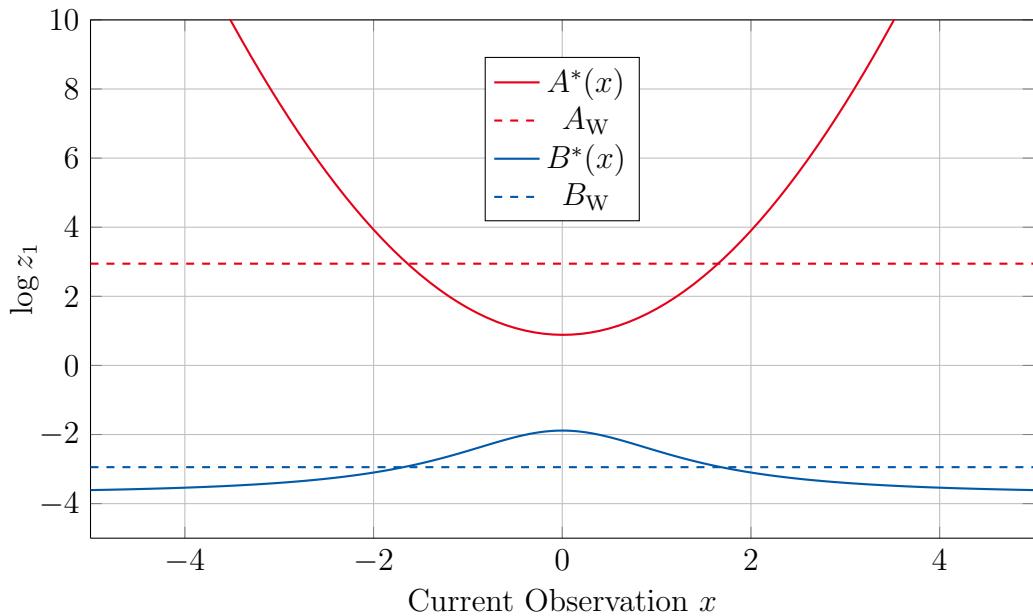


Figure 3.13: AR(1) process: Optimal and approximated log-likelihood ratio thresholds as functions of the current observation x for target error probabilities $\alpha = \beta = 0.05$.

ties of a test using Wald's approximations with target error probabilities $\alpha = \beta = 0.1$. The idea is to compare the strictly optimal test to the optimal constant threshold test.

The expected run-length of the optimal test is 7.45, compared to 7.73 for the test with constant thresholds. This corresponds to a reduction of about 3.6%. Whether this improvement is worth the increased complexity surely depends on the actual application. However, calculating the optimal constant thresholds is a non-trivial problem in itself so that the effort might as well be invested in solving the problem exactly.

3.6 Summary

In this chapter, the optimal sequential test for stochastic processes with Markovian representations was derived. This was done by first formulating the design problem as an optimal stopping problem, based on which a cost minimizing testing policy was obtained as the solution of a nonlinear integral equation. It was then shown that the partial generalized derivatives of the optimal cost function are, up to a scaling factor, identical to the error probabilities of the cost minimizing test. This relation was used to formulate the problem of designing optimal sequential tests as a problem of solving an integral equation under constraints on the partial derivatives of the solution function. Moreover, it was shown that the latter problem can be solved by means of standard linear programming techniques without the need to calculate the partial derivatives explicitly. Numerical examples were given to illustrate this procedure.

Chapter 4

Fundamentals II

Before going into the details of the design of minimax optimal sequential hypothesis tests, some additional concepts and preliminary results need to be introduced. This is the purpose of the current chapter. First, the idea of robust statistics and the minimax principle are introduced and the characterization of minimax optimal solutions via saddle points is discussed. This relation will turn out to be crucial for the proof of minimax optimality of the proposed tests. Subsequently, uncertainty sets are introduced as a means of formulating hypothesis tests under incomplete knowledge of the underlying distributions. For sequential hypothesis test, both the uncertainty sets and the true distributions additionally need to be allowed to depend on the test statistic in order to obtain strictly minimax optimal solutions. This connection between the test and the underlying random process is explained in more detail in Section 4.4. Finally, some technical aspects of convex optimization in Banach spaces are revised, in particular the method Lagrangian multipliers in infinite dimensional spaces.

4.1 Statistical Robustness and the Minimax Principle

In the previous chapter, it was assumed that the distributions under both hypotheses are known exactly. This, however, is rarely the case in practice. Even if an accurate model for P_0 and P_1 exists, a certain degree of mismatch between model and reality is usually unavoidable. Put the other way around, specifying the hypotheses exactly requires the test designer to have access to a complete probabilistic description of all possible sources of uncertainty, which is a highly unrealistic assumption. Consider, for example, a simple energy detector that is used to establish the presence or absence of a radio signal. Even if the signal of interest is deterministic and known, the performance of the detector depends on factors like the noise characteristic of the sensors, the propagation path between the transmitter and the detector, possible interference from other transmitters and even the weather conditions [MLN09]. A *robust* statistical hypothesis tests is designed to be insensitive to such random deviations from the underlying model. A more formal definition of statistical robustness and robust hypothesis tests is given in this and the upcoming sections.

Taking model mismatches into account when designing statistical tests results in *distributional uncertainties*, meaning that under either hypothesis the distribution of the observed random variables is only known approximately. Each hypothesis is hence represented by a set or class of possible distributions. Hypotheses of this kind are called *composite hypotheses*, in contrast to the *simple* hypotheses considered in Chapter 3. For a binary test, composite hypothesis are in general of the form

$$\begin{aligned}\mathcal{H}_0: P_X &\in \mathcal{P}_0, \\ \mathcal{H}_1: P_X &\in \mathcal{P}_1,\end{aligned}\tag{4.1}$$

where $\mathcal{P}_0, \mathcal{P}_1 \subset \mathcal{M}_\mu(\Omega_X, \mathcal{F}_X)$ are referred to as the uncertainty sets. Here $\mathcal{M}_\mu(\Omega, \mathcal{F})$ denotes the set of all distributions on a measurable space (Ω, \mathcal{F}) that admit a positive density with respect to the measure μ , i.e.,

$$\mathcal{M}_\mu(\Omega, \mathcal{F}) := \left\{ P : \exists p > 0 : \int_{\mathcal{E}} p(\omega) d\mu(\omega) = P(\mathcal{E}) \quad \forall \mathcal{E} \in \mathcal{F} \right\}.\tag{4.2}$$

The restriction that $p > 0$ is introduced to guarantee that all likelihood ratios are well defined. Moreover, in order to exclude cases where \mathcal{H}_0 and \mathcal{H}_1 are statistically indistinguishable, the uncertainty sets are further assumed to be disjoint, i.e., $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. A more rigorous definition of \mathcal{P}_0 and \mathcal{P}_1 is given in Section 4.3.

Tests for composite hypothesis, with fixed and random sample sizes, have been studied extensively in the literature. An in-depth treatment can be found, for example, in [Pap91] or [LR05]. Here, the discussion is limited to a brief introduction to different approaches in composite hypothesis testing with emphasis on the minimax approach.

In the previous chapter it was shown how to design an optimal sequential test for a pair of distributions (P_0, P_1) . The fundamental problem in composite hypothesis testing is that typically no test exists that is optimal for *every* possible pair of distributions $(P_0, P_1) \in \mathcal{P}_0 \times \mathcal{P}_1$. Hence, an additional criterion is necessary in order to define a meaningful objective for tests between composite hypotheses. The existing approaches can roughly be grouped into three categories: *Bayesian* methods, *adaptive* methods and *minimax* methods.

The idea of Bayesian methods is to use a weighted average of the test performance as a global objective function. More precisely, it is assumed that under each hypothesis the true distribution is generated according to some probabilistic law, or *prior* distribution. The hypothesis test is then designed so as to minimize the expected error probabilities, where the expectation is taken with respect to the prior probabilities of the individual distributions. As a consequence, Bayesian methods perform very well if the unknown distributions indeed occur with the assumed prior probabilities, but can perform poorly,

if this is not the case. Particularly critical are scenarios that are highly unlikely to occur under the assumed prior and, hence, have a negligible influence on the test design. Such corner cases can deteriorate the performance of Bayesian tests significantly. In summary, Bayesian tests are designed to yield good performance *on average*, but not for every feasible pair of distributions.

The problem of blind spots in the test performance can be avoided by using adaptive methods. In contrast to Bayesian methods, they rely less on *a priori* assumptions, but rather try to infer as much information as possible from the data itself. A typical adaptive test procedure first estimates the most likely distribution under each hypothesis and then performs an optimal test for the estimated pair [ZZM92]. In sequential testing, the estimates of the distributions are usually updated on the fly and the stopping criterion is adjusted to the increasing accuracy of the estimates [LLY14]. In theory, adaptive tests yield close to optimal performance under all possible distribution pairs (P_0, P_1) . However, this performance is only achieved if the estimated distributions are sufficiently close to the truth. This cannot be guaranteed if, for example, the sample size is small, the parameters that need to be estimated are high-dimensional, or the parameters fluctuate at a rate that is close to the sampling frequency. Moreover, performance guarantees for adaptive sequential tests have only been established in an asymptotic sense, i.e., for vanishingly small error probabilities $\alpha, \beta \rightarrow 0$. Obtaining strict performance bounds for non-vanishing error probabilities is still an open problem [Tar13].

The idea to guarantee a certain performance under all feasible circumstances leads to the minimax design principle. Its objective is to minimize the maximum (mini-max) error probabilities of a test over all pairs $(P_0, P_1) \in \mathcal{P}_0 \times \mathcal{P}_1$. This results in a test that performs optimally in the worst case and has guaranteed performance bounds in all other cases. Similar to the adaptive test, little *a priori* knowledge is incorporated into the test design. The advantage of the minimax approach is that it yields tests that are predictable and robust, in the sense that they do not suffer performance degradation over the entire uncertainty set $\mathcal{P}_0 \times \mathcal{P}_1$. The disadvantage is that the minimax approach results in highly conservative tests that are optimized for a worst case scenario that may never actually happen, while possibly performing mediocre under typical scenarios. This problem motivates the use of Bayesian methods...

Many more design principles have been suggested and various hybrids of different approaches, such as robust Bayesian [dOPdG10] or robust adaptive methods [ASZS14], can be found in the literature. In summary, there is no one method or technique that fits all needs. Nevertheless, the notion of robustness is closely tied to the minimax approach and its property to guarantee a certain performance.

The field of robust statistics, and robust hypothesis testing in particular, was developed foremost by Huber in the mid-1960s [Hub64]. He was the first to derive the famous clipped likelihood ratio test [Hub65], which is minimax optimal under ε -contamination type, i.e., infrequent, but grossly corrupted outliers. The corresponding uncertainty set \mathcal{P}_ε is given by

$$\mathcal{P}_\varepsilon = \{Q \in \mathcal{M}_\mu : Q = (1 - \varepsilon)P + \varepsilon H, H \in \mathcal{M}_\mu\},$$

where P is referred to as the nominal distribution and H is an arbitrary outlier distribution. This kind of contamination is particularly critical since a single corrupted observation can be enough to alter the outcome of a non-robust test [Hub81]. The result that a simple clipping of the test statistic yields a minimax optimal test is one of the most significant contributions to the field of robust hypothesis testing and probably the one with the highest impact on practical applications. Huber further showed that the clipped likelihood ratio test is in fact a regular likelihood ratio test, but for the so-called *least favorable* instead of the nominal, uncontaminated distributions. The idea to reduce the design of a minimax test for composite hypotheses to the design of an optimal test for carefully chosen simple hypotheses is underlying most robust testing schemes and is used in this dissertation as well. It is based on the characterization of minimax optimal solutions via saddle points. This aspect is briefly discussed in the next section.

4.2 Saddle Points and Minimax Optimality

In general, the question whether or not a minimax optimal procedure exists, depends on the existence of a *saddle point*. Consider a function $J: \Omega_U \times \Omega_V \rightarrow \mathbb{R}$, where Ω_U and Ω_V are subsets of linear vector spaces. A point $(u^*, v^*) \in \Omega_U \times \Omega_V$ is called a saddle point if

$$J(u^*, v) \leq J(u^*, v^*) \leq J(u, v^*) \tag{4.3}$$

for all $(u, v) \in \Omega_U \times \Omega_V$. If a saddle point (u^*, v^*) exists, it holds that

$$\min_{u \in \Omega_U} \max_{v \in \Omega_V} J(u, v) = \max_{u \in \Omega_U} \min_{v \in \Omega_V} J(u, v) = J(u^*, v^*). \tag{4.4}$$

A proof for the saddle point property can be found in [Lev08]. In essence, it states that (u^*, v^*) being a saddle point is a sufficient condition for it to be optimal in the sense of (4.4). This property is used in the derivation of the minimax optimal sequential test in Chapter 5. Before delving deeper into the formulation of the minimax sequential testing problem, however, some additional groundwork needs to be laid.

4.3 Uncertainty Sets

Some assumptions need to be made about the nature of the uncertainty sets such that they match the minimax design approach. As discussed in Section 4.1, the idea underlying robust statistics is that the distributional uncertainties arise from some kind of unpredictable unsystematic deviation from a given model. Here it is assumed that this deviation is

- random,
- possibly time varying,
- nonparametric
- and small.

The assumption that the uncertainty sets are nonparametric is rather standard in robust statistics and is based on the consideration that unknown sources of uncertainty cannot realistically be assumed to follow a particular parametric model. In the same spirit, the deviations cannot be assumed to be time-invariant. It is important to note, though, that the time varying property refers to the deviation itself and not the uncertainty sets, which are assumed to be time-independent. This aspect is covered in more detail later. Finally, the assumption that the deviations are small is to be understood in the sense that applying the minimax approach to the problem is sensible in the first place. The larger the uncertainty sets, the more difficult it is to guarantee a worse case performance without disproportionately sacrificing performance under nominal conditions. The exact definition of “small” depends on the uncertainty model. For the ε -contamination model, for example, ε values larger than 0.5 are infeasible [Hub65] and the number of required samples increases drastically for values close to this bound.

In order to have a well defined problem formulation for the minimax sequential test it is essential to carefully specify which distribution is subject to uncertainty. In the most general case, this is the joint distribution of the random process X . However, defining meaningful uncertainty sets for the distribution of a stochastic process is an intricate task and usually not necessary. An approach that is more useful in practice is to assume that for every $n \in \mathbb{N}$ the *marginal* or *conditional* distribution of X_n is subject to uncertainty. The collection of uncertainty sets for every X_n then induces an uncertainty set of feasible distributions P_X .

In this dissertation it is assumed that the conditional distributions P_θ , as defined in (2.10), are subject to uncertainty. More precisely, for each $\theta \in \Omega_\theta$ the conditional distribution P_θ is replaced by a set of feasible distributions \mathcal{P}_θ . Every set \mathcal{P}_θ is further assumed to be a convex subset of \mathcal{M}_μ . This uncertainty model corresponds to a situation where exact knowledge of the history of the process is available in form of θ_n , but the distribution of X_{n+1} conditioned on θ_n is subject to uncertainty. An example for this scenario is an autoregressive process whose innovations follow a distribution that is not known precisely. Alternatively, the uncertainty in the conditional distributions might stem from an uncertainty about the true model parameters. Types of uncertainty that are *not* covered by this model are, for example, mismatches between the samples that are *observed* and the samples that are actually *generated* and determine the true value of the sufficient statistic θ .

Assuming that the conditional distributions P_θ are subject to uncertainty, the hypotheses of a binary test are given by

$$\begin{aligned}\mathcal{H}_0: P_X &\in \mathcal{P}_0, \\ \mathcal{H}_1: P_X &\in \mathcal{P}_1,\end{aligned}$$

where \mathcal{P}_0 and \mathcal{P}_1 are defined in terms of two uncertainty sets \mathcal{P}_θ^0 and \mathcal{P}_θ^1 via

$$\mathcal{P}_i = \{ P_X : P_{X_n|\theta} \in \mathcal{P}_\theta^i \quad \forall n \in \mathbb{N} \}, \quad i = 0, 1. \quad (4.5)$$

In what follows, it is further assumed that for every $\theta \in \Omega_\theta$ the sets \mathcal{P}_θ^0 and \mathcal{P}_θ^1 can equivalently be defined in terms of the density function corresponding to $P_{X_n|\theta}$. More precisely, it is assumed that the uncertainty set can be defined via convex constraints on $p_{X_n|\theta}$. The notation $p_{X_n|\theta} \in \mathcal{P}_\theta$ is occasionally used to make this assumption explicit. It is discussed in more detail at the end of Section 4.5.

The assumption that the true distributions can be time-varying is in contrast to many works on robust sequential tests where it is assumed that the distribution of every X_n is unknown, but that all X_n are *identically* distributed—compare the references in Section 1.2. This assumption, however, is somewhat in conflict with the principles of minimax robustness, which targets small, but random deviations from a given model. In the i.i.d. case, in contrast, the deviations admit a certain structure and the true distribution can in fact be estimated from the data. This raises the question whether a minimax approach should be used in the first place. In general, an i.i.d. assumption should not be introduced in the design of minimax procedures unless it is well justified by external reasons.¹

¹As shown in [Hub65, HS73], the least favorable distributions for fixed sample size tests are indeed identical for every X_n . However, this property is a result of the minimax optimization, not an *a priori*

4.4 Distributions and Uncertainty Sets Depending on the Test Statistic

In the design of minimax robust sequential tests, distributions arise that are parameterized in the test statistic, meaning that the distribution of X_n depends on the value of the test statistic T^{n-1} . The notation

$$P_{X_{n+1}|x_1, \dots, x_n} = P_{X_{n+1}|t_n} \quad (4.6)$$

is used to indicate this dependence. For test statistics of the form (3.15), which will be shown to be optimal in the minimax case as well, this implies that P_{X_n} not only depends on θ , but also on the likelihood ratios z_0 and z_1 , i.e.,

$$P_{X_{n+1}|x_1, \dots, x_n} = P_{X_{n+1}|z^n, \theta_n}. \quad (4.7)$$

Irrespective of the additional complexity, optimal sequential tests for distributions depending on the test statistic can still be analyzed and designed within the framework presented in the previous chapter, namely, by choosing the sufficient statistic for the distribution of X_{n+1} as

$$\tilde{\theta}_n := (z^n, \theta_n). \quad (4.8)$$

Consequently, the results obtained in the previous sections can be extended to distributions depending on the test statistic by simply substituting $\tilde{\theta}$ for θ in the respective equations. Interestingly, for sufficient statistics of the form (4.8), no additional test statistic is required, or, more precisely, the test statistic and the Markov statistic coincide so that $T^n = \tilde{\Theta}_n$. Nevertheless, the explicit distinction between z and θ is maintained in the upcoming sections, meaning that the conditional distributions are written as $P_{z, \theta}$.

A result that is of particular importance for the derivation of minimax robust sequential tests is the function ρ_λ in Theorem 5. Allowing P_0 , P_1 and P to depend on z in addition to θ , (3.12) becomes

$$\rho_\lambda(z, \theta) = \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda \left(z_0 \frac{p_{z, \theta}^0(x)}{p_{z, \theta}(x)}, z_1 \frac{p_{z, \theta}^1(x)}{p_\theta(x)}, \xi_{z, \theta}(x) \right) dP_{z, \theta}(x) \right\}. \quad (4.9)$$

This definition of ρ_λ is used repeatedly in the upcoming sections.

assumption. Given the permutation invariance of the fixed sample size test (the outcome of the test does not depend on the ordering of the samples), the i.i.d. property is intuitively reasonable. For the sequential test, however, this symmetry argument no longer holds since rearranging the samples can indeed alter the outcome of the test—compare ordering the observations by significance for the alternative hypotheses to ordering the samples by significance for the null hypothesis.

Another central result that carries over to distributions depending on the test statistic are the integral equations (2.41)–(2.43) in Section 2.4.3. The extension follows immediately from the fact that $T^n = (Z^n, \Theta_n)$ still satisfies property (2.24), i.e., it provides all information necessary to determine the conditional distribution of T^{n+1} .

Given that θ is a component of the test statistic, the uncertainty sets \mathcal{P}_θ^0 , \mathcal{P}_θ^1 and \mathcal{P}_θ are functions of the test statistic as well. In order to present results for tests based on test statistics different form (z, θ) , the generic notations \mathcal{P}_t^0 , \mathcal{P}_t^1 and \mathcal{P}_t are used to indicate this dependence. In general, this notation should be read in the sense that the test statistic provides sufficient information to determine the uncertainty set for the distribution of the random variable generating the next sample, i.e.,

$$\mathcal{P}_{X_{n+1}|x_1, \dots, x_{n-1}} = \mathcal{P}_{X_{n+1}|\theta_n} = \mathcal{P}_{X_n|t_n}.$$

Although incorporating dependencies of the form (4.6) in the optimal sequential testing framework is technically straightforward, the question why the behavior of a stochastic process should depend on the test statistic deserves a more detailed answer. Put another way: How can the test design have an influence on the process that is being tested? Even though examples exist for which the testing procedure does indeed affect the stochastic process,² in typical signal processing applications the distribution of X is entirely independent of the test design. Nevertheless, state dependent distributions are a crucial concept for the design of minimax robust tests, irrespective of whether or not they arise in practice. The reason for this is that the least favorable distribution of the random variable generating the next observation depends on the current state of the test statistic. Assume, for example, that the test statistic is close to the stopping region corresponding to a decision for \mathcal{H}_0 . In this state, another observation in favor of \mathcal{H}_0 most likely causes the test to stop. Hence, a distribution that is least favorable *with respect to the expected run-length*, needs to place as much probability mass as possible on observations in favor of \mathcal{H}_1 . The opposite is true if the test statistic is close to the stopping region corresponding to a decision for \mathcal{H}_1 . In this case, the distribution that maximizes the expected run-length is concentrated on observations in favor of \mathcal{H}_0 . This idea is illustrated in Figure 4.1 using the example of the sequential probability ratio test with constant thresholds. The red arrows indicate the direction of the drift that is least favorable in the current state.

²A typical example is a test for the efficacy of a new drug or treatment. If such a test is based on features that the participant has access to (blood pressure, eyesight, allergic reactions, etc.), this knowledge can influence the participant's response to the treatment. It is well known, for example, that good or bad results in the beginning can change a participant's confidence in the treatment which in turn affects the expected progress [Kne14]. More details on the interesting topic of sequential medical trials can be found in [Cho11] and [BLS13].

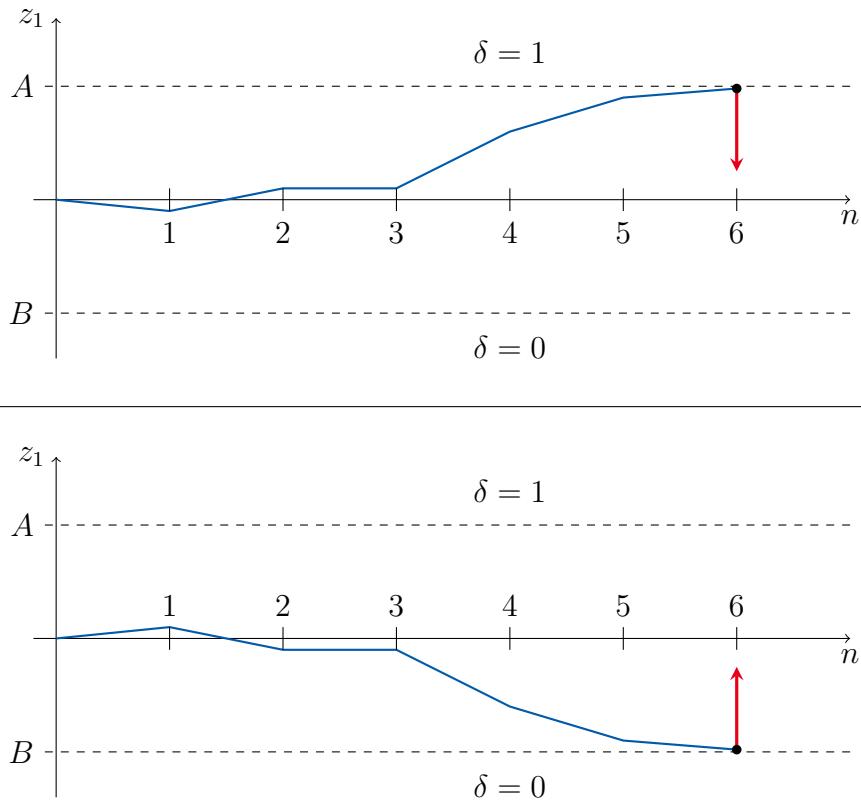


Figure 4.1: Illustration of state dependent least favorable distributions for the sequential probability ratio test. For z_1 close to the upper threshold, observations in favor of \mathcal{H}_0 are least favorable with respect to the expected run-length. For z_1 close to the lower threshold, observations in favor of \mathcal{H}_1 are least favorable.

The same considerations hold true for the error probabilities. Since the set of states that can be reached with the next update of the test statistic depends on its current state, the least favorable distributions with respect to the error probabilities also depend on the current state. In practice, however, the effect is more pronounced for the expected run-length than for the error probabilities. A more detailed discussion of this phenomenon is deferred to Section 5.8.

In light of these considerations, the property that the test statistic coincides with the Markov statistic is a characteristic of minimax procedures. In essence, $T^n = \tilde{\Theta}_n$ implies that the test designer and his virtual adversary, who “designs” the stochastic process X , have access to the same information: The sequential test is designed based on knowledge of the Markov statistic, which determines the expected behavior of the stochastic process. The stochastic process, in turn, is designed based on knowledge of the test statistic, which determines the expected performance of the test. This constellation is a necessary consequence of the minimax principle: Distributions that do not adapt to the test statistic cannot be guaranteed to be least favorable and testing pro-

cedures that are based on insufficient information about the stochastic process cannot be guaranteed to be optimal.

In the next Chapter, least favorable distributions are introduced in a more formal and precise manner. In order to do so, however, some basics of convex optimization in Banach spaces need to be introduced.

4.5 Aspects of Convex Optimization in Banach Spaces

In this section, some fundamental aspects of infinite dimensional optimization in Banach spaces are reviewed. In particular, the space $\mathcal{L}_1 = \mathcal{L}_1(\Omega, \mu)$ of all μ integrable functions on Ω is important for the upcoming sections. The focus here is on Lagrangian duality, the Fréchet-derivative and first order optimality conditions, which are used in the next Chapter to characterize the densities of the least favorable distributions. All of the presented results can be found in standard textbooks, such as [Ulb09, Bot14].

Several symbols are reused in this section that have already been defined in previous sections. However, since the results in this section are self-sufficient and do not rely on previously defined quantities, the confusion should be minimal.

4.5.1 The Dual Space

The dual space of a Banach space \mathcal{L} is defined as the space of all linear functionals on \mathcal{L} , i.e., all linear operators that map from \mathcal{L} to \mathbb{R} . For σ -finite measures μ , the dual space of \mathcal{L}_1 is given by

$$\mathcal{L}_1^* = \left\{ U: \mathcal{L}_1 \rightarrow \mathbb{R} : U(y) = \int u(\omega)y(\omega) d\mu(\omega), u \in \mathcal{L}_\infty \right\}, \quad (4.10)$$

where $\mathcal{L}_\infty = \mathcal{L}_\infty(\Omega, \mu)$ denotes the space of all essentially bounded functions on Ω , i.e.,

$$\mathcal{L}_\infty = \left\{ u: \Omega \rightarrow \mathbb{R} : \sup_{\substack{\mathcal{E} \in \mathcal{F} \\ \mu(\mathcal{E}) > 0}} \sup_{\omega \in \mathcal{E}} u(\omega) < \infty \right\}. \quad (4.11)$$

Each functional $U \in \mathcal{L}_1^*$ can, hence, be identified with an element $u \in \mathcal{L}_\infty$ so that \mathcal{L}_∞ is usually referred to as the dual space of \mathcal{L}_1 .

4.5.2 Fréchet Differentials and Subdifferentials

Let $J: \mathcal{L}_1 \rightarrow \mathbb{R}$ be a function. Given some $y \in \mathcal{L}_1$, the function J admits a directional derivative at y in the direction $d \in \mathcal{L}_1$ if the limit

$$J'(y; d) := \lim_{t \rightarrow 0} \frac{J(y + td) - J(y)}{t}, \quad t \in \mathbb{R}$$

exists. If the directional derivative exists for all $d \in \mathcal{L}_1$, J is called Gâteaux differentiable at y . If, in addition, $J'(y; \cdot) : \mathcal{L}_1 \rightarrow \mathbb{R}$ is linear, it is by definition an element of the dual space \mathcal{L}_1^* and there exist a function $J'_y \in \mathcal{L}_\infty$ such that

$$J'(y; d) = \int J'_y(\omega) d(\omega) \, d\mu(\omega). \quad (4.12)$$

The function J'_y is called the Fréchet-derivative of J at y and can be considered the infinite-dimensional equivalent to the derivative. Accordingly, the function J has a stationary point at y if

$$J'_y(\omega) = 0 \quad \text{for all } \omega \in \Omega.$$

If J is convex, its Fréchet subdifferential at y is defined as

$$\partial J(y) = \left\{ u \in \mathcal{L}_\infty : J(y + d) - J(y) \geq \int u(\omega) d(\omega) \, d\mu(\omega) \quad \forall d \in \mathcal{L}_1 \right\}.$$

Accordingly, a function $J': \mathcal{L}_1 \rightarrow \mathcal{L}_\infty$ is a generalized Fréchet derivative, if for every $y \in \mathcal{L}_1$

$$J'(y) \in \partial J(y).$$

Now let J be a function of K elements $y = (y_1, \dots, y_K) \in \mathcal{L}_1$, i.e., $J: \mathcal{L}_1^K \rightarrow \mathbb{R}$. The partial directional derivative of J with respect to y_k in the direction d is defined as the limit

$$J'_{y_k}(y; d) := \lim_{t \rightarrow 0} \frac{J((y_1, \dots, y_k + td, \dots, y_K)) - J(y_1, \dots, y_K)}{t}, \quad t \in \mathbb{R}$$

and, if it is linear, can be written as

$$J'_{y_k}(y; d) = \int J'_{y_k}(\omega) d(\omega) \, d\mu(\omega),$$

where $J'_{y_k} \in \mathcal{L}_\infty$ is called the partial Fréchet-derivative of J with respect to y_k . The function J has a stationary point at y if

$$J'_{y_k}(\omega) = 0 \quad \text{for all } \omega \in \Omega \quad \text{and all } k \in \{1, \dots, K\}.$$

If J is convex in y , its partial Fréchet subdifferential at y is defined as

$$\partial_{y_k} J(y) = \left\{ u \in \mathcal{L}_\infty : J(y_1, \dots, y_k + d, \dots, y_K) - J(y) \geq \int u(\omega) d(\omega) d\mu(\omega), d \in \mathcal{L}_1 \right\}.$$

A function $J'_{y_k} : \mathcal{L}_1 \rightarrow \mathcal{L}_\infty$ is a partial generalized Fréchet derivative if for every $y \in \mathcal{L}_1^K$

$$J'_{y_k}(y) \in \partial_{y_k} J(y).$$

A special case, that will be of interest in the next chapter, are functionals of the form

$$J(y) = \int f(y(\omega)) d\mu(\omega), \quad (4.13)$$

where $f: \mathbb{R}^K \rightarrow \mathbb{R}$ is convex function. If f_k is a partial subderivative of f with respect to its k th argument, it is the case that

$$J'_{y_k}(\omega) = f_k(y(\omega)) \quad (4.14)$$

is a partial subderivative of J with respect to y_k . The identity (4.14) can be obtained in a straightforward manner from the definitions of $\partial_{y_k} J$ and ∂f . A much more general and detailed discussion of convex functionals and their generalized derivatives can be found in [Roc68, Roc71].

4.5.3 First Order Optimality Conditions

Using the concept of dual spaces and Fréchet-differentials, the constrained minimization of convex functions in \mathcal{L}_1 can be carried out in close analogy to the finite dimensional case. Consider the optimization problem

$$\min_{y \in \mathcal{L}_1^K} J(y) \quad \text{s.t.} \quad g_i(y) \leq 0, \quad h_j(y) = 0, \quad (4.15)$$

where $i \in \{1, \dots, M_g\}$, $j \in \{1, \dots, M_h\}$ and $g_i: \mathcal{L}_1 \rightarrow \mathcal{L}_g$, $h_j: \mathcal{L}_1 \rightarrow \mathcal{L}_h$ are mappings from \mathcal{L}_1 into some Banach spaces \mathcal{L}_g and \mathcal{L}_h . The first order optimality conditions used in this thesis are the Karush–Kuhn–Tucker (KKT) conditions that are necessary for y^* to solve the constraint problem (4.15). They are given by

$$\begin{aligned} 0 &\in \partial_{y_k} J(y^*) + \sum_{i=1}^{M_g} \partial_{y_k} U_i(g_i(y^*)) + \sum_{j=1}^{M_h} \partial_{y_k} V_j(h_j(y^*)) && \text{(stationarity)} \\ g_i(y^*) &\leq 0, \quad h_j(y^*) = 0 && \text{(primal feasibility)} \\ U_i(g_i(y^*)) &\geq 0 && \text{(dual feasibility)} \\ U_i(g_i(y^*)) &= 0 && \text{(complementary slackness)} \end{aligned}$$

where $U_i \in \mathcal{L}_g^*$ and $V_i \in \mathcal{L}_h^*$ are elements of the dual spaces of \mathcal{L}_g and \mathcal{L}_h . The notation used to state the dual feasibility condition is shorthand for $U_i(g_i) \geq 0$ for all $g_i \geq 0$. If J is convex, g_i is convex for all i and h_j is affine for all j , the KKT conditions are also sufficient, i.e., every y^* that satisfies the KKT conditions solves problem (4.15).

As briefly mentioned in Section 4.3, it is assumed that the uncertainty sets \mathcal{P}_θ can be defined in term of convex constraints on a density function. Now this assumption can be made more precise. Namely, it is assumed that for every uncertainty set \mathcal{P}_θ , functions g_i and h_j exist such that the constrained $P \in \mathcal{P}_\theta$ can equivalently be written as

$$g_i(p) \leq 0, \quad h_j(p) = 0, \quad i \in \{1, \dots, M_g\}, \quad j \in \{1, \dots, M_h\}, \quad (4.16)$$

where p denotes the density function of P . In general, this assumption is rather mild and holds for all common uncertainty models, compare [KP85, GZa, GZb]

An obvious, but important implication of the KKT conditions is that if the two convex functions J and \tilde{J} have the same partial generalized derivatives, then they share the same minimizer. For easier reference, this result is fixed in a corollary.

Corollary 6 *Let $J, \tilde{J}: \mathcal{L}_1^K \rightarrow \mathbb{R}$ be two convex functions and let y^* be a solution of Problem (4.15). If it is the case that for all $k \in \{1, \dots, K\}$*

$$\partial_{y_k} J(y^*) = \partial_{y_k} \tilde{J}(y^*),$$

then y^ also solves (4.15) with J replaced by \tilde{J} .*

Corollary 6 follows immediately from the KKT conditions, which depend on J only via its partial differentials. Hence, if J and \tilde{J} share the same partial differentials at y^* , then the KKT conditions are satisfied for both and y^* is a joint minimizer of J and \tilde{J} . This result is used later in the derivation of the minimax optimal sequential test.

Chapter 5

Minimax Robust Sequential Hypothesis Tests

The derivation of the minimax sequential test in this chapter roughly follows the line of arguments in Chapter 3. However, while in Chapter 3 the reduction to an optimal stopping problem was treated in detail, this step is avoided here. Instead, the established results are used to prove the minimax property of the given solutions directly. This approach circumvents some of the technical difficulties that arise when formulating the minimax sequential testing problem from scratch in an optimal stopping framework.

5.1 Problem Formulation

The problem considered in this chapter is to design a sequential test that guarantees certain error probabilities for all $P_X \in \mathcal{P}_0$ and $P_X \in \mathcal{P}_1$ and minimizes the *maximum* expected run-length among all $P_X \in \mathcal{P}$. All uncertainty sets are assumed to be of the form (4.5) and \mathcal{P} does not need to be identical to either \mathcal{P}_0 or \mathcal{P}_1 . The corresponding optimization problem reads as

$$\min_{(\psi, \delta) \in \Pi} \max_{(P_0, P_1, P) \in \mathcal{P}_0 \times \mathcal{P}_1 \times \mathcal{P}} E_P[\tau(\psi)] \quad \text{s.t.} \quad \begin{aligned} \max_{(P_0, P_1, P) \in \mathcal{P}_0 \times \mathcal{P}_1 \times \mathcal{P}} E_{P_0}[\delta_\tau] &\leq \alpha, \\ \max_{(P_0, P_1, P) \in \mathcal{P}_0 \times \mathcal{P}_1 \times \mathcal{P}} E_{P_1}[1 - \delta_\tau] &\leq \beta. \end{aligned} \quad (5.1)$$

The distinction between the uncertainty under both hypotheses and the uncertainty in the distribution P is made for two reasons. First, it is a natural generalization of problem (3.2) under the assumption that all involved distributions are subject to uncertainty. Second, it accounts for the three performance measures in sequential testing, namely the error probabilities and the expected run-length. While the error probabilities are necessarily specified under certain hypotheses, the expected run-length can be defined for arbitrary distributions. Wald [Wal47] was the first to point out that the expected run-length of sequential tests can increase drastically, if the true distribution is indicative neither of \mathcal{H}_0 nor \mathcal{H}_1 and many authors have investigated this problem since [KW57, Lor76, LLL15]. Therefore, a test designer might want to bound the maximum expected run-length under a more relaxed uncertainty model which includes distributions that are not associated with either hypothesis.

Problem (5.1) includes several special cases that are of interest in themselves. If there is no uncertainty under either hypothesis, i.e., $\mathcal{P}_0 = \{ P_0 \}$ and $\mathcal{P}_1 = \{ P_1 \}$, Problem (5.1) is a generalized version of the Kiefer-Weiss problem [KW57], which has received considerable attention in the literature [Lor76, DN88, Pav91, ZMS13]. For $\mathcal{P} = \{ P \}$, on the other hand, Problem (5.1) corresponds to the design of a sequential test that is robust in terms of the error probabilities, but not in terms the expected run-length. More precisely, the expected run-length is minimized under the distribution P , but can be larger in case the true distribution deviates from P .

Key to understanding and solving problem (5.1) is the solution of two subproblems: how to design an optimal sequential test and how to identify least favorable distributions. While the design of optimal tests was addressed in detail in Chapter 3, the characterization of least favorable distributions is the subject of the next section. The minimax robust sequential test can then be derived by combining both results.

Before getting into the details, some additional notations need to be introduced. For the sake of compactness, the product uncertainty set $\mathcal{P}_0 \times \mathcal{P}_1 \times \mathcal{P}$ is occasionally referred to as \mathcal{U} , i.e,

$$\mathcal{U} := \mathcal{P}_0 \times \mathcal{P}_1 \times \mathcal{P}. \quad (5.2)$$

Analogously, the uncertainty set of the distributions of X_n conditioned on θ is denoted by

$$\mathcal{U}_\theta := \mathcal{P}_\theta^0 \times \mathcal{P}_\theta^1 \times \mathcal{P}_\theta. \quad (5.3)$$

The fact that the sets of optimal policies $\Pi_{\alpha,\beta}^*$ and Π_λ^* depend on the distributions underlying the sequential test is made more explicit in this section by writing

$$\Pi_{\alpha,\beta}^*(P_0, P_1, P) \quad \text{and} \quad \Pi_\lambda^*(P_0, P_1, P),$$

meaning that, for example, $\Pi_\lambda(P_0, P_1, P)$ denotes the set of all policies that solve (3.4) given the distributions (P_0, P_1, P) . Analogously, the dependence of the functions α_π^n , β_π^n and γ_π^n on the given distributions is made explicit by writing

$$\alpha_{\pi,P_0}^n(t) = E_{P_0}[\delta_\tau \mid T^n = t, \tau \geq n], \quad (5.4)$$

$$\beta_{\pi,P_1}^n(t) = E_{P_1}[1 - \delta_\tau \mid T^n = t, \tau \geq n], \quad (5.5)$$

$$\gamma_{\pi,P}^n(t) = E_P[\tau - n \mid T^n = t, \tau \geq n]. \quad (5.6)$$

Note that the subscript on the left hand side denotes the distribution with respect to which the expected value on the right hand side is taken, but *not* the distributions that might be used to calculate the test statistic T^n . The importance of the difference will become clear in the next section.

5.2 Least Favorable Distributions

As discussed in Section 4.1, the idea of the minimax approach is to design a procedure that yields the best worst-case performance among all possible procedures and all feasible scenarios. Since the performance of sequential hypothesis tests is characterized by the two error probabilities and the expected run-length of the test, this worst-case has to be specified with respect to one of the three performance measures. How to characterize the corresponding least favorable distributions is the subject of this section. First, a formal definition of “least favorable” needs to be in place.

Definition 3 Let \mathcal{P}_0 , \mathcal{P}_1 and \mathcal{P} be uncertainty sets of the form (4.5) and let $\pi = (\psi, \delta)$ be a given testing policy. A distribution

- Q_0 is least favorable with respect to the type I error probability, if

$$Q_0 \in \arg \max_{P_0 \in \mathcal{P}_0} E_{P_0}[\delta_\tau].$$

- Q_1 is least favorable with respect to the type II error probability, if

$$Q_1 \in \arg \max_{P_1 \in \mathcal{P}_1} E_{P_1}[1 - \delta_\tau].$$

- Q is least favorable with respect to the expected run-length, if

$$Q \in \arg \max_{P \in \mathcal{P}} E_P[\tau(\psi)].$$

It is important to note that the least favorable distributions are independent of each other in the sense that there is no direct coupling between them. However, all three least favorable distributions depend on the test statistic and the testing policy so that there is an indirect coupling once the policy becomes subject to the optimization. For now, it is assumed that the sequential test has already been designed, i.e., its test statistic $T^n: \Omega_X^n \rightarrow \Omega_T$ and testing policy $\pi \in \Pi$ are given and fixed.

In what follows, the least favorable distributions are stated for truncated and time-invariant tests. In both cases, the results are obtained by means of the functions α_π , β_π and γ_π . For truncated tests, the least favorable distributions are defined via a recursive maximization problem. For time-invariant tests, a sufficient optimality condition is given. In both cases the least favorable distributions Q_0 , Q_1 and Q are defined in terms of the conditional distributions $Q_{X_n|t}^0$, $Q_{X_n|t}^1$ and $Q_{X_N|t}$ introduced in Section 4.4. For time-invariant tests, whose least favorable distributions are independent of n , the notation Q_t^0 , Q_t^1 and Q_t is used.

Theorem 11 Let \mathcal{P}_0 , \mathcal{P}_1 and \mathcal{P} be uncertainty sets of the form (4.5) and let π be a truncated testing policy with horizon $N \geq 1$ and test statistic $T^n: \Omega_X^n \rightarrow \Omega_T$ that satisfies (2.24).

- The distribution Q_0 is least favorable with respect to the type I error probability if for all $n \in \{1, \dots, N-1\}$ and $t \in \Omega_T$ it holds that

$$Q_{X_n|t}^0 \in \arg \max_{P_t^0 \in \mathcal{P}_t^0} E_{P_t^0} [\alpha_{\pi, Q_0}^n(T^n) \mid T^{n-1} = t]$$

and

$$Q_{X_N|t}^0 \in \arg \max_{P_t^0 \in \mathcal{P}_t^0} E_{P_t^0} [\delta_N(t)].$$

- The distribution Q_1 is least favorable with respect to the type II error probability if for all $n \in \{1, \dots, N-1\}$ and $t \in \Omega_T$

$$Q_{X_n|t}^1 \in \arg \max_{P_t^1 \in \mathcal{P}_t^1} E_{P_t^1} [\beta_{\pi, Q_1}^n(T^n) \mid T^{n-1} = t]$$

and

$$Q_{X_N|t}^1 \in \arg \max_{P_t^1 \in \mathcal{P}_t^1} E_{P_t^1} [1 - \delta(t)].$$

- The distribution Q is least favorable with respect to the expected run-length if for all $n \in \{1, \dots, N-2\}$ and $t \in \Omega_T$

$$Q_{X_n|t} \in \arg \max_{P_t \in \mathcal{P}_t} E_{P_t} [\gamma_{\pi, Q}^n(T^n) \mid T^{n-1} = t]$$

and

$$Q_{X_{N-1}|t} \in \arg \max_{P_t \in \mathcal{P}_t} E_{P_t} [1 - \psi(t)].$$

The distribution $Q_{X_N|t} \in \mathcal{P}_t$ can be chosen arbitrarily for all $t \in \Omega_T$.

A proof is detailed in Appendix A.13. Theorem 11 states that at each time instant and for each possible state of the test statistic the least favorable distributions maximize the expected value of the respective performance measure at the next time instant, given that the test is continued under the least favorable distribution. It gives a recursive definition of the least favorable distributions in the sense that the expected values on the right hand side only depend on the distribution of $(X_m)_{m > n}$ and can be calculated according to (2.31)–(2.33).

An interesting observation is that the least favorable distribution of X_N with respect to the expected run-length is not specified. This is the case since the probability

of reaching time instant N only depends on the distribution of (X_1, \dots, X_{N-1}) . If the test indeed reaches the N th time instant, it is stopped in any case so that the last observation does not have an influence on the expected run-length. This is not the case for the error probabilities, which are affected by the last observation as well.

Such intricacies of finite horizon tests do not need to be taken into account if a time-invariant testing policy is used. The least favorable distributions for time-invariant tests are stated in the next theorem.

Theorem 12 *Let \mathcal{P}_0 , \mathcal{P}_1 and \mathcal{P} be uncertainty sets of the form (4.5) and let π be a time-invariant testing policy with test statistic $T^n: \Omega_X^n \rightarrow \Omega_T$ that satisfies (2.24) and (2.39).*

- *The distribution Q_0 is least favorable with respect to the type I error probability if for all $t \in \Omega_T$*

$$Q_t^0 \in \arg \max_{P_t^0 \in \mathcal{P}_t^0} E_{P_t^0} [\alpha_{\pi, Q_0}(T^1) \mid T^0 = t]$$

- *The distribution Q_1 is least favorable with respect to the type II error probability if for all $t \in \Omega_T$*

$$Q_t^1 \in \arg \max_{P_t^1 \in \mathcal{P}_t^1} E_{P_t^1} [\beta_{\pi, Q_1}(T^1) \mid T^0 = t]$$

- *The distribution Q is least favorable with respect to the expected run-length if for all $t \in \Omega_T$*

$$Q_t \in \arg \max_{P_t \in \mathcal{P}_t} E_{P_t} [\gamma_{\pi, Q}(T^1) \mid T^0 = t]$$

where α_{π, Q_0} , β_{π, Q_1} and $\gamma_{\pi, Q}$ are defined in (5.4)–(5.6) and can be calculated by solving the integral equations (2.41)–(2.43).

A formal proof of Theorem 12 is given in Appendix A.14. As can be seen, the time-invariance of the testing policy is reflected in time-invariant least favorable distributions, which is a significant simplification compared to the truncated sequential test. In particular the proof that a given distribution is indeed least favorable is reduced from an N -step recursion to a two-step procedure: First, the function α_{π, Q_0} , β_{π, Q_1} or $\gamma_{\pi, Q}$ is calculated by solving the respective integral equation. Second, it is checked whether or not Q_0 , Q_1 or Q also solves the corresponding maximization problem in Theorem 12.

It is important to note that Theorem 12 states a sufficient condition for Q_0 , Q_1 and Q to be least favorable, but does not guarantee the existence of least favorable distributions,

nor the corresponding functions α_{π, Q_0} , β_{π, Q_0} or γ_{π, Q_0} . In contrast to truncated tests, time-invariant tests are not guaranteed to have a finite expected run-length in general.

The sufficient conditions in Theorem 12 are used in the next section to derive the policies and least favorable distributions for minimax optimal sequential tests.

5.3 Minimax Optimal Sequential Tests

In this section, a sufficient condition for a sequential test to be minimax optimal is derived by combining the results on optimal sequential tests and least favorable distributions. In analogy to the procedure in Chapter 3, Problem (5.1) is first relaxed by replacing the explicit constraints on the error probabilities with weighted penalty terms. The solution of (5.1) is then shown to be contained in the set of solutions of the relaxed problem with approximately chosen weights.

The unconstrained version of problem (5.1) is given by

$$\min_{(\psi) \in \Pi} \max_{(P_0, P_1, P) \in \mathcal{U}} E_P[\tau(\psi, \delta)] + \lambda_0 E_{P_0}[\delta_\tau] + \lambda_1 E_{P_1}[1 - \delta_\tau], \quad (5.7)$$

with positive weights (λ_0, λ_1) . The relation between (5.1) and (5.7) is stated in the next Theorem.

Theorem 13 *If a policy $\pi \in \Pi$ and a triplet of distributions $(Q_0, Q_1, Q) \in \mathcal{U}$ solve (3.4) and it holds that*

$$\max_{P_0 \in \mathcal{P}_0} E_{P_0}[\delta_\tau] = E_{Q_0}[\delta_\tau] = \alpha,$$

$$\max_{P_1 \in \mathcal{P}_1} E_{P_1}[1 - \delta_\tau] = E_{Q_1}[1 - \delta_\tau] = \beta,$$

then π and (Q_0, Q_1, Q) also solve (5.1).

Theorem 13 is the equivalent of Theorem 3 in Chapter 3 and can be proven in an analogous way. The details are given in Appendix A.15.

In order to state the minimax testing policies and distributions that solve (5.7), it is helpful to introduce a version of the function ρ_λ with *unique* partial derivatives.

Definition 4 Let $\lambda > 0$ and distribution (P_0, P_1, P) be given. Let further $\Pi_\lambda^*(P_0, P_1, P)$ be the set of policies that solve (5.7). For every $\pi \in \Pi_\lambda^*(P_0, P_1, P)$, the function $\rho_\pi: \Omega_\rho \rightarrow \mathbb{R}_+$ is defined as

$$\rho_\pi = \rho_\lambda,$$

with unique partial generalized derivatives

$$\partial_{z_0}\rho_\pi = \lambda_0\alpha_\pi \quad \text{and} \quad \partial_{z_1}\rho_\pi = \lambda_1\beta_\pi.$$

Theorem 8 guarantees that $\lambda_0\alpha_\pi$ and $\lambda_1\beta_\pi$ are indeed feasible generalized derivatives of ρ_π . The importance of policy-dependent derivatives in the minimax case follows from a coupling between the testing policy and the least favorable distributions. In Chapter 3, where P_0 , P_1 and P are assumed to be given and fixed, it was the case that all policies satisfying the conditions in Theorem 5 were equivalent with respect to the weighted sum-cost of the sequential test. Therefore, the exact policy only became important when constraints on the error probabilities needed to be considered. Now, in the minimax case, optimal policies and least favorable distributions always need to be considered jointly. Even if the sum-cost stays constant, every change in policy that has an effect on any of the three performance measures, implies a possible change in the least favorable distributions. The notation ρ_π is used to make this dependence more explicit and the upcoming statements more concise.

The solution of (5.7) is given in the next theorem. It is the minimax equivalent to Theorem 5 in Chapter 3.

Theorem 14 Let $\lambda > 0$ and uncertainty sets \mathcal{P}_0 , \mathcal{P}_1 and \mathcal{P} of the form (4.5) be given. A policy $\pi^* = (\delta^*, \psi^*)$ and a triplet of distributions (Q_0, Q_1, Q) solve (5.7) if

1. $\pi^* \in \Pi_\lambda^*(Q_0, Q_1, Q)$, i.e., π^* is an optimal testing policy for (Q_0, Q_1, Q) in the sense of (3.4). This is the case if δ^* is of the form (3.8) and ψ^* of the form (3.16), where ρ_λ solves

$$\rho_\lambda(z, \theta) = \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) dQ_{z,\theta}(x) \right\}$$

with g_λ defined in (3.10).

2. For every $(z, \theta) \in \Omega_\rho$ it holds that the conditional distributions $Q_{z,\theta}^0$, $Q_{z,\theta}^1$ and $Q_{z,\theta}$ are first order optimal solutions of the problem

$$\max_{(P_{z,\theta}^0, P_{z,\theta}^1, P_{z,\theta}) \in \mathcal{U}_\theta} \int \rho_{\pi^*} \left(z_0 \frac{p_{z,\theta}^0(x)}{p_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{p_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) dP_{z,\theta}(x), \quad (5.8)$$

where ρ_{π^*} is given in Definition 4.

A proof is detailed in Appendix A.16. In fact, Theorem 14 merely states the idea of the minimax principle in a mathematical form. An *optimal* test for the *least favorable* distributions is minimax optimal. Consequently, the condition on the minimax testing policy is to be optimal for (Q_0, Q_1, Q) in the sense of Chapter 3. This implies that the test statistic is chosen according to Corollary 1. On the other hand, the condition on the least favorable distributions is to maximize the cost of the sequential test. The exact meaning and implications of distributions being least favorable or cost maximizing in the sense of Theorem 14 deserves a more detailed discussion, which is deferred to the next section. For now, the technical definition of the least favorable distributions in Theorem 14 is sufficient.

It is worth noting that the two conditions in Theorem 14 can be combined into the single integral equation

$$\rho_\pi(z, \theta) = \min \left\{ g_\lambda(z), 1 + \max_{(P_{z,\theta}^0, P_{z,\theta}^1, P_{z,\theta}) \in \mathcal{U}_\theta} \int \rho_\pi \left(z_0 \frac{p_{z,\theta}^0(x)}{p_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{p_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) dP_{z,\theta}(x) \right\},$$

which makes the joint optimization over the optimal policy and the least favorable distributions more explicit. In this condensed form it captures all elements of the minimax robust test and the mutual coupling between them: The *stopping rule* is optimized by requiring ρ_π to solve the outer integral equation. The *test statistic* is optimized by requiring it to be the likelihood ratio of the least favorable distributions. The *distributions* are optimized by requiring them to maximize the cost of the optimal test. This extensive coupling causes the design of minimax robust sequential tests to be significantly more involved than the design of minimax robust fixed sample size tests, for which the least favorable distributions do not depend on the optimal decision rule—compare [Hub65, HS73, FZ15b] and the first example in Section 5.5.2.

Using the solution of the relaxed problem (5.7), it is straightforward to give a sufficient condition for a policy and a triplet of distributions to be a solution of the constrained minimax sequential testing problem (5.1).

Corollary 7 *Let \mathcal{P}_0 , \mathcal{P}_1 and \mathcal{P} be uncertainty sets of the form (4.5). A policy π^* and a triplet of distributions (Q_0, Q_1, Q) solve (5.1) if they satisfy the conditions in Theorem 14 and it additionally holds that*

$$\partial_{z_0} \rho_{\pi^*}(1, 1, \theta_0) = \alpha \quad \text{and} \quad \partial_{z_1} \rho_{\pi^*}(1, 1, \theta_0) = \beta.$$

Corollary 7 follows immediately from Theorem 13 and Theorem 14.

With Theorem 14 and Corollary 7 in place, minimax robust sequential tests are completely specified in terms of sufficient conditions on their testing policies and test statistics. However, before addressing the actual design of minimax robust sequential tests, it is worthwhile to inspect the maximization problem (5.8) in more detail. This is done in the next section, where ρ_π is identified as a statistical similarity measure between triplets of distributions. While this detour into the field of statistical similarities and distances is not necessary from an optimization point of view, it adds significantly to the understanding and interpretation of the rather technical results presented in this section. A brief overview of statistical similarity measures and their relation to the minimax robust sequential test is given next.

5.4 Statistical Similarity Measures

The subject of this section is the characterization of least favorable distributions and how to interpret them in a more intuitive sense. In the previous section, the least favorable triplet (Q_0, Q_1, Q) was defined as the maximizer of a functional of the form

$$I_f(P_0, P_1, P) = \int f\left(\frac{p_0(x)}{p(x)}, \frac{p_1(x)}{p(x)}, x\right) dP(x), \quad (5.9)$$

where f is determined by ρ_π and additionally depends on z and θ . Functionals of this form, i.e., the expected value of a convex/concave function applied to a likelihood ratio, are well known to define statistical similarity measures between distributions. Before addressing the particular similarity measure induced by ρ_π , a short introduction to the subject in general is given.

Measures for statistical similarity are frequently encountered in many fields related to probability theory and statistics. Well-known examples are the Kullback-Leibler divergence, the total variation distance and the Hellinger distance. All of them are special cases of a class of statistical similarity measures known as *f-divergences*.

Definition 5 Let $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ be a convex function with $f(1) = 0$ and let P_1, P_2 be two probability measures on some measurable space (Ω, \mathcal{F}) . The functional

$$D_f(P_1, P_2) = \int_{\{p_2>0\}} f\left(\frac{p_1(\omega)}{p_2(\omega)}\right) p_2(\omega) d\mu(\omega) + f'(\infty) \int_{\{p_2=0\}} p_1(\omega) d\mu(\omega),$$

where

$$f'(\infty) := \lim_{t \rightarrow \infty} \frac{f(t)}{t} \in (-\infty, \infty],$$

is called *f*-divergence of P_1 and P_2 .

f -divergences as a measure for the similarity of two distributions were introduced almost simultaneously by Csiszár [Csi63], Morimoto [Mor63] and Ali and Silvey [AS66] and have applications in information theory, signal processing , machine learning and many more areas. See [LV87,Par05] and references therein for a detailed and up-to-date treatment of f -divergences and their applications.

A natural generalization of the f -divergence is known as *f -dissimilarity* and extends the concept of a convex distance measure to multiple distributions. In order to define f -dissimilarities, homogeneous functions needs to be introduced.

Definition 6 A function $f: \mathbb{R}_+^K \rightarrow \mathbb{R}$, $K \in \mathbb{N}$, is called homogeneous if it holds that

$$f(cy) = cf(y). \quad (5.10)$$

for all $y \in \mathbb{R}_+^K$ and all $c \in [0, \infty)$.

More elaborate definitions of homogeneity exist in the literature, but in the context of this work the one given above is sufficient. The associations of homogeneity with linearity that are invoked by this definition can be misleading. Homogeneous functions are in general neither linear nor monotonic.

Definition 7 Let $f: \mathbb{R}_+^K \rightarrow \mathbb{R}$, $K \in \mathbb{N}$, be a convex homogeneous function and let P_1, \dots, P_K , be probability measures on some measurable space (Ω, \mathcal{F}) . The functional

$$D_f(P_1, \dots, P_K) = \int f(p_1(\omega), \dots, p_K(\omega)) d\mu(\omega) =: \int f(p_0, \dots, p_K) d\mu$$

is called (K -dimensional) f -dissimilarity of P_1, \dots, P_K .

Since the class of f -dissimilarities includes the class of f -divergences,¹ the same notation is used for both. Unless explicitly stated otherwise, D_f refers to Definition 7. Györfi and Nemetz introduced and studied f -dissimilarities in a sequence of publications in the 1970s [GN75, GN77, GN78] and showed that all essential properties of f -divergences carry over to f -dissimilarities.

f -dissimilarities play an important role in the general theory of statistical decision making since they emerge naturally as measures for the *separability* of hypotheses.

¹For distributions with positive densities, this follows immediately by factoring out one of the densities. A general proof and more details on the relation between f -divergences and f -dissimilarities can be found in [GN77].

The derivation of the least favorable distributions in the previous section is an example for this property. In particular, the connection between f -dissimilarities and Bayesian risks has been a topic of high interest in statistics [NWJ09], signal processing [Var11] and machine learning [RW11]. In a nutshell, it has been shown that for every concave cost function the Bayesian risk of a decision making policy defines an f -dissimilarity.

In signal processing and communications, the connection between statistical distance measures and the performance of detectors has long been known and exploited. The design of maximally separable signals, for example, can be carried out by maximizing the Bhattacharyya distance between their distributions [Kai67]. In sequential detection, Wald showed that asymptotically the expected run-length of a sequential test is inversely proportional to the Kullback-Leibler divergence of P_0 and P_1 [Wal47].² The relation of convex similarity measures to robust decision making has been addressed by Huber [HS73], Poor [Poo80], Kassam [Kas81] and more recently Guntuboyina [Gun11] to name just a few.

For the definition of least favorable distributions in the context of minimax robust sequential tests, a minor variation on the concept of f -dissimilarities is useful. First, since ρ_λ is concave, it is more natural to define the functional in (5.9) in terms of a concave function f instead of a convex function. Second, the update of the history of the random process, i.e., the $\xi_{z,\theta}(x)$ term in the argument of ρ_π , needs to be incorporated in the definition. In general, this can be done by allowing f to depend on the integration variable directly and not only via the density functions.

Definition 8 Let P_1, \dots, P_K be probability measures on some measurable space (Ω, \mathcal{F}) and let

$$\begin{aligned} f: \mathbb{R}_+^K \times \Omega &\rightarrow \mathbb{R} \\ (y_1, \dots, y_K, \omega) &\mapsto f(y_1, \dots, y_K, \omega) \end{aligned}$$

be homogeneous and concave in (y_1, \dots, y_K) . The functional

$$I_f(P_1, \dots, P_K) = \int f(p_1(\omega), \dots, p_K(\omega), \omega) d\mu(\omega)$$

is called (K -dimensional) f -similarity of P_1, \dots, P_K .

Introducing f -similarities in addition to f -dissimilarities is by no means necessary. However, it simplifies the discussion of minimax robust sequential test in the sense

²Since neither the Kullback–Leibler divergence nor the concept of f -divergences had been introduced by the time Wald published his works on sequential analysis, he did not refer to the expressions as such.

that using a positive measure for the similarity of distributions is more convenient than using a negative measure for their dissimilarity. In other words, it allows one to define the least favorable distributions as being *most similar* instead of being *least dissimilar*. Using the concept of f -similarities, an intuitive definition of the least favorable distributions can be given.

Corollary 8 *At every time instant $n \in \mathbb{N}_0$, the least favorable distributions $Q_{z,\theta}^0$, $Q_{z,\theta}^1$ and $Q_{z,\theta}$ of X_{n+1} , conditioned on the state $(z^n, \theta_n) = (z, \theta)$, are the feasible distributions that are most similar with respect to the f -similarity defined by*

$$f_\rho(y_0, y_1, y_2, \omega) = \rho_\pi \left(z_0 \frac{y_0}{y_2}, z_1 \frac{y_1}{y_2}, \xi_{z,\theta}(\omega) \right) y_2,$$

with ρ_π given in Theorem 14.

In Appendix A.17, it is shown that f_ρ is indeed a valid f -similarity.

The result that the least favorable distributions for the minimax robust sequential test are most similar with respect to some convex/concave similarity measure sits well in the wider context of robust hypothesis tests and allows for some interesting comparisons to other robust methods, in particular the minimax fixed sample size test. In [HS73] it has been shown that a pair of distributions (Q_0, Q_1) is least favorable for a minimax fixed sample size test, if it minimizes *all* f -divergences $D_f(P_0, P_1)$, irrespective of the particular choice of f . The existence of distribution pairs with this property is not guaranteed for all uncertainty models, even if the uncertainty sets are convex—see [FZ15b] and the references therein. However, for many useful models such as the ε -contamination model or the density band model, which will be introduced in Section 5.5, least favorable pairs can indeed be found.

The fact that least favorable pairs for fixed sample size tests minimize all f -divergences simultaneously implies that the design of the least favorable distributions and the optimal decision rule can be separated. First, Q_0 and Q_1 can be determined, then an optimal test for this pair can be designed using standard techniques. For the minimax sequential test, this separation is no longer possible. Instead, the least favorable distributions need to maximize a very particular f -similarity that depends on the optimal testing policy, the current state of the test statistic and the target error probabilities α and β . On the one hand, this coupling significantly complicates the design of minimax robust sequential tests. On the other hand, it can be seen how the same principles are underpinning the test design: Both tests achieve minimax optimality by using a policy that leads to the *best separation* of the *most similar* distributions. In this sense, the

minimax robust sequential test is a straightforward extension of its fixed sample size counter part, despite the technical differences.

A question that warrants further investigation is whether distributions exist that minimize all (K -dimensional) f -dissimilarities over a given uncertainty set. In the context of robust sequential hypothesis testing, this question translates into whether uncertainty sets exists that result in least favorable distributions that are *independent of the state of the test statistic*. The results in the previous sections and the examples in Section 5.8 show that the latter cannot be the case in general. Nevertheless, state independent least favorable distributions might exists for special cases.

It becomes clear at this point that obtaining more tangible results is difficult without defining a more specific uncertainty model. For this reason, the density band uncertainty model is introduced and analyzed in the next section. It was first suggested by Kassam [Kas81] and restricts the true, unknown densities to lie within a band with given upper and lower bounds. Several popular uncertainty models can be shown to be special cases of the band model, most prominently the ε -contamination model. The density band model is discussed in more detail in the next section.

In the broader context of the dissertation introducing the band model serves two purposes: First, it is used to provide numerical examples for minimax robust sequential tests under reasonably realistic distributional uncertainties. Second, it also offers additional insight into the properties of least favorable distributions and their relation to the minimax principle. Although the validity of these findings is obviously limited to the band model, the underlying ideas can be used as guidelines for the design of heuristics and robust sequential tests in general.

5.5 The Density Band Uncertainty Model

In [Kas81], an uncertainty model for probability distributions was proposed that later became known as the band model. It allows each hypothesis to be formulated in terms of a density band, within which the true density is supposed to lie, and generalizes the outlier models suggested by Huber [Hub65], Österreicher [Ö78], Levy [Lev08] and others. It can further be interpreted as both an ε -contamination model with bounded outlier distributions, or a model for general uncertainties in the distribution shape that can be specified without introducing nominals.

Let (Ω, \mathcal{F}) be a measurable space and let \mathcal{M}_μ denote the set of all probability measures on (Ω, \mathcal{F}) that admit densities with respect to the measure μ . The density band uncertainty model covers composite hypotheses of the form

$$\mathcal{P}^= = \{P \in \mathcal{M}_\mu : p' \leq p \leq p''\}, \quad (5.11)$$

where p' and p'' fulfill

$$0 \leq p' \leq p'', \quad \int_{\Omega} p'(x) d\mu(x) =: P'(\Omega) \leq 1, \quad \int_{\Omega} p''(x) d\mu(x) =: P''(\Omega) \geq 1.$$

The band constraint is indicated by the notation $\mathcal{P}^=$. Note that P' and P'' are measures on (Ω, \mathcal{F}) , but not necessarily probability measures, and P'' does not need to be finite.

Alternatively, the band model can be interpreted as an ε -contamination model with bounded outlier distribution. In order to see this, note that every $p \in \mathcal{P}^=$ can be written as $p = p' + \varepsilon h$, where

$$\varepsilon = 1 - P'(\Omega)$$

and h denotes the density of an outlier distribution $H \in \mathcal{M}_\mu$. In contrast to the ε -contamination model, however, not every H is feasible under the density band model. More precisely, h has to be chosen such that

$$p' \leq p' + \varepsilon h \leq p''$$

which yields the constraint

$$\varepsilon h \leq p'' - p'. \quad (5.12)$$

By definition of \mathcal{M}_μ , both sides of (5.12) can be integrated over all $B \in \mathcal{F}$ so that (5.12) can equivalently be written as

$$\varepsilon H \leq P'' - P'$$

and (5.11) can be written as

$$\mathcal{P}^= = \{P \in \mathcal{M}_\mu : P = P' + \varepsilon H, \varepsilon H \leq P'' - P'\}. \quad (5.13)$$

In this regard, the band model is an ε -contamination model that allows the incorporation of *a priori* knowledge in form of additional constraints on the outlier distribution.

The least favorable distributions for a minimax robust fixed sample size test under density band uncertainties were first derived in [Kas81] and more recently in [FZ15b].

The definition of the least favorable distributions in [Kas81] is explicit, but rather cumbersome.³ The definition in [FZ15b] is more concise, but only given implicitly. In the next section, the latter approach is extended to general f -similarities and an algorithm is presented that makes use of the implicit definition to iteratively approximate the density functions of the most similar distributions.

5.5.1 Most Similar Distributions

In this section, a sufficient condition for distributions to be most similar under density band uncertainties is derived. That is, it is assumed that an f -similarity and uncertainty sets of the density band form are given and that the goal is to determine the corresponding most similar, i.e., least favorable, distributions. In the context of robust sequential hypothesis testing, this problem can be seen as the counterpart to the design of optimal sequential tests discussed in Chapter 3. While the latter deals with the derivation of optimal testing policies for known distributions, this section deals with the derivation of least favorable distributions for known cost functions ρ_π . Being able to solve both subproblems is an important step towards solving the joint problem in Theorem 14.

The discussion in this section is not limited to the three-dimensional case that arose in the derivation of the binary minimax robust sequential hypothesis test, but treats the general problem of maximizing K -dimensional f -similarities. On the one hand, this significantly increases the generality of the result without substantially increasing the technical difficulties. On the other hand, the K -dimensional case can be seen as an outlook on robust tests between *multiple* hypotheses, which are briefly discussed in Chapter 6. Throughout this section $q = (q_1, \dots, q_K)$ denotes the vector of most similar densities and the notation \bar{y}_k is used to denote a vector whose k th element has been removed, i.e., $\bar{y}_k = (y_1, \dots, y_{k-1}, y_{k+1}, \dots, y_K)$.

The most similar distributions with respect to a K -dimensional f -similarity are defined by the optimization problem

$$\max_{P_k \in \mathcal{P}_k^=} I_f(P_1, \dots, P_K), \quad (5.14)$$

where $k \in \{1, \dots, K\}$, all $\mathcal{P}_k^=$ are of the form (5.11) and I_f is defined in Definition 8.

³The theorem spans the space of a column and distinguishes between four special cases, each involving a piecewise definition of the densities. In order to know which case holds, one has to check the existence or non-existence of in total six constants that have to be chosen such that the solutions are valid densities. In some cases the solution involves a function that is guaranteed to exist, but is not specified explicitly.

In order to proceed with the optimization, it is helpful to express the constraints on the densities explicitly. The resulting optimization problem reads as

$$\begin{aligned} & \max_{p_k \in \mathcal{L}_1} \int f(p_1(\omega), \dots, p_K(\omega), \omega) d\mu(\omega) \\ \text{s.t. } & p_k - p'_k \geq 0, \quad p_k - p''_k \leq 0, \quad 1 - \int p_k d\mu = 0, \end{aligned} \tag{5.15}$$

where $\mathcal{L}_1 = \mathcal{L}_1(\Omega, \mu)$ denotes the set of all μ -integrable functions on Ω . A sufficient condition for (Q_1, \dots, Q_K) to be a solution of (5.14) is stated in the next Theorem. It constitutes the main result of this section.

Theorem 15 *The distributions (Q_1, \dots, Q_K) are most similar in the sense of (5.14), if for all $k \in \{1, \dots, K\}$ their densities satisfy*

$$q_k(\omega) = \begin{cases} p''_k(\omega) & f_{y_k}(q(\omega), \omega) > c_k \\ f_{y_k}^{-1}(\bar{q}_k(\omega), \omega, c_k) & f_{y_k}(q(\omega), \omega) = c_k \\ p'_k(\omega) & f_{y_k}(q(\omega), \omega) < c_k \end{cases} \tag{5.16}$$

for some $c_1, \dots, c_K \in \mathbb{R}$ and some

$$f_{y_k} \in \partial_{y_k} f,$$

where $f_{y_k}^{-1}$ denotes the inverse of f_{y_k} with respect to y_k and

$$f_{y_k}^{-1}(\bar{y}_k, \omega, c_k) \in \{y_k \in \mathbb{R}_+ : f_{y_k}(y_1, \dots, y_K, \omega) = c_k\}.$$

A proof of Theorem 15 is detailed in Appendix A.18. From (5.16) it can be seen that the partial generalized derivatives of f play a crucial role not only in determining the regions where the least favorable densities equal their bounds, but also in determining their shape outside these regions. This becomes more obvious if the result in Theorem 15 is stated in an alternative, more expressive form.

Corollary 9 *The distributions (Q_1, \dots, Q_K) are most similar in the sense of (5.14), if their densities satisfy*

$$q_k(\omega) = \min\{p''_k(\omega), \max\{f_{y_k}^{-1}(\bar{q}_k(\omega), \omega, c_k), p'_k(\omega)\}\} \tag{5.17}$$

for some $c_1, \dots, c_K \in \mathbb{R}$ and some $f_{y_k} \in \partial_{y_k} f$.

The corollary is proven in Appendix A.19. Stating the least favorable distributions in the form (5.17) is advantageous for several reasons. First, it eliminates the dependence of q_k on itself, which is still present in the explicitly piecewise condition (5.16). Corollary 9 gives an expression for q_k solely in terms of the remaining least favorable densities \bar{q}_k and the scalar c_k . Knowing \bar{q}_k , the missing density q_k can, hence, be found via a search over c_k . In the upcoming section, an iterative algorithm for the calculation of the least favorable distributions is proposed that is based on this idea. Second, it can be seen from (5.17) that q_k is in fact a projection of $f_{y_k}^{-1}(\bar{q}_k(\omega), \omega, c_k)$ on the corridor of feasible densities defined by p'_k and p''_k . In the limit, i.e., $p''_k \rightarrow \infty$ and $p'_k \rightarrow 0$, it hence holds that $q_k(\omega) = f_{y_k}^{-1}(\bar{q}_k(\omega), \omega, c_k)$, so that the latter is the *unconstrained* least favorable density for the given \bar{q} .

5.5.2 Examples

In this section, two examples for similarity measures are given whose most similar densities can be calculated analytically. First, an existing expression for the least favorable distributions of binary fixed sample size tests under band uncertainties is re-derived using the general result in Theorem 15. For this particular case, the least favorable distributions are independent of the function f . This is no longer the case in the second example, where the objective function is chosen as a weighted sum of Kullback–Leibler divergences. The latter example is used to highlight the dependence of the least favorable distributions on f , but also to show how the presented result can be used in a context that is unrelated to the density band model.

In [HS73] it was shown that the least favorable distributions of a minimax robust fixed sample size test jointly minimize all f -divergences, i.e.,

$$\int \varphi\left(\frac{q_1}{q_0}\right) q_0 d\mu \leq \int \varphi\left(\frac{p_1}{p_0}\right) p_0 d\mu$$

for all $(P_0, P_1) \in \mathcal{P}_0^= \times \mathcal{P}_1^=$ and all convex functions φ . In this example, it is demonstrated how q_0 and q_1 can be obtained from the general characterization of most similar distributions in Theorem 15.

One-dimensional f -divergences are obtained from the class of f -similarities introduced in Definition 8 by choosing

$$f(y_1, y_2) = -\varphi\left(\frac{y_1}{y_2}\right) y_2,$$

where $\varphi: \mathbb{R}_+ \rightarrow \mathbb{R}$ is convex and $\varphi(1) = 0$. The partial generalized derivatives of f calculate to

$$f_{y_1}(y_1, y_2) = -\partial\varphi\left(\frac{y_1}{y_2}\right) =: \tilde{f}_{y_1}\left(\frac{y_1}{y_2}\right)$$

and

$$f_{y_2}(y_1, y_2) = \partial\varphi\left(\frac{y_1}{y_2}\right)\frac{y_1}{y_2} - \varphi\left(\frac{y_1}{y_2}\right) =: \tilde{f}_{y_2}\left(\frac{y_1}{y_2}\right),$$

where $\partial\varphi$ denotes the generalized derivative of φ . Note that both derivatives depend on y_1 and y_2 only via the ratio y_1/y_2 .

Since the partial derivatives f_{y_1} and f_{y_2} are nondecreasing, their inverse functions in the sense of Theorem 15 are well defined and it holds that

$$\begin{aligned} \tilde{f}_{y_k}\left(\frac{y_1}{y_2}\right) &= \tilde{c}_k \\ \frac{y_1}{y_0} &= \tilde{f}_{y_k}^{-1}(\tilde{c}_k) \\ y_1 &= \tilde{f}_{y_k}^{-1}(\tilde{c}_k)y_2 \\ y_1 &= c_k y_2, \end{aligned}$$

where $k \in \{1, 2\}$ and $c_k := f_{y_k}^{-1}(\tilde{c}_k)$. Therefore, the two densities that are most similar with respect to all f -divergences are of the form

$$\begin{aligned} q_1(\omega) &= \min\{ p_1''(\omega), \max\{ c_1 q_2(\omega), p_1'(\omega) \} \}, \\ q_2(\omega) &= \min\{ p_2''(\omega), \max\{ c_2 q_1(\omega), p_2'(\omega) \} \}, \end{aligned}$$

which is in agreement⁴ with the result in [FZ15b].

For $K > 2$, this simple relation no longer holds since in this case the fact that f_{y_k} is constant does not imply that the likelihood ratios are constant as well. Instead, the vector of likelihood ratios can and has to follow a contour line of f_{y_k} .

This dependence on f is illustrated in the next example, where it is assumed that I_f is a weighted sum of Kullback-Leibler divergences, i.e.,

$$f(y_1, \dots, y_N) = \sum_{k=1}^{K-1} a_n \log\left(\frac{y_k}{y_K}\right) y_K, \quad (5.18)$$

with convex combination weights

$$a_1, \dots, a_K \geq 0 \quad \text{and} \quad \sum_{k=1}^K a_k = 1.$$

⁴See [FZ15b] for an extension of this result to densities that do not share the same support.

The corresponding (concave) functional is given by

$$I_f(P_1, \dots, P_K) = - \sum_{k=1}^{K-1} a_k D_{\text{KL}}(P_k \| P_K),$$

where D_{KL} denotes the Kullback-Leibler divergence. The partial derivatives of f in (5.18) calculate to

$$f_{y_k}(y_1, \dots, y_K) = a_n \frac{y_K}{y_k} =: \tilde{f}_{y_k}\left(\frac{y_k}{y_K}\right)$$

for $n \in \{1, \dots, N-1\}$ and

$$\begin{aligned} f_{y_K}(y_1, \dots, y_K) &= \sum_{k=1}^{K-1} a_k \left(\log \frac{y_k}{y_K} - 1 \right) \\ &=: \tilde{f}_{y_K}\left(\frac{y_1}{y_K}, \dots, \frac{y_{K-1}}{y_K}\right). \end{aligned}$$

The inverse functions are again obtained by solving

$$\tilde{f}_{y_k} = \tilde{c}_k$$

for y_k and read as

$$y_k = \frac{a_k}{\tilde{c}_k} y_K =: c_k y_K, \quad k < K$$

and

$$y_K = e^{-(\tilde{c}_K+1)} \prod_{k=1}^{K-1} y_k^{a_k} =: c_K \prod_{k=1}^{K-1} y_K^{a_k}.$$

The most similar densities are, hence, of the form

$$q_k(\omega) = \min\{p''_k(\omega), \max\{c_k q_K(\omega), p'_k(\omega)\}\} \quad (5.19)$$

for $k \in \{1, \dots, K-1\}$ and

$$q_K(\omega) = \min\{p''_K(\omega), \max\{c_K q_1(\omega) \cdots q_{K-1}(\omega)^{a_{K-1}}, p'_K(\omega)\}\}. \quad (5.20)$$

This means that q_1, \dots, q_{K-1} are the projections of q_K onto the corridors $\mathcal{P}_1^\pm, \dots, \mathcal{P}_{K-1}^\pm$ and q_K in turn is the projection of the weighted geometric mean of q_1, \dots, q_{K-1} onto \mathcal{P}_K^\pm .

The above result can be useful beyond the context of the density band model and robust sequential detection. In [Gun11], a lower bound on the sum of f -divergences has been derived that involves the calculation of a quantity J_f defined as

$$J_f := \min_{P_K \in \mathcal{M}_\mu} \frac{1}{K} \sum_{k=1}^{K-1} D_f(P_k \| P_K), \quad (5.21)$$

where D_f denotes the corresponding f -divergence. For the sum of Kullback-Leibler divergences, the author of [Gun11] resorts to an approximation for J_f . Using the expression for the least favorable density q_K in (5.20), the exact result can be derived.

Problem (5.21) can be cast as a special case of (5.14) by choosing p_1, \dots, p_{K-1} fixed and relaxing the band constraints on p_K enough to be nonbinding, i.e., $p'_K \rightarrow 0$ and $p''_K \rightarrow \infty$. For the uniformly weighted sum of Kullback-Leibler divergences it follows from (5.20) that

$$q_K(\omega) = \frac{1}{c_K} (p_1(\omega) \cdots p_{K-1}(\omega))^{\frac{1}{K-1}}, \quad (5.22)$$

where c_K needs to be chosen such that q_K is a valid density function, i.e.,

$$c_K = \int (p_1(\omega) \cdots p_{K-1}(\omega))^{\frac{1}{K-1}} d\mu(\omega),$$

which is a generalized Bhattacharyya coefficient [Par05]. Inserting (5.22) back into (5.21) it is not hard to show that

$$J_{\text{KL}} = \min_{P_K \in \mathcal{M}_\mu} \frac{1}{K} \sum_{n=1}^{K-1} D_{\text{KL}}(P_n \| P_K) = -\log(c_K),$$

which is the exact solution of (5.21) expressed in terms of the generalized Bhattacharyya coefficient of the given distributions P_1, \dots, P_{K-1} .

In most cases, however, closed form solutions like the ones presented here are not available. For such cases, a numerical algorithm for the calculation of most similar distributions is required.

5.5.3 Calculation of the Most Similar Distributions

In this section, an algorithm is proposed that makes use of the sufficient condition stated in Corollary 9 to iteratively approximate the most similar densities. In each iteration, the roots of K nonincreasing functions of a scalar variable need to be determined. The algorithm is detailed in Table 5.1 on the next page.

The most similar densities are not necessarily unique and the solutions of Algorithm 5.1 can differ for different initial densities. The optimality of the solution is not affected by this dependence.

The termination criterion in line 7 is intentionally left vague. Any reflexive distance measure between two functions can be used to determine whether all of the densities are sufficiently close to convergence, i.e., solve the system of equations (5.17).

```

1: input: partial derivatives  $(f_{y_1}, \dots, f_{y_K})$ , lower density bounds  $(p'_1, \dots, p'_K)$ ,  

   upper density bounds  $(p''_1, \dots, p''_K)$   

2: initialize: Choose feasible initial densities  $q^0 = (q_1^0, \dots, q_K^0) \in \mathcal{P}_1^- \times \dots \times \mathcal{P}_K^-$   

   and set  $i \leftarrow 0$ .  

3: repeat  

4:   Set  $i \leftarrow i + 1$   

5:   Set  $q^i \leftarrow q^{i-1}$   

6:   for  $k = 1$  to  $K$  do  

7:     Solve

$$\int \min\{p''_k(\omega), \max\{f_{y_k}^{-1}(\bar{q}_k^i(\omega), \omega, c_k), p'_k(\omega)\}\} d\mu(\omega) = 1 \quad (5.23)$$

   for  $c_k$  and set

$$q_k^i \leftarrow \min\{p''_n(\omega), \max\{f_{y_k}^{-1}(\bar{q}_k^i(\omega), \omega, c_k), p'_k(\omega)\}\}.$$

8:   end for  

9: until  $q_k^i \approx q_k^{i-1}$  for all  $k \in \{1, \dots, K\}$   

10: return  $(q_1^i, \dots, q_K^i)$ 

```

Table 5.1: Iterative algorithm for the calculation of the least favorable densities.

The only non-generic step of the algorithm in Table 5.1 is the calculation of $f_{y_k}^{-1}$. In some cases, see the previous section, this inverse can be calculated analytically. In cases where this is not possible, the functions f_{y_k} have to be inverted numerically by solving

$$f_{y_k}(q_1(\omega), \dots, q_K(\omega), \omega) = c_k \quad (5.24)$$

for $q_k(\omega)$. Since f_{y_k} is the partial derivative of a concave function, it is guaranteed to be nonincreasing in $q_k(\omega)$, which allows for the use of efficient numerical root-finding algorithms [PG02]. This step can be simplified by first checking if either

$$f_{y_k}(q(\omega), \omega) \Big|_{q_k(\omega)=p'_k(\omega)} < c_k \quad (5.25)$$

or

$$f_{y_k}(q(\omega), \omega) \Big|_{q_k(\omega)=p''_k(\omega)} > c_k. \quad (5.26)$$

If (5.25) holds, $f_{y_k}^{-1}(\bar{q}(\omega), \omega, c_k)$ has to be smaller than $p'_k(\omega)$ so that $q_k(\omega) = p'_k(\omega)$. If (5.26) holds, $f_{y_k}^{-1}(\bar{q}(\omega), \omega, c_k)$ has to be larger than $p''_k(\omega)$ and $q_k(\omega) = p''_k(\omega)$. If neither (5.25) nor (5.26) hold, root-finding can be performed to determine $q_k(\omega)$ on the interval $[p'_k(\omega), p''_k(\omega)]$. In the worst case, this procedure has to be repeated for every $\omega \in \Omega$.

By construction, it is the case that if the algorithm in Table 5.1 converges, it converges to a global minimizer of I_f . The question under which conditions the algorithm is guaranteed to converge is more intricate. The type of fixed-point iteration it implements goes by the names coordinate minimization [HB91], coordinate descent [Wri15] or nonlinear Gauss-Seidel method [GS00] and has been studied under various assumptions in the literature. However, general convergence results for convex functions in Banach spaces do not exist. In [GS00], it is shown that coordinate minimization is guaranteed to converge if

- the objective function is differentiable and pseudoconvex in each variable,
- the optimization variables are finite-dimensional,
- and the set of feasible solutions is compact convex.

This implies convergence of the algorithm in Table 5.1 if

- the function f is differentiable and
- the sample spaces Ω is finite.

The first assumption makes sure that the coordinate descent method cannot get stuck in a “corner”, where f is increasing in some direction, but nonincreasing in each individual argument. The second assumption reduces the density functions $q_k \in \mathcal{L}_1$ to vectors $q_k \in \mathbb{R}^{|\Omega|}$, which eliminates possible corner cases that arise in general Banach spaces. In fact, the convergence of a class of subgradient algorithms for the minimization of convex functions in Banach spaces has been shown in [AIS97]. However, the results do not cover coordinate descent methods.

While the above restrictions are quite strong in theory, they are typically not critical in practice. In particular, the assumption that the sample space is finite and compact, holds automatically whenever a finite grid is introduced to represent the density functions numerically. In general, the known cases where coordinate descent methods fail to converge are carefully designed counter examples and quite unlikely to occur in practice [Wri15]. The numerical results in Section 5.8 were obtained by means of the presented algorithm.

For the special case of minimizing f -divergences, the convergence of the algorithm in Table 5.1 has been proven in [FZ15b]. Whether or not this result extends to higher-dimensional cases is a delicate question that requires a careful analysis and is beyond the scope of this work.

5.6 Minimax Robust Sequential Tests under Density Band Uncertainties

In this section, the general result on distributions that maximize f -similarities under density band constraints is applied to the minimax robust sequential test. This yields an alternative characterization of the least favorable distributions in terms of the functions α_π , β_π and γ_π , instead of the characterization via f -similarities. Moreover, the least favorable distributions in the sequential case are compared to the least favorable distributions in the fixed sample size case and shown to follow the same underlying design principles.

Using the density band model, the uncertainty sets of the distributions of X_n conditioned on θ are given by

$$\begin{aligned}\mathcal{P}_\theta^0 &= \{P \in \mathcal{M}_\mu : p'_{\theta,0} \leq p \leq p''_{\theta,0}\}, \\ \mathcal{P}_\theta^1 &= \{P \in \mathcal{M}_\mu : p'_{\theta,1} \leq p \leq p''_{\theta,1}\}, \\ \mathcal{P}_\theta &= \{P \in \mathcal{M}_\mu : p'_\theta \leq p \leq p''_\theta\},\end{aligned}$$

where $\mathcal{M}_\mu = \mathcal{M}_\mu(\Omega_X, \mathcal{F}_X)$ is defined in (4.2). The upper and lower density bounds are time-invariant, but allowed to depend on the past observations via θ .

In Section 5.3, the distributions that are least favorable with respect to the weighted sum-cost (5.7) were shown to maximize the f -similarity

$$f_\rho(y_0, y_1, y_2, x) = \rho_\pi \left(z_0 \frac{y_0}{y_2}, z_1 \frac{y_1}{y_2}, \xi_{z,\theta}(x) \right) y_2, \quad (5.27)$$

where x is used instead of ω to indicate that the observation is generated by X . For the density band uncertainty model the first order optimality conditions were then calculated explicitly. They are stated in Theorem 15 in the previous section and can be applied readily to the problem of minimax robust sequential testing.

In Appendix A.16, it is shown that the partial derivatives of f_ρ in (5.27) are given by

$$\begin{aligned}\partial_{y_0} f_\rho(y_0, y_1, y_2, x) &= \lambda_0 z_0 \alpha_\pi \left(z_0 \frac{y_0}{y_2}, z_1 \frac{y_1}{y_2}, \xi_{z,\theta}(x) \right), \\ \partial_{y_1} f_\rho(y_0, y_1, y_2, x) &= \lambda_1 z_1 \beta_\pi \left(z_0 \frac{y_0}{y_2}, z_1 \frac{y_1}{y_2}, \xi_{z,\theta}(x) \right), \\ \partial_{y_2} f_\rho(y_0, y_1, y_2, x) &= \gamma_\pi \left(z_0 \frac{y_0}{y_2}, z_1 \frac{y_1}{y_2}, \xi_{z,\theta}(x) \right).\end{aligned}$$

Inserting these expressions into the optimality conditions (5.16) yields

$$q_{z,\theta}^0(x) = \begin{cases} p''_{0,\theta}(x), & \alpha_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) > c_0, \\ \alpha_\pi^{-1}(x), & \alpha_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) = c_0, \\ p'_{0,\theta}(x), & \alpha_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) < c_0, \end{cases} \quad (5.28)$$

$$q_{z,\theta}^1(x) = \begin{cases} p''_{1,\theta}(x), & \beta_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) > c_1, \\ \beta_\pi^{-1}(x), & \beta_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) = c_1, \\ p'_{1,\theta}(x), & \beta_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) < c_1, \end{cases} \quad (5.29)$$

$$q_{z,\theta}(x) = \begin{cases} p''_\theta(x), & \gamma_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) > c, \\ \gamma_\pi^{-1}(x), & \gamma_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) = c, \\ p'_\theta(x), & \gamma_\pi \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}(x)}{q_{z,\theta}^0(x)}, \xi_{z,\theta}(x) \right) < c, \end{cases} \quad (5.30)$$

where the additional arguments of the inverse functions have been omitted for the sake of a more compact notation. Also note that c_0 and c_1 are scaled with $\frac{1}{\lambda_0 z_0}$ and $\frac{1}{\lambda_1 z_1}$, respectively, compared to their definition in Theorem 15.

The constraints (5.28)–(5.30) allow for some deeper insight into what properties make the most similar distributions least favorable. It can be seen that in any state (z, θ) , their shape depends on the error-probabilities and the expected run-length in the *next state*, given that the likelihood ratios of the least favorable distributions themselves are used as a test statistic. Consider, for example, the constraints on $Q_{z,\theta}^0$. From (5.28), it follows that the least favorable distribution places as much probability mass as possible ($q_{z,\theta}^0 = p''_{0,\theta}$) on regions of the sample space for which the following holds: If the next sample is taken from this region, it leads to a new state of the test statistic in which the type I error probability is larger than c_0 . In this way, the least favorable distribution maximizes the chances that the test enters a state with large ($> c_0$) type I error probability in the next time instant. Analogously, the least favorable distribution

places as little probability mass as possible ($q_{z,\theta}^0 = p'_{0,\theta}$) on regions of the sample space that lead to a state with type I error probability smaller than c_0 . This minimizes the chances of entering a state with small ($< c_0$) error probability. Finally, on the remaining region of the sample space, the least favorable distribution is chosen such that the type I error probability in the next state is constant and equal to c_0 . In contrast to the first two properties, the rationale for this particular assignment of probability mass might not be obvious at first glance, but is critical for minimax procedures. It implies that all observations from the region where $\alpha_\pi = c_0$ holds are equivalent in terms of the type I error probability. Ultimately, it is this equivalence that makes the minimax sequential test robust against distributional uncertainties. By choosing the test statistic such that observations from whole regions of the sample space lead to the very same performance, the exact distribution become less important. In fact, as long as probability mass is shifted within the region of constant performance, the test becomes entirely insensitive to changes in the true distribution and hence to deviations from the underlying model. This strategy of designing a procedure such that its performance profile is flat over large regions of the state space is at the heart of every minimax procedure.

The same reasoning holds for the type II error probability and the expected-length in (5.29) and (5.30). Again, the least favorable distributions are chosen according to whether the respective performance measure is larger or smaller than a certain threshold and such that it is constant everywhere else.

5.7 Test Design

In the previous sections, an implicit characterization of optimal policies and least favorable distributions was derived. This section deals with the question how to make use of these results for the actual design of minimax robust sequential tests. In general, designing a minimax optimal test means solving the optimization problem

$$\begin{aligned} & \max_{(P_{z,\theta}^0, P_{z,\theta}^1, P_{z,\theta}) \in \mathcal{U}_\theta} \int \rho_\pi \left(z_0 \frac{p_{z,\theta}^0(x)}{p_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{p_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) dP_{z,\theta}(x) \\ \text{s.t. } & \rho_\pi(z, \theta) = \min \left\{ g_\lambda(z), 1 + \int \rho_\pi \left(z_0 \frac{p_{z,\theta}^0(x)}{p_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{p_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) dP_{z,\theta}(x) \right\} \\ & \pi \in \Pi_\lambda^*(P_0, P_1, P), \quad \partial_{z_0} \rho_\pi(1, 1, \theta_0) = \alpha, \quad \partial_{z_1} \rho_\pi(1, 1, \theta_0) = \beta, \end{aligned} \quad (5.31)$$

where the solution of the maximization is defined in terms of its first order optimality conditions in the sense of Theorem 14. Solving (5.31) directly is a formidable task and,

if done numerically, requires considerable processing power and memory. In contrast to the problem of optimal test design in Chapter 3, not only ρ_π needs to be discretized, but also the distribution families $(P_{z,\theta}^0, P_{z,\theta}^1, P_{z,\theta})$. The fact that the latter depend on the state of the test statistic means that an unknown triplet of distributions has to be introduced for every sample point of the state space Ω_ρ . The number of unknown variables is hence given by $M_\rho M_X$, where M_ρ denotes the number of sample points in Ω_ρ and M_X the number of sample points in Ω_X .

An obvious way to simplify the process of solving (5.31) is to use iterative techniques. The procedure suggested here is to alternate between the design of the function ρ_π and the most similar distributions $(Q_{z,\theta}^0, Q_{z,\theta}^1, Q_{z,\theta})$. First, an optimal testing policy π^0 for an initial triplet of distributions $(Q_{z,\theta}^0, Q_{z,\theta}^1, Q_{z,\theta})^0$ is designed. Then, in a second step, the least favorable distributions $(Q_{z,\theta}^0, Q_{z,\theta}^1, Q_{z,\theta})^1$ with respect to ρ_{π^0} are calculated. Based on $(Q_{z,\theta}^0, Q_{z,\theta}^1, Q_{z,\theta})^1$, the policy is updated to π^1 and so on. These steps are repeated until both ρ_{π^i} and $(Q_{z,\theta}^0, Q_{z,\theta}^1, Q_{z,\theta})^i$ have sufficiently converged, i.e.,

$$\rho_{\pi^i}(z_m, \theta_m) \approx \rho_{\pi^{i-1}}(z_m, \theta_m)$$

and

$$(Q_{z_m, \theta_m}^0, Q_{z_m, \theta_m}^1, Q_{z_m, \theta_m})^i \approx (Q_{z_m, \theta_m}^0, Q_{z_m, \theta_m}^1, Q_{z_m, \theta_m})^{i-1}$$

for all $(z_m, \theta_m)_{m \in \{1, \dots, M_\rho\}}$. The advantage of this approach is that the large problem (5.31) is split into smaller, more tractable subproblems. For the design of optimal testing policies for given distributions, the methods discussed in Section 3.4 can be used. The linear programming approach presented in Section 3.4.2 is usually most efficient. The corresponding problem formulation is straightforward and state-of-the-art linear programming solvers can reliably solve problems involving millions of variables [GO15]. In case the obtained solution is still too coarse, it can be used as an initial guess for a Newton-like iteration scheme. Once ρ_π is calculated, the least favorable distributions can be determined subsequently. The fact that ρ_π is kept constant during this step means that there is no coupling between the least favorable distributions so that they can be designed independently for different states (z, θ) . This decoupling is a significant simplification and can make the difference between being able to calculate an optimal solution or not. The decoupled problem (5.31) can either be solved in parallel for each triplet $(Q_{z_m, \theta_m}^0, Q_{z_m, \theta_m}^1, Q_{z_m, \theta_m})$, which reduces the computation time, or it can be solved sequentially, which reduces the memory consumption.

It is worth noting that iteratively designing ρ_{π^*} and $(Q_{z,\theta}^0, Q_{z,\theta}^1, Q_{z,\theta})$ does *not* correspond to iteratively designing optimal testing policies and least favorable distributions. Unless convergence has occurred, the distributions $(Q_{z,\theta}^0, Q_{z,\theta}^1, Q_{z,\theta})^i$ are not least favorable for the policy π^i . This follows from the fact that the triplet $(Q_0, Q_1, Q)^i$ only

satisfies the criteria for least favorable distributions in Theorem 14 if π^i is the corresponding optimal policy—compare also the proof in Appendix A.16. $(Q_{z,\theta}^0, Q_{z,\theta}^1, Q_{z,\theta})^i$ being least favorable with respect to π^i , however, implies convergence. There is no “clean” separation between the optimal test and the least favorable distributions since the optimal testing policy is affected by both steps of the iteration: The shape of ρ_π determines the stopping rule for a given test statistic. The latter, however, is determined by the distributions. Hence, the iterative procedure does not alternate between the design of optimal policies and least favorable distributions, but between the design of optimal policies and a combined step that jointly updates the test statistic and the distributions.

For the density band uncertainty model, the iterative solution method is still demanding in terms of computational power, but at the same time composed of simple individual steps. The iterative algorithm detailed in Section 5.5.3 reduces the maximization in (5.31) to the problem of finding roots of nonincreasing functions of scalar variables, which is a standard problem in numerical mathematics.

In summary, the results presented in this dissertation provide all the building blocks necessary to solve the minimax sequential hypothesis testing problem (5.31) under density band uncertainties.

5.8 Examples and Numerical Results

In this section, the design of minimax optimal sequential tests is illustrated by means of numerical results. Owing to the computational complexity of the design procedure, only one example is presented. Nevertheless, the example is sufficient to highlight the characteristic properties of minimax robust sequential tests.

The hypotheses testing problem considered in this section is a variation on the testing problem discussed in Section 3.5.1. The nominal hypotheses are again given by

$$\begin{aligned}\mathcal{H}_0 : \quad & X_n \sim \mathcal{N}(-0.5, 1), \\ \mathcal{H}_1 : \quad & X_n \sim \mathcal{N}(0.5, 1),\end{aligned}$$

where $\mathcal{N}(\eta, \sigma)$ denotes a Gaussian distribution with mean η and standard deviation σ and all X_n are assumed to be i.i.d. While in Section 3.5.1 the optimal test was designed under ideal conditions, here the possibility of model mismatches is taken into account. More precisely, the density band uncertainty model introduced in Section 5.5 is used

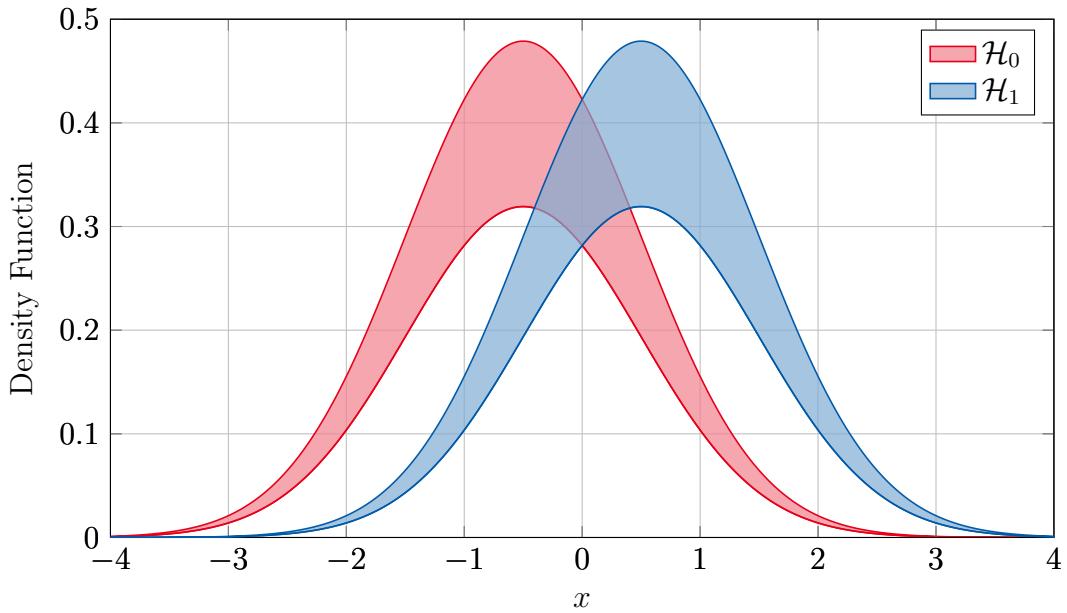


Figure 5.1: Area of feasible densities under \mathcal{H}_0 and \mathcal{H}_1 .

to relax the exact densities to corridors of feasible densities. The uncertainty under each hypothesis is modeled as

$$\mathcal{P}^i = \{ P \in \mathcal{M}_\mu : 0.8p_i \leq p \leq 1.2p_i \}, \quad i \in \{0, 1\}$$

for all X_n , where p_0 and p_1 denote the densities corresponding to $\mathcal{N}(-0.5, 1)$ and $\mathcal{N}(0.5, 1)$, respectively. The two corridors of feasible density functions are shown in Figure 5.1. The corresponding composite hypotheses are given by

$$\begin{aligned} \mathcal{H}_0: \quad & \mathcal{P}_{X_n} \in \mathcal{P}^0, \\ \mathcal{H}_1: \quad & \mathcal{P}_{X_n} \in \mathcal{P}^1. \end{aligned} \tag{5.32}$$

The maximum expected run-length is minimized over the uncertainty set

$$\mathcal{P} = \{ P \in \mathcal{M}_\mu : 0.8p_0 \leq p \}.$$

This choice guarantees that the expected run-length under \mathcal{H}_0

- is upper bounded by the expected run-length under the least favorable family of distributions Q_z since $\mathcal{P}^0 \subset \mathcal{P}$.
- will not exceed the expected run-length under Q_z , even if the observations are contaminated by up to 20% outliers.

This choice of uncertainty sets is suitable for applications where there is little uncertainty about the distributions under either hypothesis, but at the same time a decision needs to be taken reliably and quickly, even if a fraction of the observations is lost or corrupted.

Numerical results are presented for target error probabilities $\alpha = \beta = 0.05$. In order to perform the calculations, the log-likelihood plane was discretized on $[-5, 2] \times [-5, 2]$, with step sizes $\Delta \log z_0 = \Delta \log z_1 = 0.1$, and the iterative design procedure outlined in the previous section was used. The cost function ρ_λ was first calculated by means of the linear programming approach detailed in Section 3.4.2. The result was then used as an initial guess for Broyden's methods, which was applied to calculate ρ_λ on a finer grid with step sizes $\Delta \log z_0 = \Delta \log z_1 = 0.05$. The conditional distributions P_z^0 , P_z^1 and P_z were discretized on the interval $[-6, 6]$ with step sizes $\Delta x = 0.05$. In order to better trace the rather sharp peaks of the least favorable densities an even finer discretization would have been preferable in some cases, but in general the obtained densities are sufficiently accurate. The least favorable distributions of the fixed sample size test, see [FZ15b], were used as initial guesses for the least favorable densities of the sequential test under \mathcal{H}_0 and \mathcal{H}_1 . The projection of $\mathcal{N}(0, 1)$ on \mathcal{P} was used as an initial guess for the least favorable densities with respect to the expected run-length, i.e,

$$q(x) = \max\{cp(x), p_0(x)\},$$

where p denotes the density corresponding to $\mathcal{N}(0, 1)$ and c needs to be chosen such that q is a valid density.

The iterative design of ρ_λ and (Q_z^0, Q_z^1, Q_z) was stopped after three iterations, when the maximum difference between the second and third iterate of the cost function was of the order of 10^{-3} . The last iterate of ρ_λ is depicted in Figure 5.2.

It can be seen that the cost function is similar in shape to the first function in Figure 3.1, which corresponds to an optimal test for the nominal hypothesis under $P = P_0$. The minimax test in this section is designed under the relaxed null hypothesis $\mathcal{P} \supset \mathcal{P}^0$, which explains the similarity. The artifacts that can be seen along the left “mountain-side” are due to numerical inaccuracies at the line where g_λ and d_λ intersect. A finer grid needs to be used in order to obtain more accurate results in this area.

The boundary of the stopping region is depicted in Figure 5.3. Again, some concessions have to be made in terms of numerical accuracy, but most parts of the boundary are reasonably smooth. In general, the boundary admits a higher curvature compared to the non-robust tests in Section 3.5. An intuitive explanation for this effect is given later in this section.

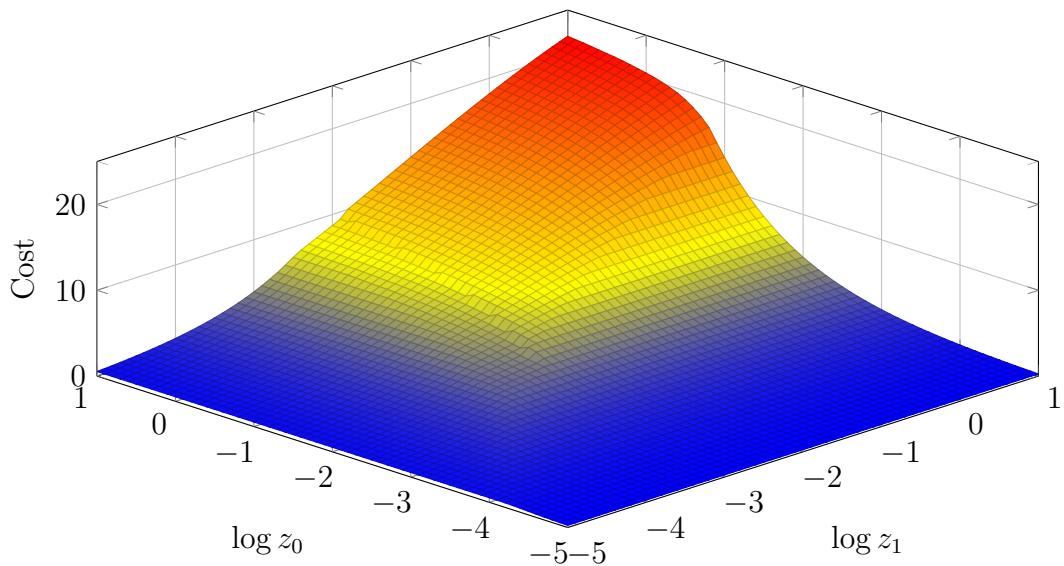


Figure 5.2: Approximate cost function $\rho_\lambda(z)$ of a minimax test for the hypotheses in (5.32) with target error probabilities $\alpha = \beta = 0.05$.

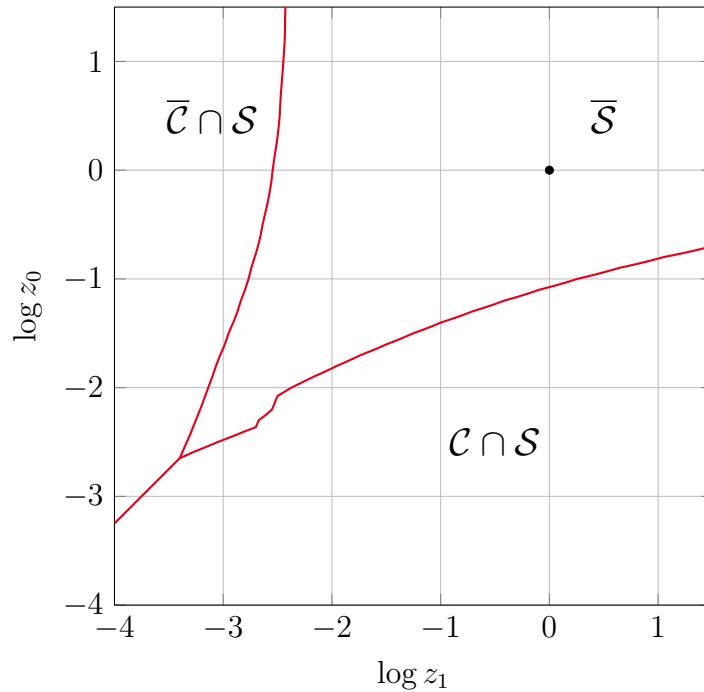


Figure 5.3: Approximate boundaries of the stopping region of the minimax sequential test for the hypotheses in (5.32) with error probabilities $\alpha = \beta = 0.05$.

Examples for least favorable densities corresponding to different states of the test statistic are depicted on the left-hand side of Figure 5.4. On the right-hand side, the log-likelihood ratios of the respective densities are plotted, i.e., the test statistics used by the minimax test to update z_0 and z_1 .

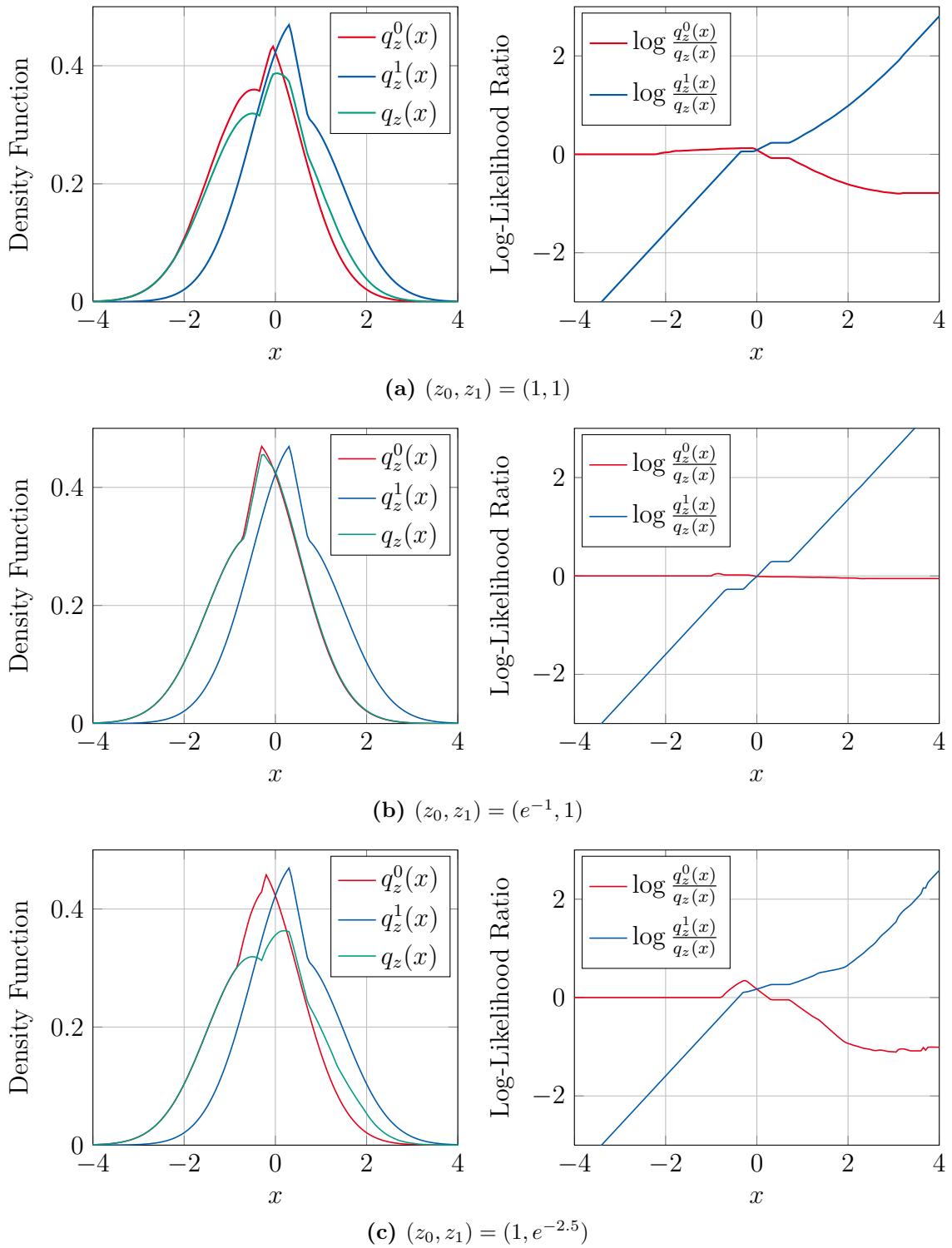


Figure 5.4: Examples for least favorable densities (left) and corresponding test statistics (right) of a minimax sequential test in different states of the test statistic.

Interestingly, q_z^1 in this example turns out to be independent of the test statistic, while the shapes of q_z^0 and q_z change for different values of z . In the initial state,

$z = (1, 1)$, the least favorable densities are chosen such that $q_{(1,1)}^0$ is close to $q_{(1,1)}^1$ and $q_{(1,1)}$ is located between $q_{(1,1)}^0$ and $q_{(1,1)}^1$. This choice is reasonable, given that the least favorable densities are chosen in order to make testing \mathcal{H}_0 against \mathcal{H}_1 as difficult as possible. More insight can be gained by inspecting states that are closer to the stopping region. Consider the state $z = (e^{-1}, 1)$, which is very close to the part of the stopping region that leads to a decision for \mathcal{H}_1 (compare Figure 5.3). In this state, the least favorable density with respect to the expected run-length is chosen almost identical to the least favorable density with respect to the type I error probability, i.e., $q_{(e^{-1},0)} \approx q_{(e^{-1},0)}^0$. The reason for this is that in order to prolong the test, $q_{(e^{-1},0)}$ needs to be chosen such that the test statistic is driven away from the stopping region. In the given state, where a decision for \mathcal{H}_1 is imminent, this means generating samples according to \mathcal{H}_0 . The opposite is the case in the state $z = (1, e^{-2.5})$, which is close to the part of the stopping region that leads to decision for \mathcal{H}_0 . In order to maximize the expected run-length, $q_{(1,e^{-2.5})}$ is now chosen as similar as possible to $q_{(1,e^{-2.5})}^1$. However, since the uncertainty set only allows for 20% outliers, the densities do not overlap completely and only a certain degree of similarity can be realized.

These considerations allow for an intuitive interpretation of the shape of the stopping region. Since the virtual adversary of the test designer has to stick to densities within the uncertainty sets, the test statistic cannot be pushed away from the \mathcal{H}_1 -boundary as effectively as it can be pushed away from the \mathcal{H}_0 -boundary. Consequently, the test designer can relax the former and has to tighten the latter. This is reflected in the asymmetry that can clearly be seen in Figure 5.3. The boundary corresponding to a decision for \mathcal{H}_1 is shifted much more towards negative log-likelihood ratios than the boundary corresponding to a decision for \mathcal{H}_0 . This also highlights an important difference between state-dependent minimax solutions and state-independent or asymptotic minimax solutions. While the latter result in a *minimum drift towards the thresholds* (compare the zero-drift example in Section 3.5.1), the former result in a *maximum drift away from the thresholds*.

An intuition for the testing strategy that is used to counteract this effect can be gained from the test statistics depicted in Figure 5.4. It can be seen that the most significant component of the test statistic is z_1 , which is allowed to perform rather big steps with every new observation. The z_0 component, on the other hand, changes in much smaller steps. Therefore, its role in the test can be interpreted as a kind of “correction term” that determines the thresholds for z_1 , but does not contribute significantly to the decision itself. This strategy is in line with the assumption that the distribution under \mathcal{H}_1 is subject to uncertainty, but not contaminated with outliers. Accordingly, the z_1 component can be trusted more, whereas the z_0 component might be corrupted and is clipped as a precaution.

5.9 Summary

In this chapter, the design of minimax robust sequential hypothesis tests was discussed and sufficient conditions were derived for distributions to be least favorable with respect to the expected run-length and error probabilities of a sequential test. Combining results on optimal sequential tests and least favorable distributions, a sufficient condition for a sequential test to be minimax optimal under general distributional uncertainties was derived. The cost function of the minimax optimal test was further identified as a convex statistical similarity measure and the least favorable distributions as the distributions that are most similar with respect to this measure. In order to obtain more specific results, the density band model was introduced as an example for a nonparametric uncertainty model. The corresponding least favorable distributions were given in an implicit form, based on which a simple algorithm for their numerical calculation was derived. Finally, the minimax robust sequential test under density band uncertainties was detailed and shown to admit the characteristic minimax property of a maximally flat performance profile over its state space. A numerical example for a minimax optimal sequential test concluded the section.

Chapter 6

Conclusions and Outlook

In this dissertation, a framework for the design, analysis and implementation of optimal and minimax robust sequential hypothesis tests was developed. A close connection between the error probabilities of optimal sequential tests and the partial derivatives of the cost function of the corresponding optimal stopping problem was shown. Based on this result, efficient numerical techniques for the design of optimal sequential tests were derived. For the design of minimax robust tests, the general case of convex uncertainty sets and the particular case of the density band model were discussed. For the latter, a simple iterative algorithm for the calculation of the least favorable distributions was proposed. It was further shown that the least favorable distributions are defined as being most similar with respect to a similarity measure induced by the cost function of the corresponding sequential test. The results complement and extend the state-of-the-art in robust sequential hypothesis testing, which is mostly focused on asymptotically or approximately optimal procedures and parametric uncertainty models.

Possible extensions of the work and open problems are listed below. Some of them are left open, some are the subject of ongoing research.

6.1 Tests for Multiple Hypotheses

Although this topic has not been touched upon in the body of the dissertation, the presented approach to the design of sequential tests can be extended to multiple hypotheses in a rather straightforward manner. Optimal sequential test for multiple hypotheses were derived in [Nov08] and the results in Chapter 3 can be adapted accordingly. More precisely, it can be shown that the optimal sequential test for \mathcal{H}_0 against K alternatives $\mathcal{H}_1, \dots, \mathcal{H}_K$ is characterized by a cost function $\rho_\lambda: \mathbb{R}_+^{K+1} \times \Omega_\theta \rightarrow \mathbb{R}_+$, where

$$\rho_\lambda = \rho_\lambda(z_0, \dots, z_K, \theta)$$

and $\lambda \in \mathbb{R}_+^{K+1}$. For the design of optimal sequential tests for multiple hypotheses, an integral equation of the form (3.12) needs to be solved. For the minimax test, the least favorable distributions are defined by a $K + 1$ -dimensional similarity measure. The idea of this extension should be clear without going into technical details. A systematic presentation of the results for multi-hypotheses tests is work in progress.

6.2 Asymptotic Results

Another subject that warrants further investigation is the connection between the optimal results in this work and the asymptotic results in, for example, [BD08b]. By intuition, it should be possible to derive the asymptotic solutions from the optimal solutions. In case of optimal tests for i.i.d. processes, this derivation is indeed straightforward. Letting the error probabilities go to zero, i.e.,

$$\max\{\alpha, \beta\} \rightarrow 0,$$

implies that the probability of stopping in any state (z, θ) goes to zero, i.e.,

$$H_z(\mathcal{S}) \rightarrow 0 \quad \forall z \in \mathbb{R}_+^2.$$

This implies that asymptotically

$$\rho_\lambda(z) = 1 + \int \rho_\lambda \left(z_0 \frac{p_0(x)}{p(x)}, z_1 \frac{p_1(x)}{p(x)} \right) dP(x) \quad (6.1)$$

since stopping the test at any finite time instant is no longer a valid option. It is not hard to show that this equation is solved by any function of the form

$$\rho_\lambda(z) = a_0 \log \lambda_0 z_0 + a_1 \log \lambda_1 z_1, \quad (6.2)$$

where a_0 and a_1 have to be chosen such that

$$a_0 D_{\text{KL}}(P_0 \| P) + a_1 D_{\text{KL}}(P_1 \| P) = 1$$

and D_{KL} denotes the Kullback-Leibler divergence. Minimizing the cost of the asymptotic test over a_0 and a_1 yields

$$\rho_\lambda^*(1) = \min \left\{ \frac{\log \lambda_0}{D_{\text{KL}}(P_0 \| P)}, \frac{\log \lambda_1}{D_{\text{KL}}(P_1 \| P)} \right\} \quad (6.3)$$

The minimum expression in (6.3) is in line with the results in the literature. The interpretation of (6.3) is that the stopping region is asymptotically determined by two separate thresholds for z_0 and z_1 and the expected run-length is determined by the one that is reached first, on average.

In general, the similarity measure induced by ρ_λ converges towards the Kullback-Leibler divergence in the asymptotic i.i.d. case. However, a systematic analysis of this convergence and how it extends to dependent data and the state dependent minimax case still needs to be carried out.

6.3 Investigation of Special Cases

The results in this work are intentionally based on rather weak and general assumptions about the stochastic process X and the nature of the two hypotheses. The question whether analytic results for least favorable distributions can be obtained by introducing stricter assumptions is certainly worth investigating. The same holds for the question whether collapsing one or two of the uncertainty sets to singletons yields significant simplifications. In general, introducing more and stronger assumptions seems to be necessary in order to obtain minimax procedures that do not require extensive computations for their design and implementation.

6.4 Existence of Minimax Optimal Tests

The sufficient conditions given in Chapter 5 can be used for the design of minimax sequential tests in the sense that the solution of a system of integral equations determines the optimal test. However, the question under which conditions and uncertainty models this solution exists and is well defined requires further investigation.

6.5 Comparison to Alternative Procedures

The numerical results given in this dissertation are of limited scope. A more detailed analysis of how large the performance differences are between strictly minimax and asymptotically minimax tests, or state-dependent and state-independent tests would be useful. In addition, the minimax sequential test should be compared in more detail to the minimax fixed sample size test and tests based on different design approaches.

A comparison between sequential and fixed sample size minimax tests is currently being worked on. In fact, it can be shown that for $\mathcal{P} = \mathcal{M}_\mu$, i.e., the true distribution can be any distribution, the minimax sequential test reduces to the minimax fixed sample size test. However, since the fixed sample size test is not time-invariant and, therefore, not covered by the framework in Chapter 5, a formal proof is beyond the scope of this dissertation.

Appendix

A.1 Recursive Definition of Performance Measures

By definition of α_π^n , δ and ψ , it holds that

$$\begin{aligned}\alpha_\pi^n(t) &= E_{P_0}[\delta_\tau \mid T^n = t, \tau \geq n] \\ &= \psi_n(t)E_{P_0}[\delta_\tau \mid T^n = t, \tau = n] + (1 - \psi_n(t))E_{P_0}[\delta_\tau \mid T^n = t, \tau > n] \\ &= \psi_n(t)\delta_n(t) + (1 - \psi_n(t))E_{P_0}[\delta_\tau \mid T^n = t, \tau > n].\end{aligned}\quad (4)$$

Using property (2.24) and the fact that $\{\tau > n\} = \{\tau \geq n+1\}$, it further follows that

$$\begin{aligned}E_{P_0}[\delta_\tau \mid T^n = t, \tau > n] &= \int_{\Omega_T} E_{P_0}[\delta_\tau \mid T^{n+1} = t', \tau > n] dP_0(T^{n+1} = t' \mid T^n = t, \tau > n) \\ &= \int_{\Omega_T} E_{P_0}[\delta_\tau \mid T^{n+1} = t', \tau \geq n+1] dP_0(T^{n+1} = t' \mid T^n = t) \\ &= E_{P_0}[\alpha_\pi^{n+1}(T^{n+1}) \mid T^n = t]\end{aligned}$$

which yields (2.31) when inserted back into (4). The recursion for β_π^n can be shown analogously.

For γ_π^n it holds that

$$\begin{aligned}\gamma_\pi^n(t) &= E_P[\tau - n \mid T^n = t, \tau \geq n] \\ &= \psi_n(t)E_P[\tau - n \mid T^n = t, \tau = n] + (1 - \psi_n(t))E_P[n - \tau \mid T^n = t, \tau > n] \\ &= (1 - \psi_n(t))E_P[\tau - n \mid T^n = t, \tau > n].\end{aligned}\quad (5)$$

Again, using property (2.24), it follows that

$$\begin{aligned}E_P[\tau - n \mid T^n = t, \tau > n] &= \int_{\Omega_T} E_P[\tau - n \mid T^{n+1} = t', \tau > n] dP(T^{n+1} = t' \mid T^n = t, \tau > n) \\ &= \int_{\Omega_T} E_P[\tau - n \mid T^{n+1} = t', \tau \geq n+1] dP(T^{n+1} = t' \mid T^n = t) \\ &= 1 + E_P[\gamma_\pi^{n+1}(T^{n+1}) \mid T^n = t]\end{aligned}$$

which yields (2.31) when inserted back into (5).

A.2 Proof of Theorem 3

Let $\pi^* = (\psi^*, \delta^*)$ be as defined in Theorem 3, i.e., π^* solves (3.4) and it holds that

$$E_{P_0}[\delta_\tau^*] = \alpha \quad \text{and} \quad E_{P_1}[1 - \delta_\tau^*] = \beta$$

What needs to be shown is that

$$\min_{(\psi, \delta) \in \Pi_{\alpha, \beta}} E_P[\tau(\psi)] = E_P[\tau(\psi^*)],$$

where $\Pi_{\alpha, \beta}$ denotes the set of all testing policies with type I and type II error probabilities smaller than α and β , respectively. A proof can be given by contradiction. Assume that a policy $\pi^\diamond = (\psi^\diamond, \delta^\diamond)$ exists such that

$$E_P[\tau(\psi^\diamond)] < E_P[\tau(\psi^*)]$$

and

$$E_{P_0}[\delta_\tau^\diamond] \leq \alpha, \quad E_{P_1}[1 - \delta_\tau^\diamond] \leq \beta.$$

This, however, implies that

$$\begin{aligned} E_P[\tau(\psi^*)] + \lambda_0 E_{P_0}[\delta_\tau^*] + \lambda_1 E_{P_1}[1 - \delta_\tau^*] &= E_P[\tau(\psi^*)] + \lambda_0 \alpha + \lambda_1 \beta \\ &> E_P[\tau(\psi^\diamond)] + \lambda_0 E_{P_0}[\delta_\tau^\diamond] + \lambda_1 E_{P_1}[1 - \delta_\tau^\diamond], \end{aligned}$$

which contradicts the assumption that π^* solves (3.4).

A.3 Proof of Theorem 5

Theorem 5 is essentially a corollary of the optimal stopping results in Theorem 1 and Theorem 2, with cost functions J_n chosen according to (3.9).

First, the finite-horizon case is considered. Let N be the finite horizon of the truncated version of problem (3.4), i.e.,

$$\min_{\psi \in \Delta^N} E_P[n + g_\lambda(z)],$$

where the optimization is performed over $(\psi_n)_{0 \leq n \leq N-1}$ and $\psi_N = 1$. Via induction, it can be shown that $J_{n,N}^*(x_1, \dots, x_n)$ in Theorem 1 is of the form

$$J_{n,N}^*(x_1, \dots, x_n) = n + \rho_\lambda^{n,N}(z^n, \theta_n).$$

The inductive step is given by

$$\begin{aligned}
J_{n-1,N}^*(x_1, \dots, x_{n-1}) &= \min\{J_{n-1}(x_1, \dots, x_{n-1}), V_{n-1,N}(x_1, \dots, x_n)\} \\
&= \min\left\{(n-1) + g_\lambda(z^{n-1}), E_P\left[n + \rho_\lambda^{n,N}(z^n, \theta_n) \mid x_1, \dots, x_{n-1}\right]\right\} \\
&= \min\left\{(n-1) + g_\lambda(z^{n-1}), n + E_P\left[\rho_\lambda^{n,N}(z^n, \theta_n) \mid z^{n-1}, \theta_{n-1}\right]\right\} \\
&= (n-1) + \min\left\{g_\lambda(z^{n-1}), 1 + E_P\left[\rho_\lambda^{n,N}(z^n, \theta_n) \mid z^{n-1}, \theta_{n-1}\right]\right\} \\
&= (n-1) + \rho_\lambda^{n-1,N}(z^{n-1}, \theta_{n-1}),
\end{aligned}$$

where

$$\rho_\lambda^{n-1,N}(z^{n-1}, \theta_{n-1}) := \min\left\{g_\lambda(z^{n-1}), 1 + E_P\left[\rho_\lambda^{n,N}(z^n, \theta_n) \mid z^{n-1}, \theta_{n-1}\right]\right\}$$

and

$$\begin{aligned}
E_P\left[\rho_\lambda^{n,N}(z^n, \theta_n) \mid z^{n-1}, \theta_{n-1}\right] \\
&= \int \rho_\lambda^{n,N}\left(z_0^{n-1} \frac{p_{\theta_{n-1}}^0(x)}{p_{\theta_{n-1}}(x)}, z_1^{n-1} \frac{p_{\theta_{n-1}}^1(x)}{p_{\theta_{n-1}}(x)}, \xi_{\theta_{n-1}}(x)\right) dP_{\theta_{n-1}}(x).
\end{aligned}$$

The induction basis is given by

$$J_{N,N}^*(x_1, \dots, x_N) = J_N(x_1, \dots, x_N) = N + g_\lambda(z^N, \theta_N)$$

so that $\rho_\lambda^{N,N} = g_\lambda$. Hence, the minimum cost of the N -truncated test is

$$J_{0,N}^* = \rho_\lambda^{0,N}(z^0, \theta_0) = \rho_\lambda^{0,N}(1, 1, \theta_0)$$

for all $N \geq 1$.

To determine the limit $\rho_\lambda^{n,\infty} = \lim_{N \rightarrow \infty} \rho_\lambda^{n,N}$, note that $\rho_\lambda^{n,N}$ is obtained from $\rho_\lambda^{n+1,N}$ by applying the transformation

$$\mathcal{T}\{\rho(z, \theta)\} = \min\left\{g_\lambda(z), 1 + \int \rho\left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x)\right) dP_\theta(x)\right\}, \quad (6)$$

which is monotonic in ρ . Since $\rho_\lambda^{N,N} = g_\lambda$, independent of N , $\rho_\lambda^{n,N}$ can be expressed as

$$\rho_\lambda^{n,N} = \mathcal{T}^{N-n}\{g_\lambda\},$$

where \mathcal{T}^n denotes an n times repeated application of \mathcal{T} . The transition to the infinite-horizon case yields

$$\lim_{N \rightarrow \infty} \rho_\lambda^{n,N} = \lim_{N \rightarrow \infty} \mathcal{T}^{N-n}\{g_\lambda\} = \lim_{N \rightarrow \infty} \mathcal{T}^N\{g_\lambda\} =: \rho_\lambda \quad (7)$$

for all $n \in \mathbb{N}$. The next step is to show that $\lim_{N \rightarrow \infty} \mathcal{T}^N\{g_\lambda\}$ exists and is unique. Since every bounded monotonic sequence converges, [Rud76, Theorem 3.14], and the limit of converging sequences in metric spaces is unique, [Muk05, Theorem 3.1.4.] (compare also [Nov09]), it suffices to show that $\mathcal{T}^n\{g_\lambda\} \geq 0$ for all $n \geq 1$ and that the sequence $\mathcal{T}^n\{g_\lambda\}$ is monotonically nonincreasing. The fact that $\mathcal{T}^n\{g_\lambda\} \geq 0$ follows directly from $g_\lambda \geq 0$ and the definition of \mathcal{T} . The monotonic property can again be established by induction. Assume that $\mathcal{T}^n\{g_\lambda\} \leq \mathcal{T}^{n-1}\{g_\lambda\}$. By monotonicity of \mathcal{T} it then holds that

$$\mathcal{T}^n\{g_\lambda\} = \mathcal{T}\{\mathcal{T}^{n-1}\{g_\lambda\}\} \leq \mathcal{T}\{\mathcal{T}^n\{g_\lambda\}\} = \mathcal{T}^{n+1}\{g_\lambda\}.$$

The induction basis $\mathcal{T}\{g_\lambda\} \leq g_\lambda$ holds trivially since $\mathcal{T}\{\rho\} \leq g_\lambda$ by definition. The fixed-point solution of

$$\rho_\lambda = \mathcal{T}\{\rho_\lambda\}$$

yields (3.13). This concludes the proof.

A.4 Proof of Corollary 3

The time-invariance property (2.38) of the optimal stopping and decision rules follows directly from the fact that the limit in (7) is independent of the time instant n . It remains to show that the test statistic $T^n = (z^n, \theta_n)$ satisfies (2.24) and (2.39). By definition of θ , it holds that for all $\mathcal{E} \in \mathcal{F}_T$ and $n \in \mathbb{N}_0$

$$\begin{aligned} P_X(T^{n+1} \in \mathcal{E} | (T^0, \dots, T^n)) &= ((z^0, \theta_0), \dots, (z^n, \theta_n)), \tau \geq n \\ &= P_X\left(\left(z_0^n \frac{p_{\theta_n}^0(x_{n+1})}{p_{\theta_n}(x_{n+1})}, z_1^n \frac{p_{\theta_n}^0(x_{n+1})}{p_{\theta_n}(x_{n+1})}, \xi_{\theta_n}(x_{n+1})\right) \in \mathcal{E}\right) \\ &= P_X(T^{n+1} \in \mathcal{E} | T^n = (z^n, \theta_n)) \end{aligned}$$

so that property (2.24) is satisfied. Finally, since for all $m, n \in \mathbb{N}$

$$P_{X_n|\theta} = P_{X_m|\theta},$$

it is the case that

$$\begin{aligned} P_X(T^{n+1} \in \mathcal{E} | T^n = (z, \theta)) &= P_X\left(\left(z_0 \frac{p_{\theta_n}^0(x_{n+1})}{p_{\theta_n}(x_{n+1})}, z_1 \frac{p_{\theta_n}^0(x_{n+1})}{p_{\theta_n}(x_{n+1})}, \xi_{\theta_n}(x_{n+1})\right) \in \mathcal{E}\right) \\ &= P_{X_{n+1}|\theta}\left(\left(z_0 \frac{p_{\theta_n}^0(x_{n+1})}{p_{\theta_n}(x_{n+1})}, z_1 \frac{p_{\theta_n}^0(x_{n+1})}{p_{\theta_n}(x_{n+1})}, \xi_{\theta_n}(x_{n+1})\right) \in \mathcal{E}\right) \\ &= P_{X_{m+1}|\theta}\left(\left(z_0 \frac{p_{\theta_n}^0(x_{m+1})}{p_{\theta_n}(x_{m+1})}, z_1 \frac{p_{\theta_n}^0(x_{m+1})}{p_{\theta_n}(x_{m+1})}, \xi_{\theta_n}(x_{m+1})\right) \in \mathcal{E}\right) \\ &= P_X(T^{m+1} \in \mathcal{E} | T^m = (z, \theta)). \end{aligned}$$

A.5 Proof of Lemma 1

Uniform convergence of monotonic sequences often follows immediately from Dini's theorem [Rud76, Theorem 7.13]. In this case, however, neither is the state space Ω_ρ compact, nor is ρ_λ necessarily continuous in θ . Nevertheless, uniform convergence can still be shown via a detour over almost uniform convergence.

Define the measure

$$H^*(\mathcal{E}) = \sup_{(z,\theta) \in \Omega_\rho} H_{z,\theta}(\mathcal{E}), \quad \mathcal{E} \in \mathcal{F}_\rho.$$

By Theorem 5, the sequence $(\rho_\lambda^n)_{n \in \mathbb{N}_0}$ converges pointwise on Ω_ρ and hence H^* almost everywhere. Egorov's theorem [WWZ77, Theorem 4.17] states that this implies almost uniform convergence with respect to H^* , i.e., for every $\varepsilon > 0$, there exists a set $\mathcal{E}_\varepsilon \in \mathcal{F}_\rho$ such that $H^*(\mathcal{E}_\varepsilon) < \varepsilon$ and $(\rho_\lambda^n)_{n \geq 0}$ converges uniformly on $\Omega_\rho \setminus \mathcal{E}_\varepsilon$. In the following it is shown that for ρ_λ almost uniform convergence implies uniform convergence.

Since $(\rho_\lambda^n)_{n \geq 0}$ is monotonically nonincreasing, ρ_λ^n can be written as $\rho_\lambda^n = \rho_\lambda + \Delta\rho_\lambda^n$ for every $n \in \mathbb{N}_0$, where $(\Delta\rho_\lambda^n)_{n \geq 0}$ is a nonincreasing sequence of nonnegative functions. In order to guarantee uniform convergence it suffices to show that

$$\lim_{n \rightarrow \infty} \sup_{\Omega_\rho} \Delta\rho_\lambda^n = 0.$$

By definition of ρ_λ^n , it holds that

$$\begin{aligned} \sup_{\Omega_\rho} \Delta\rho_\lambda^n &= \sup_{(z,\theta) \in \Omega_\rho} \left\{ \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda^{n-1} dH_{z,\theta} \right\} - \rho_\lambda(z, \theta) \right\} \\ &\leq \sup_{(z,\theta) \in \Omega_\rho} \left\{ \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda dH_{z,\theta} \right\} - \rho_\lambda + \int \Delta\rho_\lambda^{n-1} dH_{z,\theta} \right\} \\ &= \sup_{(z,\theta) \in \Omega_\rho} \left\{ \int \Delta\rho_\lambda^{n-1} dH_{z,\theta} \right\}. \end{aligned}$$

With \mathcal{E}_ε defined as above, it further follows that

$$\begin{aligned} \sup_{\Omega_\rho} \Delta\rho_\lambda^n &\leq \sup_{(z,\theta) \in \Omega_\rho} \left\{ \int \Delta\rho_\lambda^{n-1} dH_{z,\theta} \right\} \\ &\leq \int_{\Omega_\rho \setminus \mathcal{E}_\varepsilon} \sup_{\Omega_\rho \setminus \mathcal{E}_\varepsilon} \Delta\rho_\lambda^{n-1} dH^* + \int_{\mathcal{E}_\varepsilon} \sup_{\mathcal{E}_\varepsilon} \Delta\rho_\lambda^{n-1} dH^* \\ &< \sup_{\Omega_\rho \setminus \mathcal{E}_\varepsilon} \Delta\rho_\lambda^{n-1} + \varepsilon \sup_{\Omega_\rho} \Delta\rho_\lambda^{n-1}. \end{aligned}$$

Since

$$\lim_{n \rightarrow \infty} \sup_{\Omega_\rho \setminus \mathcal{E}_\varepsilon} \Delta\rho_\lambda^n = 0,$$

the sequence $(\sup_{\Omega_\rho} \Delta \rho_\lambda^n)_{n \in \mathbb{N}_0}$ converges to zero for every $\varepsilon < 1$, given that $\sup_{\Omega_\rho} \Delta \rho_\lambda^n$ is bounded for some n . The latter is guaranteed by the pointwise convergence of ρ_λ^n on Ω_ρ .

A.6 Proof of Theorem 6

The proof of Theorem 6 is given by showing that for $\pi \in \Pi_\lambda^*$ the function $\gamma_\pi + \lambda_0 z_0 \alpha_\pi + \lambda_1 z_1 \beta_\pi$ solves the integral equation that defines ρ_λ . Since ρ_λ is unique, this implies that both functions are identical.

First, note that for $\pi = (\psi, \delta) \in \Pi_\lambda^*$ it holds that

$$\begin{aligned} \rho_\lambda(z, \theta) &= \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta(x) \right\} \\ &= \psi(z, \theta) g_\lambda(z) + (1 - \psi(z, \theta)) \left(1 + \int \rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta(x) \right) \\ &= \lambda_0 z_0 \psi(z, \theta) \delta(z) + \lambda_1 z_1 \psi(z, \theta) (1 - \delta(z)) \\ &\quad + (1 - \psi(z, \theta)) \left(1 + \int \rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta(x) \right). \end{aligned}$$

Using $H_{z, \theta}$ as defined in (3.19) and omitting explicit function arguments, this equation can be written more compactly as

$$\rho_\lambda(z, \theta) = \lambda_0 z_0 \psi \delta + \lambda_1 z_1 \psi (1 - \delta) + (1 - \psi) \left(1 + \int \rho_\lambda dH_{z, \theta} \right). \quad (8)$$

In order to show that

$$\rho_\lambda(z, \theta) = \gamma_\pi + \lambda_0 z_0 \alpha_\pi(z, \theta) + \lambda_1 z_1 \beta_\pi(z, \theta), \quad (9)$$

it suffices to show that $\gamma_\pi + \lambda_0 z_0 \alpha_\pi + \lambda_1 z_1 \beta_\pi$ solves (8). Inserting (9) into (8) yields

$$\begin{aligned} \gamma_\pi + \lambda_0 z_0 \alpha_\pi + \lambda_1 z_1 \beta_\pi &= \lambda_0 z_0 \psi \delta + \lambda_1 z_1 \psi (1 - \delta) \\ &\quad + (1 - \psi) \left(1 + \int \gamma_\lambda dH_{z, \theta} + \lambda_0 z_0 \int \alpha_\lambda dH_{z, \theta}^0 + \lambda_1 z_1 \int \beta_\lambda dH_{z, \theta}^1 \right), \quad (10) \end{aligned}$$

where the changes in measure on the right hand side follow from the fact that for every function of the form $\tilde{f}(z, \theta) = z_i f(z, \theta)$, $i \in \{0, 1\}$, it holds that

$$\begin{aligned} \int \tilde{f} dH_{z, \theta} &= \int z_i \frac{p_\theta^i(x)}{p_\theta(x)} f \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta(x) \\ &= \int z_i f \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta^i(x) \\ &= z_i \int f dH_{z, \theta}^i. \end{aligned}$$

Given the test statistic $T^n = (z^n, \theta_n)$, it further holds that

$$E_{P_0} [\alpha_\pi(T^1) \mid T^0 = (z, \theta)] = \int \alpha_\pi \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta^0(x)$$

$$E_{P_1} [\beta_\pi(T^1) \mid T^0 = (z, \theta)] = \int \beta_\pi \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta^1(x)$$

$$E_P [\gamma_\pi(T^1) \mid T^0 = (z, \theta)] = \int \gamma_\pi \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^1(x)}{p_\theta(x)}, \xi_\theta(x) \right) dP_\theta(x)$$

so that the integral equations (2.41)–(2.43) can be written as

$$\alpha_\pi = \psi\delta + (1 - \psi) \int \alpha_\pi dH_{z,\theta}^0, \quad (11)$$

$$\beta_\pi = \psi(1 - \delta) + (1 - \psi) \int \beta_\pi dH_{z,\theta}^1, \quad (12)$$

$$\gamma_\pi = (1 - \psi) \left(1 + \int \gamma_\pi dH_{z,\theta} \right). \quad (13)$$

Multiplying (11) by $\lambda_0 z_0$, (12) by $\lambda_1 z_1$ and adding (11)–(13) yields (10), which concludes the proof.

A.7 Proof of Theorem 7

Both the non-decreasing property and concavity of ρ_λ can be shown by induction. Here only the concavity of ρ_λ is proven in detail. The non-decreasing property can be shown analogously. Assume that

$$\rho_\lambda^n(\kappa z' + (1 - \kappa)z, \theta) \geq \kappa \rho_\lambda^n(z', \theta) + (1 - \kappa) \rho_\lambda^n(z, \theta)$$

holds for some element ρ_λ^n of the sequence $(\rho_\lambda^n)_{n \in \mathbb{N}_0}$ defined in Lemma 1. Since the minimum function and, therefore, $g_\lambda(z)$ are concave as well, this implies

$$\begin{aligned} \rho_\lambda^{n+1}(\kappa z' + (1 - \kappa)z, \theta) &= \min \left\{ g_\lambda(\kappa z' + (1 - \kappa)z), 1 + \int \rho_\lambda^n dH_{\kappa z' + (1 - \kappa)z, \theta} \right\} \\ &\geq \min \left\{ \kappa g_\lambda(z') + (1 - \kappa) g_\lambda(z), \right. \\ &\quad \left. 1 + \kappa \int \rho_\lambda^n dH_{z', \theta} + (1 - \kappa) \int \rho_\lambda^n dH_{z, \theta} \right\} \\ &\geq \kappa \min \left\{ g_\lambda(z'), 1 + \int \rho_\lambda^n dH_{z', \theta} \right\} \\ &\quad + (1 - \kappa) \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda^n dH_{z, \theta} \right\} \\ &= \kappa \rho_\lambda^{n+1}(z', \theta) + (1 - \kappa) \rho_\lambda^{n+1}(z, \theta). \end{aligned}$$

The induction basis is $g_\lambda(z)$. Uniform convergence of $(\rho_\lambda^n)_{n \geq 1}$ suffices to guarantee that ρ_λ is concave as well.

A.8 Proof of Theorem 8

Since ρ_λ is concave, its generalized partial derivatives exist. Moreover, for $\pi \in \Pi_\lambda^*$, ρ_λ can be written as

$$\rho_\lambda = \min\{g_\lambda, d_\lambda\} = \psi(\lambda_0 z_0 \delta + \lambda_1 z_1 (1 - \delta)) + (1 - \psi)d_\lambda \quad (14)$$

where d_λ is defined in (3.17). Exploiting the coupling between ρ_λ and d_λ is key to the proof of Theorem 8. The argument used here is based on a generalized version of Leibniz's integral rule, which is given in the next lemma.

Lemma 3 (Generalized Leibniz integral rule) *Let (Ω, \mathcal{F}) be a measurable space and $f: \mathbb{R}^K \times \Omega \rightarrow \mathbb{R}$ be a convex/concave function. If $f(y, \omega)$ is μ -integrable for all $y \in \mathbb{R}^K$, it holds that*

$$\partial_{y_k} \left(\int_{\Omega} f(y, \omega) \, d\mu(\omega) \right) = \int_{\Omega} \partial_{y_k} f(y, \omega) \, d\mu(\omega),$$

where the integral on the right hand side is a short-hand notation for the set of integrals over all feasible partial derivatives of f , i.e.,

$$\int_{\Omega} \partial_{y_k} f(y, \omega) \, d\mu(\omega) := \left\{ \int_{\Omega} f_{y_k}(y, \omega) \, d\mu(\omega) : f_{y_k} \in \partial_{y_k} f \right\}.$$

The generalized Leibniz integral rule is stated and proven in [Roc74, Theorem 23]. Extensions and variations are given in [Pap97] and [Chi09], to name just two.

Since $\rho_\lambda \leq g_\lambda$, it follows that

$$\int \rho_\lambda \, dH_{z,\theta} \leq \int g_\lambda \, dH_{z,\theta} \leq \max\{\lambda_0 z_0, \lambda_1 z_1\} < \infty$$

so that ρ_λ is $H_{z,\theta}$ -integrable for all $(z, \theta) \in \Omega_\rho$. Hence, Leibniz's integral rule applies to $d_\lambda(z, \theta)$ so that

$$\begin{aligned}\partial_{z_i} d_\lambda(z, \theta) &= \partial_{z_i} \left(\int \rho_\lambda \, dH_{z,\theta} \right) \\ &= \int \partial_{z_i} \left(\rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^i(x)}{p_\theta(x)}, \xi_\theta(x) \right) \right) \, dP_\theta(x) \\ &= \int \frac{p_\theta^0(x)}{p_\theta(x)} \partial_{z_i} \rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^i(x)}{p_\theta(x)}, \xi_\theta(x) \right) \, dP_\theta(x) \\ &= \int \partial_{z_i} \rho_\lambda \left(z_0 \frac{p_\theta^0(x)}{p_\theta(x)}, z_1 \frac{p_\theta^i(x)}{p_\theta(x)}, \xi_\theta(x) \right) \, dP_\theta^i(x) \\ &= \int \partial_{z_i} \rho_\lambda \, dH_{z,\theta}^i\end{aligned}$$

Consequently, $\partial_{z_0} \rho_\lambda$ and $\partial_{z_1} \rho_\lambda$ satisfy the set-valued integral equations [BKTG15]

$$\partial_{z_0} \rho_\lambda(z, \theta) = \lambda_0 \psi \delta + (1 - \psi) \int \partial_{z_0} \rho_\lambda \, dH_{z,\theta}^0, \quad (15)$$

$$\partial_{z_1} \rho_\lambda(z, \theta) = \lambda_1 \psi (1 - \delta) + (1 - \psi) \int \partial_{z_1} \rho_\lambda \, dH_{z,\theta}^1. \quad (16)$$

Read from “right to left”, (15) and (16) state that inserting a generalized derivative of ρ_λ into the integral on the right hand side yields another valid generalized derivative on the left hand side. Read from “left to right”, (15) and (16) state that given any $r_i \in \partial_{z_i} \rho_\lambda$ on the left hand side, a function $r'_i \in \partial_{z_i} \rho_\lambda$ exists such that the right hand side evaluates to r_i .

The above characterization of the generalized differentials, which follows solely from the concavity and integrability of ρ_λ , already implies both statements in Theorem 8.

By inspection of (2.41) and (2.42) it can be seen that $\lambda_0 \alpha_\pi$ and $\lambda_1 \beta_\pi$ are solutions of (15) and (16) for all $\pi \in \Pi_\lambda^*$. This yields the first statement in Theorem 8.

The second part of Theorem 8 is proven by showing that the sets in the statement are subsets of each other, which implies identity. The details are only given for the partial differential with respect to z_0 since the proof for the partial differential with respect to z_1 follows analogously. From the above results it is clear that $\alpha_\pi \in \partial_{z_0} \rho_\lambda / \lambda_0$ for all $\pi \in \Pi_\lambda^*$. By definition, this implies that $\alpha_\pi(z, \theta) \in \partial_{z_0} \rho_\lambda(z, \theta)$ for all $(z, \theta) \in \Omega_\rho$ and all $\pi \in \Pi_\lambda^*$, i.e.,

$$\{\alpha_\pi(z, \theta) : \pi \in \Pi_\lambda^*\} \subset \partial_{z_0} \frac{\rho_\lambda}{\lambda_0}(z, \theta).$$

The converse is shown in two steps. The first step is to show that given two policies $\pi, \pi' \in \Pi_\lambda^*$ with

$$\alpha_\pi(z^*, \theta^*) = \alpha \quad \text{and} \quad \alpha_{\pi'}(z^*, \theta^*) = \alpha'$$

for some $(z^*, \theta^*) \in \Omega_\rho$, it holds that for every $\tilde{\alpha} \in \text{co}\{\alpha, \alpha'\}$ there exists a policy $\tilde{\pi} \in \Pi_\lambda^*$ such that

$$\alpha_{\tilde{\pi}}(z^*, \theta^*) = \tilde{\alpha}.$$

This can be shown in a straightforward manner by considering a randomized policy $\tilde{\pi}$ that at each time instant is chosen to be π with probability κ and π' with probability $(1 - \kappa)$, where $\kappa \in [0, 1]$. The type I error probability of such a mixed policy is given by the integral equation

$$\alpha_{\tilde{\pi}}(z, \theta) = \kappa\psi\delta + (1 - \kappa)\psi'\delta' + \int \alpha_{\tilde{\pi}} dH_{z, \theta}$$

which has the unique solution $\alpha_{\tilde{\pi}} = \kappa\alpha_\pi + (1 - \kappa)\alpha_{\pi'}$ so that

$$\alpha_{\tilde{\pi}}(z^*, \theta^*) = \kappa\alpha_\pi(z^*, \theta^*) + (1 - \kappa)\alpha_{\pi'}(z^*, \theta^*) = \kappa\alpha + (1 - \kappa)\alpha',$$

Knowing that $\{\alpha_\pi(z^*, \theta^*) : \pi \in \Pi_\lambda^*\}$ and $\partial_{z_0} \frac{\rho_\lambda}{\lambda_0}(z^*, \theta^*)$ are both intervals on the real line, their identity can be established by comparing their endpoints. Consider the upper endpoint of $\partial\rho_\lambda(z^*, \theta^*)$, i.e.,

$$\max_{r_0 \in \partial_{z_0} \rho_\lambda} r_0(z^*, \theta^*). \quad (17)$$

Assume that some r_0 exists that solves the above maximization, but is not contained in $\{\alpha_\pi(z, \theta) : \pi \in \Pi_\lambda^*\}$. This implies that for all $\pi \in \Pi_\lambda^*$

$$r_0(z^*, \theta^*) \neq \lambda_0 \psi\delta + (1 - \psi) \int r_0 dH_{z^*, \theta^*}^0.$$

However, due to (15), a function $r'_0 \in \partial_{z_0} \rho_\lambda$ and a policy $\pi \in \Pi_\lambda^*$ are guaranteed to exist such that

$$r_0(z^*, \theta^*) = \lambda_0 \psi\delta + (1 - \psi) \int r'_0 dH_{z^*, \theta^*}^0.$$

By definition of r_0 and r'_0 it holds that

$$\hat{r}_0 := \max\{r_0, r'_0\} \in \partial_{z_0} \rho_\lambda, \quad (18)$$

with $\hat{r}_0(z^*, \theta^*) = r_0(z^*, \theta^*)$. This, in turn, implies that

$$\hat{r}_0(z^*, \theta^*) = r_0(z^*, \theta^*) = \lambda_0 \psi\delta + (1 - \psi) \int r'_0 dH_{z^*, \theta^*}^0 \leq \lambda_0 \psi\delta + (1 - \psi) \int \hat{r}_0 dH_{z^*, \theta^*}^0.$$

Since $r_0(z^*, \theta^*)$ solves (17), the last inequality has to hold with equality so that

$$\hat{r}_0(z^*, \theta^*) = \lambda_0 \psi\delta + (1 - \psi) \int \hat{r}_0 dH_{z^*, \theta^*}^0$$

solves (17) and

$$\frac{\hat{r}_0}{\lambda_0}(z^*, \theta^*) \in \{\alpha_\pi(z^*, \theta^*) : \pi \in \Pi_\lambda^*\}.$$

This implies that for every $(z, \theta) \in \Omega_\rho$ there exists $\pi \in \Pi_\lambda^*$ for which $\lambda_0 \alpha_\pi(z, \theta)$ constitutes the upper endpoint of the generalized differential $\partial_{z_0} \rho_\lambda(z, \theta)$. The analogous result for the lower endpoint can be shown by taking the minimum instead of the maximum in (18). In summary, this yields

$$\{ \alpha_\pi(z, \theta) : \pi \in \Pi_\lambda^* \} \supset \partial_{z_0} \frac{\rho_\lambda}{\lambda_0}(z, \theta),$$

for every $(z, \theta) \in \Omega_\rho$, which concludes the proof of the second statement in Theorem 8.

A.9 Proof of Theorem 9

The idea underlying the proof of Theorem 9 is to show that a function $\tilde{\rho}_\lambda: \Omega_\rho \rightarrow \mathbb{R}_+$ exists such that for all $(z, \theta) \in \Omega_\rho$ it holds that

$$\rho_\lambda(z, \theta) = \tilde{\rho}_\lambda(\lambda z, \theta), \quad (19)$$

where λz is a shorthand notation for the element-wise product, i.e., $\lambda z = (\lambda_0 z_0, \lambda_1 z_1)$. If (19) holds, it follows that

$$\begin{aligned} \partial_{z_i} \rho_\lambda(z, \theta) &= \partial_{z_i} (\tilde{\rho}_\lambda(\lambda z, \theta)) = \lambda_i \partial_i \tilde{\rho}_\lambda(\lambda z, \theta), \\ \partial_{\lambda_i} \rho_\lambda(z, \theta) &= \partial_{\lambda_i} (\tilde{\rho}_\lambda(\lambda z, \theta)) = z_i \partial_i \tilde{\rho}_\lambda(\lambda z, \theta), \end{aligned}$$

where $\partial_i \tilde{\rho}$ denotes the generalized differential of $\tilde{\rho}$ with respect to its i th argument. This immediately yields the statement in Theorem 9:

$$\partial_{z_i} \frac{\rho_\lambda}{\lambda_i}(z, \theta) = \partial_i \tilde{\rho}_\lambda(\lambda z, \theta) = \partial_{\lambda_i} \frac{\rho_\lambda}{z_i}(z, \theta).$$

The existence of $\tilde{\rho}_\lambda(\lambda z, \theta)$ can be shown via induction. Let the sequence ρ_λ^n be as defined in Lemma 1 and assume that (19) holds for some $n \geq 0$, i.e., a function $\tilde{\rho}_\lambda$ exists such that $\rho_\lambda^n(z, \theta) = \tilde{\rho}_\lambda^n(\lambda z, \theta)$. It then follows that

$$\begin{aligned} \rho_\lambda^{n+1}(z, \theta) &= \min \left\{ g_\lambda(z), 1 + \int \rho_\lambda^n \left(z_0^n \frac{p_{\theta_n}^0(x)}{p_{\theta_n}(x)}, z_1^n \frac{p_{\theta_n}^1(x)}{p_{\theta_n}(x)}, \xi_{\theta_n}(x) \right) dP_{\theta_n}(x) \right\} \\ &= \min \left\{ g(\lambda z), 1 + \int \tilde{\rho}_\lambda^n \left(\lambda_0 z_0^n \frac{p_{\theta_n}^0(x)}{p_{\theta_n}(x)}, \lambda_1 z_1^n \frac{p_{\theta_n}^1(x)}{p_{\theta_n}(x)}, \xi_{\theta_n}(x) \right) dP_{\theta_n}(x) \right\} \\ &=: \tilde{\rho}_\lambda^{n+1}(\lambda z, \theta), \end{aligned}$$

where

$$g(\lambda z) = \min\{\lambda_0 z_0, \lambda_1 z_1\} = g_\lambda(z).$$

The induction basis is given by $\rho_\lambda^0(z) = g_\lambda(z) = g(\lambda z) = \tilde{\rho}_\lambda^0(\lambda z, \theta)$.

A.10 Proof of Corollary 5

Since $L_{\alpha,\beta}(\lambda)$ is concave in λ , a sufficient condition for λ^* to be a solution of (3.27) is

$$0 \in \partial_{\lambda_i} L_{\alpha,\beta}(\lambda^*), \quad i = 0, 1.$$

By Theorem 8 and Theorem 9 it holds that

$$\begin{aligned}\partial_{\lambda_0} L_{\alpha,\beta}(\lambda) &= \partial_{\lambda_0} \rho_\lambda(1, 1, \theta_0) - \alpha \\ \partial_{\lambda_1} L_{\alpha,\beta}(\lambda) &= \partial_{\lambda_1} \rho_\lambda(1, 1, \theta_0) - \beta\end{aligned}$$

so that for $\lambda = \lambda^*$

$$\begin{aligned}\alpha &\in \partial_{\lambda_0} \rho_{\lambda^*}(1, 1, \theta_0), \\ \beta &\in \partial_{\lambda_1} \rho_{\lambda^*}(1, 1, \theta_0),\end{aligned}$$

which implies Corollary 5.

A.11 Proof of Theorem 10

Qualitatively speaking, the validity of the relaxation in Theorem 10 follows from the fact that every $\rho_\lambda(z, \theta)$ is a nondecreasing function of $\rho(\mathcal{E})$ for all $\mathcal{E} \in \Omega_\rho$. Therefore, maximizing ρ_λ at one point implies maximizing ρ_λ over the entire state space.

To formalize this, let ρ_λ^* be the solution of (3.27) and $\tilde{\rho}_\lambda$ be the solution of the corresponding relaxed problem in Theorem 10. Since ρ_λ^* is unique, $\tilde{\rho}_\lambda = \rho_\lambda^*$ whenever $\tilde{\rho}_\lambda$ fulfills the relaxed constraint with equality. Hence, only the case when equality does not hold needs closer inspection. In this case, a function $\tilde{\rho}_\lambda + \Delta\rho_\lambda$ with

$$\Delta\rho_\lambda = \min \left\{ g_\lambda, 1 + \int \tilde{\rho}_\lambda dH_{z,\theta} \right\} - \tilde{\rho}_\lambda \geq 0$$

can be constructed, without changing λ , that still fulfills the inequality constraint, but dominates $\tilde{\rho}_\lambda$. This procedure can be repeated to create a nondecreasing sequence of functions that converges to a solution of the non-relaxed problem (3.27). Since $\tilde{\rho}_\lambda$ is assumed to be optimal, this implies that $\tilde{\rho}_\lambda \leq \rho_\lambda^*$, but that $\tilde{\rho}_\lambda(1, 1, \theta_0) = \rho_\lambda^*(1, 1, \theta_0)$. For this to hold, $\tilde{\rho}_\lambda$ and ρ_λ^* must differ only on a set of states that are reached from $(1, 1, \theta_0)$ with probability zero under P . This implies the equivalence stated in Theorem 10.

A.12 Enforcing Equality in the Constraint of the Relaxed Linear Program

If numerical problems arise in the solution of (3.28) such that the inequality constraint is not fulfilled with equality, a regularization term can be added to the objective function, namely,

$$\begin{aligned} \max_{\lambda \in \mathbb{R}_+^2, \rho_\lambda \in \mathcal{L}_+^H} & \rho_\lambda(1, 1, \theta_0) - \lambda_0 \alpha - \lambda_1 \beta + c \int \rho \, d\tilde{\mu} \\ \text{s.t. } & \rho_\lambda(z, \theta) \leq \lambda_0 z_0, \\ & \rho_\lambda(z, \theta) \leq \lambda_1 z_1, \\ & \rho_\lambda(z, \theta) \leq 1 + \int \rho_\lambda \, dH_{z, \theta}. \end{aligned} \quad (20)$$

where $\tilde{\mu}$ is some strictly increasing measure on $(\Omega_\rho, \mathcal{F}_\rho)$ and c is a small positive constant. In (20), ρ is explicitly maximized over the entire state space, whereas in (3.28) this maximization resulted indirectly from maximizing $\rho(1, 1, \theta_0)$. Note that this regularization of the original problem is by no means the only way to combat numerical artifacts and is not essential to this work. Nevertheless, it is straightforward and yields good results in practice.

Since ρ is increasing in λ , c has to be chosen small enough such that the problem is still bounded. To guarantee this, the additional integral term can be upper bounded by

$$\int \rho_\lambda(z, \theta) \, d\tilde{\mu}(z, \theta) < \int g_\lambda(z) \, d\tilde{\mu}(z, \theta) < \lambda_0 \int z_0 \, d\tilde{\mu}(z, \theta) + \lambda_0 \int z_1 \, d\tilde{\mu}(z, \theta) = \lambda_0 + \lambda_1,$$

where, without loss of generality, it is assumed that $\tilde{\mu}$ is chosen such that

$$\int z_i \, d\tilde{\mu}(z, \theta) = 1, \quad i = 0, 1.$$

The regularized objective function in (20) is then bounded by

$$\rho(1, 1, \theta_0) - \lambda_0(\alpha - c) - \lambda_1(\beta - c).$$

Consequently, choosing $c < \min\{\alpha, \beta\}$ guarantees boundedness. Furthermore, this shows that the regularized problem corresponds to the original problem with smaller target error probabilities, which means that the solution, even though not strictly optimal anymore, still satisfies the original error requirements.

For the discretized problem, the additional integral term can be replaced by a weighted sum of all elements of the vector ρ and the above considerations can be used to determine the constant c so that the solution of the regularized problem is sufficiently close to the original one.

A.13 Proof of Theorem 11

Theorem 11 can be shown via induction. The proof is only given for Q_0 since the results for Q_1 and Q follow analogously.

Assume that Q_0 is chosen according to Theorem 11 and that for some $n \leq N$ it is the case that

$$\alpha_{\pi, Q_0}^n \geq \alpha_{\pi, P_0}^n. \quad (21)$$

This implies that for all $t \in \Omega_T$ and all $P_0 \in \mathcal{P}_0$

$$\begin{aligned} \alpha_{\pi, Q_0}^{n-1}(t) &= \psi_{n-1}(t)\delta_{n-1}(t) + (1 - \psi_{n-1}(t))E_{P_0}[\alpha_{\pi, P_0}^n(T^n) \mid T^{n-1} = t] \\ &\geq \psi_{n-1}(t)\delta_{n-1}(t) + (1 - \psi_{n-1}(t))E_{P_{X_n|t}}[\alpha_{\pi, Q_0}^n(T^n) \mid T^{n-1} = t] \\ &\geq \psi_{n-1}(t)\delta_{n-1}(t) + (1 - \psi_{n-1}(t))E_{Q_{X_n|t}}[\alpha_{\pi, Q_0}^n(T^n) \mid T^{n-1} = t] \end{aligned}$$

so that

$$\alpha_{\pi, Q_0}^{n-1} \geq \alpha_{\pi, P_0}^n.$$

Repeatedly applying the same procedure yields

$$\alpha_{\pi, Q_0}^0 \geq \alpha_{\pi, P_0}^0 \quad \forall P_0 \in \mathcal{P}_0$$

and hence

$$E_{Q_0}[\delta_\tau] = \alpha_{\pi, Q_0}^0(t_0) \geq \alpha_{\pi, P_0}^0(t_0) = E_{P_0}[\delta_\tau] \quad \forall P_0 \in \mathcal{P}_0.$$

The induction basis (21) is given by

$$\alpha_{\pi, P_0}^0 = \alpha_{\pi, Q_0}^0 = \delta_N.$$

A.14 Proof of Theorem 12

The proof of Theorem 12 is given by showing that the least favorable distributions for time-invariant sequential tests are determined by integral equations that are solved by Q_0 , Q_1 and Q in Theorem 12. Again, a detailed proof is only given for Q_0 .

The integral equations that characterize the least favorable distributions for time-invariant tests are obtained by letting the horizon of the corresponding truncated test go to infinity. For the least favorable distribution with respect to the type I error probability this means

$$\alpha_{\pi, Q_0} = \lim_{N \rightarrow \infty} \alpha_{\pi, Q_0}^{n, N}, \quad (22)$$

where $\alpha_{\pi, Q_0}^{n,N}$ is obtained from $\alpha_{\pi, Q_0}^{n+1,N}$ via

$$\alpha_{\pi, Q_0}^{n,N}(t) = \psi(t)\delta(t) + (1 - \psi(t)) \max_{P_t^0 \in \mathcal{P}_t^0} E_{P_t^0} \left[\alpha_{\pi, Q_0}^{n+1,N}(T^{n+1}) \mid T^n = t \right].$$

Given it exists, the limit in (22) is defined by the integral equation

$$\alpha_\pi(t) = \psi(t)\delta(t) + (1 - \psi(t)) \max_{P_t^0 \in \mathcal{P}_t^0} E_{P_t^0} \left[\alpha_\pi(T^{n+1}) \mid T^n = t \right], \quad (23)$$

where the subscript Q_0 has been omitted since the distribution is specified implicitly via the maximization on the right hand side. If Q_0 admits the properties in Theorem 12, it holds by definition that

$$\begin{aligned} \alpha_{\pi, Q_0}(t) &= \psi(t)\delta(t) + (1 - \psi(t))E_{Q_t^0} \left[\alpha_\pi(T^{n+1}) \mid T^n = t \right] \\ &= \psi(t)\delta(t) + (1 - \psi(t)) \max_{P_t^0 \in \mathcal{P}_t^0} E_{P_t^0} \left[\alpha_{\pi, Q_0}(T^{n+1}) \mid T^n = t \right] \end{aligned}$$

so that α_{π, Q_0} solves (23) and Q_0 is least favorable. Statement 2 and Statement 3 in Theorem 12 can be proven by the same arguments.

A.15 Proof of Theorem 13

A proof of Theorem 13 can be given in close analogy to the proof of Theorem 3. Let $\pi^* = (\psi^*, \delta^*)$ be as defined in Theorem 13, i.e., π^* solves (5.7) and it holds that

$$\max_{P_0 \in \mathcal{P}_0} E_{P_0} [\delta_\tau^*] = E_{Q_0} [\delta_\tau^*] = \alpha$$

$$\max_{P_1 \in \mathcal{P}_1} E_{P_1} [1 - \delta_\tau^*] = E_{Q_1} [1 - \delta_\tau^*] = \beta$$

What needs to be shown is that

$$\min_{(\psi, \delta) \in \Pi_{\alpha, \beta}} \max_{P \in \mathcal{P}} E_P [\tau(\psi)] = E_Q [\tau(\psi^*)],$$

The proof can again be given by contradiction. Assume that a policy $\pi^\diamond = (\psi^\diamond, \delta^\diamond)$ exists such that

$$\max_{P \in \mathcal{P}} E_P [\tau(\psi^\diamond)] < E_Q [\tau(\psi^*)]$$

and

$$\max_{P_0 \in \mathcal{P}_0} E_{P_0} [\delta_\tau^\diamond] \leq \alpha, \quad \max_{P_1 \in \mathcal{P}_1} E_{P_1} [1 - \delta_\tau^\diamond] \leq \beta.$$

This implies that

$$\begin{aligned}
& \max_{(P_0, P_1, P) \in \mathcal{U}} E_P[\tau(\psi^*)] + \lambda_0 E_{P_0}[\delta_\tau^*] + \lambda_1 E_{P_1}[1 - \delta_\tau^*] \\
&= \max_{P \in \mathcal{P}} E_P[\tau(\psi^*)] + \lambda_0 \max_{P_0 \in \mathcal{P}_0} E_{P_0}[\delta_\tau^*] + \lambda_1 \max_{P_1 \in \mathcal{P}_1} E_{P_1}[1 - \delta_\tau^*] \\
&= E_Q[\tau(\psi^*)] + \lambda_0 E_{Q_0}[\delta_\tau^*] + \lambda_1 E_{Q_1}[1 - \delta_\tau^*] \\
&= E_Q[\tau(\psi^*)] + \lambda_0 \alpha + \lambda_1 \beta \\
&> \max_{P \in \mathcal{P}} E_P[\tau(\psi^\diamond)] + \lambda_0 \max_{P_0 \in \mathcal{P}_0} E_{P_0}[\delta_\tau^\diamond] + \lambda_1 \max_{P_1 \in \mathcal{P}_1} E_{P_1}[1 - \delta_\tau^\diamond]
\end{aligned}$$

which contradicts the assumption that π^* solves (5.7).

A.16 Proof of Theorem 14

The proof of Theorem 14 is given by showing that if a policy π and a triplet of distributions (Q_0, Q_1, Q) satisfy the given conditions, it is a saddle point of the objective function in (5.7) and hence a solution of the minimax problem.

To show that (π^*, Q_0, Q_1, Q) is a saddle point is to show that, on the one hand,

$$\pi^* \in \arg \min_{(\psi, \delta) \in \Pi} E_Q[\tau(\psi)] + \lambda_0 E_{Q_0}[\delta_\tau] + \lambda_1 E_{Q_1}[1 - \delta_\tau] \quad (24)$$

and, on the other hand,

$$(Q_0, Q_1, Q) \in \arg \max_{(P_0, P_1, P) \in \mathcal{U}} E_P[\tau(\psi^*)] + \lambda_0 E_{P_0}[\delta_\tau^*] + \lambda_1 E_{P_1}[1 - \delta_\tau^*]. \quad (25)$$

Since there is no coupling between the individual terms in (25), the maximum operator distributes and (25) is equivalent to

$$Q_0 \in \arg \max_{P_0 \in \mathcal{P}_0} E_{P_0}[\delta_\tau^*], \quad (26)$$

$$Q_1 \in \arg \max_{P_1 \in \mathcal{P}_1} E_{P_1}[1 - \delta_\tau^*], \quad (27)$$

$$Q \in \arg \max_{P \in \mathcal{P}} E_P[\tau(\psi^*, \delta^*)]. \quad (28)$$

The evidence of (24) follows directly from Theorem 5 in Section 3.2 and the construction of π^* as an optimal testing policy for (Q_0, Q_1, Q) .

A proof of (26)–(28) can be given by combining the results from Section 3.3 with the characterization of least favorable distributions in Section 5.2. First, if π^* is chosen

according to Theorem 14, it is the case that the sequential test is time-invariant and uses the test statistic

$$T^n(x_1, \dots, x_n) = (z_0^n, z_1^n, \theta_n), \quad (29)$$

where

$$z_0^n = \frac{dQ_0}{dQ}(x_1, \dots, x_n) \quad \text{and} \quad z_1^n = \frac{dQ_1}{dQ}(x_1, \dots, x_n).$$

Let the focus for now be on Q_0 . By Theorem 12, Q_0 is least favorable with respect to the type I error probability if for all $t \in \Omega_T$ it holds that

$$Q_0 \in \arg \max_{P_0 \in \mathcal{P}_\theta^0} E_{P_0} [\alpha_{\pi, Q_0}(T^n) \mid T^{n-1} = t], \quad (30)$$

where α_{π, Q_0} is independent of n since the underlying sequential test is time-invariant. By inserting the test statistic (29) into (30) and writing the expected value integral explicitly, (30) becomes

$$Q_{z,\theta} \in \arg \max_{P_{z,\theta}^0 \in \mathcal{P}_\theta^0} \int \alpha_{\pi^*, Q_0} \left(z_0 \frac{q_\theta^0(x)}{q_\theta(x)}, z_1 \frac{q_\theta^1(x)}{q_\theta(x)}, \xi_{z,\theta}(x) \right) dP_{z,\theta}^0(x).$$

This is to say, the distribution Q_0 is least favorable, if for all $(z, \theta) \in \Omega_\rho$ the conditional distributions $Q_{z,\theta}$ solve

$$\max_{p_{z,\theta}^0} \lambda_0 z_0 I_\alpha(p_{z,\theta}^0) \quad \text{s.t.} \quad p_{z,\theta}^0 \in \mathcal{P}_\theta^0 \quad (31)$$

where

$$I_\alpha(p_{z,\theta}^0) := \int \alpha_{\pi^*, Q_0} \left(z_0 \frac{q_\theta^0(x)}{q_\theta(x)}, z_1 \frac{q_\theta^1(x)}{q_\theta(x)}, \xi_{z,\theta}(x) \right) p_{z,\theta}^0(x) d\mu(x).$$

The reason for introducing the scaling factor $\lambda_0 z_0$ will become clear soon. In any case, since λ and z are not subject to the optimization, this scaling does not affect the least favorable densities.

The next step is to show that if a conditional distribution $Q_{z,\theta}^0$ solves (5.8) in Theorem 14, it also solves (31) and is, hence, least favorable with respect to the type I error probability. Problem (5.8) is given by

$$\max_{p_{z,\theta}^0, p_{z,\theta}^1, p_{z,\theta}} I_\rho(p_{z,\theta}^0, p_{z,\theta}^1, p_{z,\theta}) \quad \text{s.t.} \quad p_{z,\theta}^0 \in \mathcal{P}_\theta^0, \quad p_{z,\theta}^1 \in \mathcal{P}_\theta^1, \quad p_{z,\theta} \in \mathcal{P}_\theta \quad (32)$$

where

$$I_\rho(p_{z,\theta}^0, p_{z,\theta}^1, p_{z,\theta}) := \int \rho_{\pi^*} \left(z_0 \frac{p_{z,\theta}^0(x)}{p_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{p_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) p_{z,\theta}(x) d\mu(x). \quad (33)$$

Since for now only $p_{z,\theta}^0$ is of interest, the maximization can be simplified by inserting the known solutions for $p_{z,\theta}^1$ and $p_{z,\theta}$ so that Problem (32) becomes

$$\max_{p_{z,\theta}^0} I_\rho(p_{z,\theta}^0, q_{z,\theta}^1, q_{z,\theta}) \quad \text{s.t.} \quad p_{z,\theta}^0 \in \mathcal{P}_\theta^0 \quad (34)$$

The next step is to show that (34) and (31) are both convex problems in $p_{z,\theta}^0$. A proof is straightforward: The sets \mathcal{P}_θ^0 are convex by definition. The functional $I_\alpha(p_{z,\theta}^0)$ is linear in $p_{z,\theta}^0$ and hence concave. The functional $I_\rho(p_{z,\theta}^0, p_{z,\theta}^1, p_{z,\theta})$, on the other hand, is the integral over the *perspective* of $\rho_{\pi^*}(z, \theta)$ with respect to z . The perspective of a convex/concave function $f: \mathbb{R}^K \rightarrow \mathbb{R}$ is defined as

$$f(y, c) = f\left(\frac{y_1}{c}, \dots, \frac{y_K}{c}\right)c,$$

where $y \in \mathbb{R}^K$ and $c \in \mathbb{R}_+$, and is well known to be jointly convex/concave in (y, c) —see, for example, [DM08]. Since $\rho_{\pi^*}(z, \theta)$ is concave in z and integration preserves convexity/concavity, [Rø04], $I_\rho(p_{z,\theta}^0, p_{z,\theta}^1, p_{z,\theta})$ is concave in $(p_{z,\theta}^0, p_{z,\theta}^1, p_{z,\theta})$.

By definition, $q_{z,\theta}^0$ is a solution of (34). Moreover, (34) and (31) differ only in terms of the objective function. Hence, it follows from Theorem 6 that $q_{z,\theta}^0$ is a joint solution of both problems if it is the case that

$$\lambda_0 z_0 \partial_{p_{z,\theta}^0} I_\alpha(q_{z,\theta}^0) = \partial_{p_{z,\theta}^0} I_\rho(q_{z,\theta}^0).$$

This identity can be shown by straightforward calculation of both expressions. The partial Fréchet differential of $I_\alpha(q_{z,\theta}^0)$ is given by

$$\begin{aligned} \partial_{p_{z,\theta}^0} I_\alpha(q_{z,\theta}^0) &= \partial_{p_{z,\theta}^0} \left(\int \alpha_{\pi^*, Q_0} \left(z_0 \frac{q_\theta^0(x)}{q_\theta(x)}, z_1 \frac{q_\theta^1(x)}{q_\theta(x)}, \xi_{z,\theta}(x) \right) p_{z,\theta}^0(x) d\mu(x) \right) \Big|_{p_{z,\theta}^0 = q_{z,\theta}^0} \\ &= \alpha_{\pi^*, Q_0} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \end{aligned}$$

so that

$$\lambda_0 z_0 \partial_{p_{z,\theta}^0} I_\alpha(q_{z,\theta}^0) = \lambda_0 z_0 \alpha_{\pi^*, Q_0} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right)$$

The Fréchet differential of $I_\rho(q_{z,\theta}^0, q_{z,\theta}^1, q_{z,\theta})$ calculates to

$$\begin{aligned} \partial_{p_{z,\theta}^0} I_\rho(q_{z,\theta}^0, q_{z,\theta}^1, q_{z,\theta}) &= \partial_{p_{z,\theta}^0} \left(\int \rho_{\pi^*} \left(z_0 \frac{p_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_\theta(x) \right) q_{z,\theta}(x) d\mu(x) \right) \Big|_{p_{z,\theta}^0 = q_{z,\theta}^0} \\ &= z_0 \frac{1}{q_{z,\theta}(x)} \partial_{z_0} \rho_{\pi^*} \left(z_0 \frac{p_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_\theta(x) \right) q_{z,\theta}(x) \Big|_{p_{z,\theta}^0 = q_{z,\theta}^0} \\ &= z_0 \partial_{z_0} \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right). \end{aligned}$$

By definition of ρ_{π^*} and Theorem 8 it further holds that

$$\partial_{z_0} \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) = \lambda_0 \alpha_{\pi^*, Q_0} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right)$$

which yields

$$\partial_{p_{z,\theta}^0} I_\rho(q_{z,\theta}^0, q_{z,\theta}^1, q_{z,\theta}) = \lambda_0 z_0 \alpha_{\pi^*, Q_0} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) = \lambda_0 z_0 \partial_{p_{z,\theta}^0} I_\alpha(q_{z,\theta}^0).$$

Consequently, $q_{z,\theta}^0$ satisfies the first order optimality conditions of (31) and is least favorable with respect to the type I error probability.

The proof that $q_{z,\theta}^1$ is least favorable with respect to the type II error probability can be given analogously. What needs to be shown is that for every $(z, \theta) \in \Omega_\rho$ it is the case that $q_{z,\theta}^1$ solves

$$\max_{p_{z,\theta}^1} \lambda_1 z_1 I_\beta(p_{z,\theta}^1) \quad \text{s.t.} \quad p_{z,\theta}^1 \in \mathcal{P}_\theta^1, \quad (35)$$

where

$$I_\beta(p_{z,\theta}^1) := \int \beta_{\pi^*, Q_1} \left(z_0 \frac{q_\theta^0(x)}{q_\theta(x)}, z_1 \frac{q_\theta^1(x)}{q_\theta(x)}, \xi_{z,\theta}(z) \right) p_{z,\theta}^1(x) d\mu(x)$$

and the objective function has again been scaled in order to simplify the comparison of the differentials later on. It is known that $q_{z,\theta}^1$ solves

$$\max_{p_{z,\theta}^1} I_\rho(q_{z,\theta}^0, p_{z,\theta}^1, q_{z,\theta}) \quad \text{s.t.} \quad p_{z,\theta}^1 \in \mathcal{P}_\theta^1,$$

with I_ρ defined in (33). Evaluating the partial differentials with respect to $p_{z,\theta}^1$ at $q_{z,\theta}^1$ yields

$$\begin{aligned} \partial_{p_{z,\theta}^1} I_\beta(q_{z,\theta}^1) &= \partial_{p_{z,\theta}^1} \left(\int \beta_{\pi^*, Q} \left(z_1 \frac{q_\theta^0(x)}{q_\theta(x)}, z_1 \frac{q_\theta^1(x)}{q_\theta(x)}, \xi_{z,\theta}(x) \right) p_{z,\theta}^1(x) d\mu(x) \right) \Big|_{p_{z,\theta}^1=q_{z,\theta}^1} \\ &= \beta_{\pi^*, Q} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \end{aligned}$$

and

$$\begin{aligned} \partial_{p_{z,\theta}^1} I_\rho(q_{z,\theta}^0, q_{z,\theta}^1, q_{z,\theta}) &= \partial_{p_{z,\theta}^1} \left(\int \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) q_{z,\theta}(x) d\mu(x) \right) \Big|_{p_{z,\theta}^1=q_{z,\theta}^1} \\ &= z_1 \frac{1}{q_{z,\theta}(x)} \partial_{z_1} \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{p_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) q_{z,\theta}(x) \Big|_{p_{z,\theta}^1=q_{z,\theta}^1} \\ &= z_1 \partial_{z_1} \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \\ &= \lambda_1 z_1 \beta_{\pi^*, Q_1} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right). \end{aligned}$$

This concludes the proof.

Finally, the proof that $q_{z,\theta}$ is least favorable with respect to the expected run-length follows the exact same steps, the only difference being that evaluating the partial differential of ρ_{π^*} with respect to $p_{z,\theta}$ is slightly more cumbersome. What needs to be shown is that for every $(z, \theta) \in \Omega_\rho$ it is the case that $q_{z,\theta}$ solves

$$\max_{p_{z,\theta}} I_\gamma(p_{z,\theta}) \quad \text{s.t.} \quad p_{z,\theta} \in \mathcal{P}_\theta \quad (36)$$

where

$$I_\gamma(p_{z,\theta}) := \int \gamma_{\pi^*, Q} \left(z_0 \frac{q_\theta^0(x)}{q_\theta(x)}, z_1 \frac{q_\theta^1(x)}{q_\theta(x)}, \xi_{z,\theta}(x) \right) p_{z,\theta}(x) d\mu(x).$$

It is known that $q_{z,\theta}$ solves

$$\max_{p_{z,\theta}} I_\rho(q_{z,\theta}^0, q_{z,\theta}^1, p_{z,\theta}) \quad \text{s.t.} \quad p_{z,\theta} \in \mathcal{P}_\theta,$$

with I_ρ defined in (33). Evaluating the partial differentials with respect to $p_{z,\theta}$ at $q_{z,\theta}$ yields

$$\begin{aligned} \partial_{p_{z,\theta}} I_\gamma(q_{z,\theta}) &= \partial_{p_{z,\theta}} \left(\int \gamma_{\pi^*, Q} \left(z_1 \frac{q_\theta^0(x)}{q_\theta(x)}, z_1 \frac{q_\theta^1(x)}{q_\theta(x)}, \xi_{z,\theta}(x) \right) p_{z,\theta}(x) d\mu(x) \right) \Big|_{p_{z,\theta}=q_{z,\theta}} \\ &= \gamma_{\pi^*, Q} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \end{aligned}$$

and

$$\begin{aligned} \partial_{p_{z,\theta}} I_\rho(q_{z,\theta}^0, q_{z,\theta}^1, q_{z,\theta}) &= \partial_{p_{z,\theta}} \left(\int \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) p_{z,\theta}(x) d\mu(x) \right) \Big|_{p_{z,\theta}=q_{z,\theta}} \\ &= \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \\ &\quad - z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)} \partial_{z_0} \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \\ &\quad - z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)} \partial_{z_1} \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \\ &= \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \\ &\quad - \lambda_0 z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)} \alpha_{\pi^*, Q_0} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \\ &\quad - \lambda_1 z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)} \beta_{\pi^*, Q_1} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right). \end{aligned} \quad (37)$$

By Theorem 6 it holds that

$$\begin{aligned} \rho_{\pi^*} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) &= \gamma_{\pi^*,Q} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \\ &\quad + \lambda_0 z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)} \alpha_{\pi^*,Q_0} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right) \\ &\quad + \lambda_1 z_1 \frac{q_{z,\theta}^1(x)}{p_{z,\theta}(x)} \beta_{\pi^*,Q_1} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right), \end{aligned}$$

which inserted back into (37) yields the desired result

$$\partial_{p_{z,\theta}} I_\rho(q_{z,\theta}^0, q_{z,\theta}^1, q_{z,\theta}) = \gamma_{\pi^*,Q} \left(z_0 \frac{q_{z,\theta}^0(x)}{q_{z,\theta}(x)}, z_1 \frac{q_{z,\theta}^1(x)}{q_{z,\theta}(x)}, \xi_{z,\theta}(x) \right).$$

A.17 Proof that ρ_π induces an f -similarity

By Theorem 7, the function ρ_π is concave in z . Since f_ρ is the perspective of ρ_π with respect to z (cf. Appendix A.16), f_ρ is concave in (y_0, y_1, y_2) . Moreover, it holds that for all $(y_0, y_1, y_2) \in \mathbb{R}_+^3$ and every scalar $c > 0$

$$\begin{aligned} f_\rho(cy_0, cy_1, cy_2, \omega) &= \rho_\pi \left(z_0 \frac{cy_0}{cy_2}, z_1 \frac{cy_1}{cy_2}, \xi_{z,\theta}(\omega) \right) cy_2 \\ &= c \rho_\pi \left(z_0 \frac{y_0}{y_2}, z_1 \frac{y_1}{y_2}, \xi_{z,\theta}(\omega) \right) y_2 \\ &= c f_\rho(y_0, y_1, y_2) \end{aligned}$$

and

$$\lim_{c \rightarrow 0} \rho_\pi \left(z_0 \frac{cy_0}{cy_2}, z_1 \frac{cy_1}{cy_2}, \xi_{z,\theta}(\omega) \right) cy_2 = \lim_{c \rightarrow 0} c \rho_\pi \left(z_0 \frac{y_0}{y_2}, z_1 \frac{y_1}{y_2}, \xi_{z,\theta}(\omega) \right) y_2 = 0$$

so that f_ρ is homogeneous in the sense of Definition 6 and I_{f_ρ} is a valid f -similarity.

A.18 Proof of Theorem 15

The proof is given by showing that every vector of distributions (Q_0, \dots, Q_K) that satisfies the conditions in Theorem 15 also satisfies the first order optimality conditions corresponding to problem (5.15) and is hence a global maximizer of the functional I_f .

Let \mathcal{L}_∞ be the dual space of \mathcal{L}_1 , as defined in (4.11). The Lagrange function $L: (\mathcal{L}_1 \times \mathcal{L}_\infty \times \mathcal{L}_\infty \times \mathbb{R})^K \rightarrow \mathbb{R}$ corresponding to (5.15) calculates to

$$L(p, u, v, c) = \int \left(f(p(\omega), \omega) + \sum_{k=1}^K \zeta_k(p_k(\omega), \omega) \right) d\mu(\omega) + \sum_{k=1}^K c_k,$$

where $p = (p_1, \dots, p_K)$, $u = (u_1, \dots, u_K)$, $v = (v_1, \dots, v_K)$ and

$$\zeta_k(p_k(\omega), \omega) := u_k(\omega)(p_k(\omega) - p''_k(\omega)) + v_k(\omega)(p'_k(\omega) - p_k(\omega)) - c_k p_k(\omega).$$

Here $u_k, v_k \in \mathcal{L}_\infty$ and $c_k \in \mathbb{R}$ denote the Lagrangian multipliers corresponding to the constraints $p_k - p''_k \leq 0$, $p'_k - p_k \leq 0$ and $\int p_k d\mu = 1$, respectively. The dual problem of (5.15) is given by

$$\min_{\substack{u, v \in \mathcal{L}_\infty^K \\ c \in \mathbb{R}^K}} \left\{ \max_{p \in \mathcal{L}_1^K} L(p, u, v, c) \right\} \quad \text{s.t.} \quad u, v \geq 0. \quad (38)$$

The partial Fréchet differentials of L with respect to p_k calculate to

$$\partial_{p_k} L(p, u, v, c)(\omega) = f_{y_k}(p(\omega), \omega) + u_k(\omega) - v_k(\omega) - c_k, \quad (39)$$

where f_{y_k} is defined in Theorem 15. The first order optimality conditions for the least favorable densities require that for all $k \in \{1, \dots, K\}$

$$\begin{aligned} f_{y_k}(q(\omega), \omega) + u_k(\omega) - v_k(\omega) - c_k &= 0 && \text{(stationarity)} \\ p'_k(\omega) \leq q_k(\omega) \leq p''_k(\omega), \int q_k(\omega) d\mu(\omega) &= 1 && \text{(primal feasibility)} \\ u_k(\omega), v_k(\omega) &\geq 0 && \text{(dual feasibility)} \\ u_k(\omega)(q_k(\omega) - p''_k(\omega)) &= v_k(\omega)(p'_k(\omega) - q_k(\omega)) = 0 && \text{(compl. slackness)} \end{aligned}$$

Since I_f is concave in (p_1, \dots, p_K) and the band constraints are affine, these conditions are sufficient for q to be a global maximizer.

Let all q_k and c_k be chosen such that they comply with the conditions in Theorem 15. By construction, this implies that q_k satisfies the primal feasibility constraints. Since f is concave, its partial generalized derivatives exist. Without violating dual feasibility, the functions v_k and u_k can be chosen as

$$\begin{aligned} v_k(\omega) &= \max\{c_k - f_{y_k}(q(\omega), \omega), 0\}, \\ -u_k(\omega) &= \min\{c_k - f_{y_k}(q(\omega), \omega), 0\} \end{aligned}$$

so that

$$v_k(\omega) - u_k(\omega) = f_{y_k}(q(\omega), \omega) - c_k. \quad (40)$$

Inserting (40) back into the stationarity constraints yields

$$f_{y_k}(q(\omega), \omega) + u_k(\omega) - v_k(\omega) - c_k(\omega) = 0$$

for all $k \in \{1, \dots, K\}$. The last step in the proof is to show that these choices for q_k , u_k and v_k also satisfy the complementary slackness constraints, i.e., that

$$\begin{aligned} v_k(\omega) > 0 &\Rightarrow q_k(\omega) = p'_k(\omega), \\ u_k(\omega) > 0 &\Rightarrow q_k(\omega) = p''_k(\omega), \end{aligned}$$

for all $\omega \in \Omega$. Again, this is guaranteed by construction: $v_k(\omega) > 0$ implies that $f_{y_k}(q(\omega), \omega) < c_k$, which in turn implies that $q_k(\omega) = p'_k(\omega)$. Analogously, $u_k(\omega) > 0$ implies $f_{y_k}(q(\omega), \omega) > c_k$ and in turn $q_k = p''_k$.

A.19 Proof of Corollary 9

Corollary 9 is a consequence of the fact that f_{y_k} , being a derivative of a concave function, is nonincreasing in y_k . For the three cases in Theorem 15 it, hence, holds that

$$\begin{aligned} f_{y_k}(q(\omega), \omega) > c_k &\Rightarrow q_k(\omega) = p''_k(\omega) \leq f_{y_k}^{-1}(\bar{q}_k(\omega), \omega, c_k), \\ f_{y_k}(q(\omega), \omega) = c_k &\Rightarrow q_k(\omega) = f_{y_k}^{-1}(\bar{q}_k(\omega), \omega, c_k), \\ f_{y_k}(q(\omega), \omega) < c_k &\Rightarrow q_k(\omega) = p'_k(\omega) \geq f_{y_k}^{-1}(\bar{q}_k(\omega), \omega, c_k). \end{aligned} \tag{41}$$

The expression for q_k given in Corollary 9 is merely a more compact way of writing the constraints on the right hand side of (41).

List of Symbols

The following list contains the most important symbols in the dissertation in alphabetical order. The remaining symbols are introduced where they are used.

$\mathbf{1}_{\mathcal{A}}$	indicator function of the set \mathcal{A}
A	upper likelihood ratio threshold of a sequential probability ratio test
B	lower likelihood ratio threshold of a sequential probability ratio test
$\mathcal{C}, \mathcal{C}_n$	critical region, critical region at time instant n
$d_{\lambda}(z, \theta)$	cost for continuing an optimal test in state (z, θ) for a given λ
$g_{\lambda}(z, \theta)$	cost for stopping an optimal test in state (z, θ) for a given λ
$\mathcal{H}_0, \mathcal{H}_1$	null hypothesis, alternative hypothesis
\mathbb{N}, \mathbb{N}_0	natural numbers excluding zero, including zero
P_0, P_1	distribution of the stochastic process X under $\mathcal{H}_0, \mathcal{H}_1$,
P_X	distribution of the stochastic process X
$P_{X_{n+1} t}$	distribution of X_{n+1} conditioned on $T^n = t$
$P_{X_{n+1} x_1, \dots, x_n}$	distribution of X_{n+1} conditioned on $(X_1, \dots, X_n) = (x_1, \dots, x_n)$
$P_{X_{n+1} z, \theta}$	distribution of X_{n+1} conditioned on $(z^n, \theta_n) = (z, \theta)$
$P_{X_{n+1} \theta}$	distribution of X_{n+1} conditioned on $\theta_n = \theta$
P_t	distribution of arbitrary X_{n+1} conditioned on $T^n = t$
$P_{z, \theta}$	distribution of arbitrary X_{n+1} conditioned on $(z^n, \theta_n) = (z, \theta)$
P_{θ}	distribution of arbitrary X_{n+1} conditioned on $\theta_n = \theta$
\mathcal{P}	uncertainty set for P_X
$\mathcal{P}_0, \mathcal{P}_1$	uncertainty sets for P_X under $\mathcal{H}_0, \mathcal{H}_1$
$\mathcal{P}_{X_{n+1} \theta}$	uncertainty set for the distribution of X_{n+1} conditioned on $\theta_n = \theta$
$\mathcal{P}_{X_{n+1} z, \theta}$	uncertainty set for the distribution of X_{n+1} conditioned on $(z^n, \theta_n) = (z, \theta)$
$\mathcal{P}_{X_{n+1} t}$	uncertainty set for the distribution of X_{n+1} conditioned on $T^n = t$
\mathcal{P}_{θ}	uncertainty set for the distribution of arbitrary X_{n+1} conditioned on $\theta_n = \theta$
$\mathcal{P}_{z, \theta}$	uncertainty set for the distribution of arbitrary X_{n+1} conditioned on $(z^n, \theta_n) = (z, \theta)$
\mathcal{P}_t	uncertainty set for the distribution of arbitrary X_{n+1} conditioned on $T^n = t$
\mathbb{R}, \mathbb{R}_+	real numbers, nonnegative real numbers
$\mathcal{S}, \mathcal{S}_n$	stopping region, stopping region at time instant n
T^n	test statistic at time instant n

x_n	realization of X_n
X_n	random variable corresponding to X at time instant n
z_0^n, z_1^n	likelihood-ratios $\frac{dP_0}{dP}(x_1, \dots, x_n), \frac{dP_1}{dP}(x_1, \dots, x_n)$
α	upper bound on type I error probability
$\alpha_\pi(t)$	type I error probability of a time-invariant test using policy π in state t
$\alpha_{\pi,P}(t)$	type I error probability of a time-invariant test using policy π in state t under distribution P
$\alpha_\pi^n(t)$	type I error probability of a test using policy π in state t at time instant n
$\alpha_{\pi,P}^n(t)$	type I error probability of a test using policy π in state t at time n under distribution P
$\beta_\pi(t)$	type II error probability of a time-invariant test using policy π in state t
$\beta_{\pi,P}(t)$	type II error probability of a time-invariant test using policy π in state t under distribution P
$\beta_\pi^n(t)$	type II error probability of a test using policy π in state t at time instant n
$\beta_{\pi,P}^n(t)$	type II error probability of a test using policy π in state t at time instant n under distribution P
$\gamma_\pi(t)$	expected remaining run-length of a time-invariant test using policy π in state t
$\gamma_{\pi,P}(t)$	expected remaining run-length of a time-invariant test using policy π in state t under distribution P
$\gamma_\pi^n(t)$	expected remaining run-length of a test using policy π in state t at time instant n
$\gamma_{\pi,P}^n(t)$	expected remaining run-length of a test using policy π in state t at time instant n under distribution P
δ, δ_n	decision rule, decision rule at time instant n
Δ	set of randomized decision rules
λ	cost coefficients
μ	reference measure for the definition of probability densities
θ_n	sufficient statistic for (x_1, \dots, x_n)
ξ	function mapping x_n and θ_n to θ_{n+1}
π	testing policy
Π	set of testing policies
Π_λ^*	set of cost minimizing testing policies for cost coefficient λ
$\Pi_{\alpha,\beta}$	set of testing policies with error probabilities α, β

$\Pi_{\alpha,\beta}^*$	set of optimal testing policies with error probabilities α, β
ρ_λ	cost function of cost minimizing sequential test for cost coefficients λ
ρ_π	cost function of optimal sequential test with policy π
τ	stopping time
ψ, ψ_n	stopping rule, stopping rule at time instant n
$(\Omega_T, \mathcal{F}_T)$	state space of the test statistic T^n
$(\Omega_X, \mathcal{F}_X)$	state space of the stochastic process X
$(\Omega_\theta, \mathcal{F}_\theta)$	state space of the sufficient statistic Θ
$(\Omega_\rho, \mathcal{F}_\rho)$	domain of the cost function ρ_λ

References

- [AIS97] Y. I. Alber, A. N. Iusem, and M. V. Solodov, “Minimization of nonsmooth convex functionals in Banach spaces.” *Journal of Convex Analysis*, vol. 4, no. 2, pp. 235–255, 1997.
- [AMR15] M. Avella Medina and E. Ronchetti, “Robust statistics: a selective overview and new directions,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 6, pp. 372–393, 2015.
- [AS66] S. M. Ali and S. D. Silvey, “A general class of coefficients of divergence of one distribution from another,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.
- [AS06] V. Azhmyakov and W. Schmidt, “Approximations of relaxed optimal control problems,” *Journal of Optimization Theory and Applications*, vol. 130, no. 1, pp. 61–78, 2006.
- [ASZS14] S. Al-Sayed, A. Zoubir, and A. Sayed, “Robust distributed detection over adaptive diffusion networks,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7233–7237.
- [Atk89] K. Atkinson, *An Introduction to Numerical Analysis*, 2nd ed. New York City, New York, USA: Wiley, 1989.
- [BD08a] B. E. Brodsky and B. S. Darkhovsky, “Minimax methods for multihypothesis sequential testing and change-point detection problems,” *Sequential Analysis*, vol. 27, no. 2, pp. 141–173, 2008.
- [BD08b] ——, “Minimax sequential tests for many composite hypotheses i,” *Theory of Probability & Its Applications*, vol. 52, no. 4, pp. 565–579, 2008.
- [BEG81] M. Basseville, B. Espiau, and J. Gasnier, “Edge detection using sequential methods for change in level—part I: A sequential edge detection algorithm,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 1, pp. 24–31, 1981.
- [BK08] A. Buja and H. R. Künsch, “A conversation with Peter Huber,” *Statistical Science*, vol. 23, no. 1, pp. 120–135, 2008.
- [BKTG15] I. M. Berenguer, H. Kunze, L. D. Torre, and R. M. Galán, *Interdisciplinary Topics in Applied Mathematics, Modeling and Computational Science*. Springer International Publishing, 2015, ch. Set-valued Nonlinear Fredholm Integral Equations: Direct and Inverse Problem, pp. 65–71.
- [BLS13] J. Bartroff, T. L. Lai, and M.-C. Shih, *Sequential Experimentation in Clinical Trials*. New York City, New York, USA: Springer, 2013.

- [BM55] J. Bussgang and D. Middleton, “Optimum sequential detection of signals in noise,” *IRE Transactions on Information Theory*, vol. 1, no. 3, pp. 5–18, 1955.
- [BMG⁺10] S. Boriah, V. Mithal, A. Garg, V. Kumar, M. Steinbach, C. Potter, and S. A. Klooster, “A comparative study of algorithms for land cover change,” in *Proceedings of the Conference on Intelligent Data Understanding*, 2010, pp. 175–188.
- [Bot14] F. Botelho, *Functional Analysis and Applied Optimization in Banach Spaces*. Cham, Switzerland: Springer, 2014.
- [Box80] J. F. Box, “R. A. Fisher and the design of experiments, 1922–1926,” *The American Statistician*, vol. 34, no. 1, pp. 1–7, 1980.
- [Boy04] L. Boyd, S. and Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.
- [Bro65] C. G. Broyden, “A class of methods for solving nonlinear simultaneous equations,” *Mathematics of Computation*, vol. 19, pp. 577–593, 1965.
- [BV15] T. Banerjee and V. Veeravalli, “Data-efficient minimax quickest change detection with composite post-change distribution,” *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5172–5184, 2015.
- [CCF95] P. Clarkson, S.-W. Chen, and Q. Fan, “A robust sequential detection algorithm for cardiac arrhythmia classification,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1995, pp. 1181–1184.
- [Chi09] N. H. Chieu, “The Fréchet and limiting subdifferentials of integral functionals on the spaces,” *Journal of Mathematical Analysis and Applications*, vol. 360, no. 2, pp. 704 – 710, 2009.
- [Cho11] S. C. Chow, *Controversial Statistical Issues in Clinical Trials*. Boca Raton, Florida, USA: CRC Press, 2011.
- [Cla83] F. H. Clarke, *Optimization and Nonsmooth Analysis*. New York City, New York, USA: Wiley, 1983.
- [Cla00] B. Clarke, “A review of differentiability in relation to robustness with application to seismic data analysis,” *Proceedings of the Indian National Science Academy*, vol. 66, A, no. 5, pp. 467–482, 2000.
- [CM65] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*. York City, New York, USA: Wiley, 1965.
- [CRS71] Y. S. Chow, H. Robbins, and D. Siegmund, *Great Expectations: The Theory of Optimal Stopping*. Boston, Massachusetts, USA: Houghton Mifflin, 1971.

- [Csi63] I. Csiszár, “Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten,” *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, vol. 8, pp. 85–107, 1963.
- [CVMM02] Q. Cheng, P. Varshney, K. Mehrotra, and C. Mohan, “Optimal bandwidth assignment for distributed sequential detection,” in *Proc. of the Fifth International Conference on Information Fusion*, vol. 2, 2002, pp. 1550–1556.
- [DeG60] M. H. DeGroot, “Minimax sequential tests of some composite hypotheses,” *Annals of Mathematical Statistics*, vol. 31, no. 4, pp. 1193–1200, 1960.
- [DKW53] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, “Sequential decision problems for processes with continuous time parameter. testing hypotheses,” *Annals of Mathematical Statistics*, vol. 24, no. 2, pp. 254–264, 1953.
- [DM65] A. Y. Dubovitskiy and A. A. Milyutin, “Extremum problems in the presence of constraints,” *Zhurnal Vychislitel’noi Matematiki i Matematicheskoi Fiziki*, vol. 5, pp. 395–453, 1965.
- [DM08] B. Dacorogna and P. Marchal, “The role of perspective functions in convexity, polyconvexity, rank-one convexity and separate convexity,” *Journal of Convex Analysis*, vol. 15, no. 2, pp. 271–284, 2008.
- [DN88] V. P. Dragalin and A. A. Novikov, “Asymptotic solution of the Kiefer–Weiss problem for processes with independent increments,” *Theory of Probability & Its Applications*, vol. 32, no. 4, pp. 617–627, 1988.
- [dOPdG10] P. de Oude, G. Pavlin, and J. de Groot, “Robust Bayesian detection: A case study,” in *Proceedings of the 13th Conference on Information Fusion (FUSION)*, 2010, pp. 1–8.
- [Dud02] R. M. Dudley, *Real Analysis and Probability*, ser. Cambridge Studies in Advanced Mathematics. Cambridge, UK: Cambridge University Press, 2002.
- [ESV79] A. El-Sawy and V. Vandelinde, “Robust sequential detection of signals in noise,” *IEEE Transactions on Information Theory*, vol. 25, no. 3, pp. 346–353, 1979.
- [Fel66] W. Feller, *An Introduction to Probability Theory and Its Applications—Vol. II*. New York City, New York, USA: Wiley., 1966.
- [Fis66] R. Fisher, *The design of experiments*. New York City, New York, USA: Hafner Publishing Company, 1966.
- [FT12a] G. Fellouris and A. G. Tartakovsky, “Almost optimal sequential tests of discrete composite hypotheses,” 2012. [Online]. Available: <http://arxiv.org/abs/1204.5291>
- [FT12b] ———, “Nearly minimax one-sided mixture-based sequential tests,” *Sequential Analysis*, vol. 31, no. 3, pp. 297–325, 2012.

- [FZ14] M. Fauss and A. Zoubir, “Designing discrete sequential tests via mixed integer programming,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3455–3459.
- [FZ15a] M. Fauß and A. M. Zoubir, “A linear programming approach to sequential hypothesis testing,” *Sequential Analysis*, vol. 34, no. 2, pp. 235–263, 2015.
- [FZ15b] ——, “Old bands, new tracks – revisiting the band model for robust hypothesis testing,” 2015, submitted for publication in the IEEE Transactions on Signal Processing. [Online]. Available: <http://arxiv.org/abs/1510.04524>
- [GL11] J. Geng and L. Lai, “Optimal sequential detection with stochastic energy constraint,” in *Proceedings of the 7th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, 2011, pp. 423–427.
- [GN75] L. Györfi and T. Nemetz, “On the dissimilarity of probability measures,” Mathematical Institute of the Hungarian Academy of Science, Tech. Rep., 1975.
- [GN77] ——, “ f -dissimilarity: a general class of separation measures of several probability distributions,” *Colloquia of the János Bolyai Mathematical Society Mathematical Society: Topics in Information Theory*, vol. 16, pp. 309–321, 1977.
- [GN78] ——, “ f -dissimilarity: A generalization of the affinity of several distributions,” *Annals of the Institute of Statistical Mathematics*, vol. 30, no. 1, pp. 105–113, 1978.
- [GO15] I. Gurobi Optimization, “Gurobi optimizer reference manual,” 2015. [Online]. Available: <http://www.gurobi.com>
- [Gri12] A. Griewank, “Broyden updating, the good and the bad!” *Optimization Stories*,, vol. Documenta Mathematica, Extra Volume ISMP, pp. 301–315, 2012.
- [GS91] B. K. Ghosh and P. K. Sen, Eds., *Handbook of Sequential Analysis*, ser. Statistics: A Series of Textbooks and Monographs. Boca Raton, Florida, USA: CRC Press, 1991.
- [GS00] L. Grippo and M. Sciandrone, “On the convergence of the block nonlinear Gauss–Seidel method under convex constraints,” *Operations Research Letters*, vol. 26, no. 3, pp. 127 – 136, 2000.
- [GS06] C. M. Grinstead and L. J. Snell, *Grinstead and Snell’s Introduction to Probability*. Providence, Rhode Island, USA: American Mathematical Society, 2006. [Online]. Available: <http://math.dartmouth.edu/~{}prob/prob/prob.pdf>

- [Gun11] A. Guntuboyina, “Lower bounds for the minimax risk using f -divergences, and applications,” *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2386–2399, 2011.
- [GZa] G. GÜL and A. M. Zoubir, “Minimax robust hypothesis testing.” [Online]. Available: <http://arxiv.org/abs/1502.00647>
- [GZb] ——, “Robust hypothesis testing with α -divergence.” [Online]. Available: <http://arxiv.org/abs/1501.05019>
- [HB91] R. Hathaway and J. Bezdek, “Grouped coordinate minimization using Newton’s method for inexact minimization in one vector coordinate,” *Journal of Optimization Theory and Applications*, vol. 71, no. 3, pp. 503–516, 1991.
- [Hol75] S. Holm, “Asymptotic minimax character of SPR tests in one-parameter exponential classes,” *Scandinavian Journal of Statistics*, vol. 2, no. 2, pp. 49–60, 1975.
- [HS73] P. J. Huber and V. Strassen, “Minimax tests and the Neyman–Pearson lemma for capacities,” *The Annals of Statistics*, vol. 1, no. 2, pp. 251–263, 1973.
- [Hub64] P. J. Huber, “Robust estimation of a location parameter,” *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.
- [Hub65] ——, “A robust version of the probability ratio test,” *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.
- [Hub81] ——, *Robust Statistics*. Hoboken, New Jersey, USA: Wiley, 1981.
- [Kai67] T. Kailath, “The divergence and Bhattacharyya distance measures in signal selection,” *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967.
- [Kas81] S. Kassam, “Robust hypothesis testing for bounded classes of probability densities (corresp.),” *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 242–247, 1981.
- [Kem58] K. W. Kemp, “Formulae for calculating the operating characteristic and the average sample number of some sequential tests,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 20, no. 2, pp. 379–386, 1958.
- [Kha02] A. Kharin, “On robustifying of the sequential probability ratio test for a discrete model under contaminations,” *Austrian Journal of Statistics*, vol. 31, no. 4, pp. 267–277, 2002.
- [KMZ15] D. Kalus, M. Muma, and A. M. Zoubir, “Distributed robust change point detection for autoregressive processes with an application to voice activity detection,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 3906–3910.

- [Kne14] J. G. Knecht, “Medication and self-efficacy of patients living with heart failure: A mixed-methods study,” Ph.D. dissertation, University of Connecticut Graduate School, 2014.
- [KP85] S. Kassam and H. Poor, “Robust techniques for signal processing: A survey,” *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433–481, 1985.
- [KW57] J. Kiefer and L. Weiss, “Some properties of generalized sequential probability ratio tests,” *Annals of Mathematical Statistics*, vol. 28, no. 1, pp. 57–74, 1957.
- [Lev08] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*, 1st ed. New York City, New York, USA: Springer, 2008.
- [LLL15] B. Liu, J. Lan, and X. Li, “Multiple-model hypothesis testing based on 2-SPRT,” in *Proceedings of the American Control Conference (ACC)*, 2015, pp. 183–188.
- [LLY14] X. Li, J. Liu, and Z. Ying, “Generalized sequential probability ratio test for separate families of hypotheses,” *Sequential Analysis*, vol. 33, no. 4, pp. 539–563, 2014.
- [Lor76] G. Lorden, “2-SPRT’S and the modified Kiefer-Weiss problem of minimizing an expected sample size,” *Annals of Statistics*, vol. 4, no. 2, pp. 281–291, 1976.
- [LR05] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. New York City, New York, USA: Springer, 2005.
- [Lue69] D. G. Luenberger, *Optimization by Vector Space Methods*. New York City, New York, USA: Wiley, 1969.
- [LV87] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig, Germany: Teubner, 1987.
- [LZL12] L. Lu, X. Zhou, and G. Li, “Optimal sequential detection in cognitive radio networks,” in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)*, 2012, pp. 289–293.
- [Mau57] R. J. Maurice, “A minimax procedure for choosing between two populations using sequential sampling,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 19, no. 2, pp. 255–261, 1957.
- [MIT03] G. G. Magaril-Ilyaev and V. M. Tikhomirov, *Convex Analysis: Theory and Applications*, ser. Translations of Mathematical Monographs. Providence, Rhode Island, USA: American Mathematical Society, 2003.
- [MLN09] Y. S. Meng, Y. H. Lee, and B. C. Ng, “The effects of tropical weather on radio-wave propagation over foliage channel,” *IEEE Transactions on Vehicular Technology*, vol. 58, no. 8, pp. 4023–4030, 2009.

- [MMY06] R. Maronna, D. Martin, and V. Yohai, *Robust Statistics: Theory and Methods*. Hoboken, New Jersey, USA: Wiley, 2006.
- [Mor63] T. Morimoto, “Markov processes and the H-theorem,” *Journal of the Physical Society of Japan*, vol. 18, no. 3, pp. 328–331, 1963.
- [Muk05] M. N. Mukherjee, *Elements of Metric Spaces*. Kolkata, India: Academic Publishers, 2005.
- [Nik94] I. Nikiforov, “Sequential optimal detection and isolation of faults in systems with random disturbances,” in *Proc. of the American Control Conference*, vol. 2, 1994, pp. 1853–1857 vol.2.
- [Nov08] A. Novikov, “Optimal sequential multiple hypothesis tests,” 2008. [Online]. Available: <http://arxiv.org/abs/0811.1297>
- [Nov09] ——, “Optimal sequential tests for two simple hypotheses,” *Sequential Analysis*, vol. 28, no. 2, p. 188217, 2009.
- [NP33] J. Neyman and E. S. Pearson, “On the problem of the most efficient tests of statistical hypotheses,” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 231, no. 694-706, p. 289337, 1933.
- [NWJ08] X. L. Nguyen, M. Wainwright, and M. Jordan, “On optimal quantization rules for some problems in sequential decentralized detection,” *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3285–3295, 2008.
- [NWJ09] X. Nguyen, M. J. Wainwright, and M. I. Jordan, “On surrogate loss functions and f -divergences,” *Annals of Statistics*, vol. 37, no. 2, pp. 876–904, 2009.
- [Ö78] F. Österreicher, “On the construction of least favourable pairs of distributions,” *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 43, no. 1, pp. 49–55, 1978.
- [Oga06] S. Ogawa, “Stochastic integral equations of Fredholm type,” (*The 7th Workshop on Stochastic Numerics, Research Institute of Mathematical Sciences, Kyoto University*), vol. 1462, pp. 35–45, 2006.
- [Pag54] E. S. Page, “An improvement to Wald’s approximation for some properties of sequential tests,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 16, no. 1, pp. 136–139, 1954.
- [Pap91] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., ser. McGraw-Hill Series in Electrical Engineering. New York City, New York, USA: McGraw-Hill, 1991.
- [Pap97] N. S. Papageorgiou, “Convex integral functionals,” *Transactions of the American Mathematical Society*, vol. 349, no. 4, pp. 1421–1436, 1997.

- [Par05] L. Pardo, *Statistical Inference Based on Divergence Measures*. Boca Raton, Florida, USA: CRC Press, 2005.
- [Pav91] I. V. Pavlov, “Sequential procedure of testing composite hypotheses with applications to the Kiefer–Weiss problem,” *Theory of Probability & Its Applications*, vol. 35, no. 2, pp. 280–292, 1991.
- [Pea00] K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *Philosophical Magazine Series 5*, vol. 50, no. 302, pp. 157–175, 1900.
- [Pet98] M. Pettersson, “Monitoring a freshwater fish population: statistical surveillance of biodiversity,” *Environmetrics*, vol. 9, no. 2, pp. 139–150, 1998.
- [PG02] P. Planinsic and M. Golob, *Soft Computing and Industry: Recent Applications*. London, UK: Springer London, 2002, ch. Root-Finding of Monotone Nonlinear Functions with Fuzzy Iterative Methods, pp. 711–722.
- [PH09] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, UK: Cambridge University Press., 2009.
- [Pol97] E. Polak, *Optimization – Algorithms and Consistent Approximations*. New York City, New York, USA: Springer, 1997.
- [Poo80] H. Poor, “Robust decision design using a distance criterion,” *IEEE Transactions on Information Theory*, vol. 26, no. 5, pp. 575–587, 1980.
- [PS06] G. Peskir and A. Shiryaev, *Optimal Stopping and Free-Boundary Problems*, ser. Lectures in Mathematics ETH Zürich. Basel, Switzerland: Birkhäuser., 2006.
- [Rø04] F. Rønning, “On the preservation of direction convexity under differentiation and integration,” *Rocky Mountain Journal of Mathematics*, vol. 34, no. 3, pp. 1121–1130, 2004.
- [Roc68] R. T. Rockafellar, “Integrals which are convex functionals.” *Pacific Journal of Mathematics*, vol. 24, no. 3, pp. 525–539, 1968.
- [Roc70] ——, *Convex Analysis*. Princeton, New Jersey, USA: Princeton University Press, 1970.
- [Roc71] ——, “Integrals which are convex functionals II,” *Pacific Journal of Mathematics*, vol. 39, no. 2, pp. 439–469, 1971.
- [Roc74] ——, *Conjugate Duality and Optimization*. Philadelphia, Pennsylvania, USA: Society for Industrial and Applied Mathematics, 1974, ch. 1, pp. 1–74.
- [Rud76] W. Rudin, *Principles of Mathematical Analysis*. New York City, New York, USA: McGraw-Hill, 1976.

- [RW11] M. D. Reid and R. C. Williamson, “Information, divergence and risk for binary experiments,” *Journal of Machine Learning Research*, vol. 12, pp. 731–817, 2011.
- [Sae96] S. Saeki, “A proof of the existence of infinite product probability measures,” *The American Mathematical Monthly*, vol. 103, no. 8, pp. 682–683, 1996.
- [Sch87] N. Schmitz, “Minimax sequential tests of composite hypotheses on the drift of a Wiener process,” *Statistische Heft*, vol. 28, no. 1, pp. 247–261, 1987.
- [Sie85] D. Siegmund, *Sequential Analysis*. New York City, New York, USA: Springer, 1985.
- [SJ15] F. Suratman and S. Jarot, “An efficient implementation of sequential detector in spectrum sensing under correlated observations,” in *Proceedings of the 3rd International Conference on Information and Communication Technology (ICoICT)*, 2015, pp. 140–145.
- [SZ13] F. Suratman and A. Zoubir, “Bootstrap based sequential probability ratio tests,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 6352–6356.
- [Tar13] A. Tartakovsky, “Sequential hypothesis tests: Historical overview and recent results,” 2013, plenary Lecture at the Fourth International Workshop in Sequential Methodologies in Athens, Georgia, USA.
- [Tij12] H. Tijms, *Understanding Probability*. Cambridge, UK: Cambridge University Press, 2012.
- [TLY03] A. G. Tartakovsky, X. R. Li, and G. Yaralov, “Sequential detection of targets in multichannel systems,” *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 425–445, 2003.
- [TNB14] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Boca Raton, Florida, USA: Chapman and Hall/CRC, 2014.
- [TV65] G. M. Tallis and M. K. Vagholfkar, “Formulae to improve Wald’s approximation for some properties of sequential tests,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 27, no. 1, pp. 74–81, 1965.
- [Ulb09] M. Ulbrich, *Optimization with PDE Constraints*. Dordrecht, Netherlands: Springer Netherlands, 2009, ch. Optimization Methods in Banach Spaces, pp. 97–156.
- [UVM11] J. Unnikrishnan, V. Veeravalli, and S. Meyn, “Minimax robust quickest change detection,” *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1604–1614, 2011.

- [Var11] K. Varshney, “Bayes risk error is a Bregman divergence,” *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4470–4472, 2011.
- [VK88] E. Voudouri and L. Kurz, “A robust approach to sequential detection,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1200–1210, 1988.
- [Vos01] H. J. Vos, “A minimax procedure in the context of sequential testing problems in psychodiagnostics,” *British Journal of Mathematical and Statistical Psychology*, vol. 54, pp. 139–159, 2001.
- [Wal47] A. Wald, *Sequential Analysis*. Hoboken, New Jersey, USA: Wiley, 1947.
- [Wal50a] ——, “Basic ideas of a general theory of statistical decision rules,” in *Proc. of the International Congress of Mathematicians*, 1950.
- [Wal50b] ——, *Statistical Decision Functions*, ser. Wiley Series in Probability and Mathematical Statistics, R. A. Bradley and J. S. Hunter, Eds. Hoboken, New Jersey, USA: Wiley, 1950.
- [Wol52] J. Wolfowitz, “Abraham Wald, 1902-1950,” *Annals of Mathematical Statistics*, vol. 23, no. 1, pp. 1–13, 1952.
- [Wri15] S. J. Wright, “Coordinate descent algorithms,” *Math. Program.*, vol. 151, no. 1, pp. 3–34, 2015.
- [WW48] A. Wald and J. Wolfowitz, “Optimum character of the sequential probability ratio test,” *Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326–339, 1948.
- [WWZ77] R. Wheeden, R. Wheeden, and A. Zygmund, *Measure and Integral: An Introduction to Real Analysis*, ser. Chapman & Hall/CRC Pure and Applied Mathematics. Boca Raton, Florida, USA: Taylor & Francis, 1977.
- [YZ99] D. Yao and S. Zheng, “Sequential quality control in batch manufacturing,” *Annals of Operations Research*, vol. 87, no. 0, pp. 3–30, 1999.
- [Zir01] C. L. Zirbel, “Optimal stopping of Markov chains or How to play Blackjack,” 2001, lecture notes. [Online]. Available: <http://www-math.bgsu.edu/~zirbel/blackjack/>
- [ZKCM12] A. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, “Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts,” *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 61–80, 2012.
- [ZMS13] M. V. Zhitlukhin, A. A. Muravlev, and A. N. Shiryaev, “The optimal decision rule in the Kiefer–Weiss problem for a Brownian motion,” *Russian Mathematical Surveys*, vol. 68, no. 2, p. 389, 2013.
- [ZZM92] O. Zeitouni, J. Ziv, and N. Merhav, “When is the generalized likelihood ratio test optimal?” *IEEE Transactions on Information Theory*, vol. 38, pp. 1597–1602, 1992.

Curriculum Vitae

Name: Michael Fauß
Date of birth: 04.02.1986
Place of birth: Seeheim-Jugenheim, Germany
Family status: single

Education

09/2007 - 12/2010	Technische Universität München, Germany Electrical Engineering and Information Technology Diplom (Dipl.-Ing.)
03/2010 - 12/2010	University of Edinburgh, Scotland, United Kingdom Diplomarbeit (Diploma Thesis): “On the Design of Transceive Filters for the Two-Way Amplify-and-Forward Relay Channel”
09/2005 - 09/2007	Technische Universität Darmstadt, Germany Electrical Engineering and Information Technology
06/2005	High school degree (Abitur) at Schuldorf Bergstraße, Seeheim-Jugenheim, Germany

Work experience

since 11/2011	Research Associate at Signal Processing Group Technische Universität Darmstadt, Germany
03/2011 - 09/2011	Research Associate at Methoden der Signalverarbeitung Technische Universität München, Germany
03/2009 - 06/2009	Internship at ITK Engineering, Martinsried, Germany.

Erklärung laut §9 der Promotionsordnung

Ich versichere hiermit, dass ich die vorliegende Dissertation allein und nur unter Verwendung der angegebenen Literatur verfasst habe. Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 08. Februar 2016,

