DECISION-MAKING UNDER STATISTICAL UNCERTAINTY

BY

JAYAKRISHNAN UNNIKRISHNAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Venugopal V. Veeravalli, Chair
Professor Sean P. Meyn, Director of Research
Professor Bruce Hajek
Associate Professor Pramod Viswanath

ABSTRACT

Statistical decision-making procedures are used in a wide range of contexts varying from communication receiver design to environment monitoring systems. Although such procedures have been studied for a long time, much of the focus has been restricted to systems where the underlying probabilistic model is known accurately. In this thesis we consider the setting where there is some uncertainty about the probabilistic model. We focus on two different problems and present approaches to dealing with statistical uncertainty in each of these cases.

For the problem of universal hypothesis testing, we study tests that improve upon the known optimal solution in two different aspects. Firstly, we study the generalized likelihood ratio test (GLRT) that exploits partial knowledge about the alternate distribution to improve finite-sample performance over the Hoeffding test. Although the Hoeffding test is universally optimal in an asymptotic sense, we show that it suffers from high bias and variance which leads to a poor performance over finite observation lengths. The performance degradation of the Hoeffding test is particularly significant for the testing of large alphabet distributions. We also show that the test statistic used in the GLRT is a relaxation of the Kullback-Leibler divergence statistic used in the Hoeffding test. We present results on the asymptotic behavior of the two test statistics to explain the advantage of the GLRT. We then study robust procedures for universal hypothesis testing when there is uncertainty about the null hypothesis. We present new results on the asymptotic behavior of the proposed test statistic which can be used to obtain procedures for setting thresholds in these tests for a target false alarm requirement.

We also study the problem of quickest change detection under statistical uncertainty. We formulate a new problem in robust quickest change detection, in which one seeks to minimize the worst-case delay over all possible

instances of the uncertain distributions subject to false alarm constraints. We adopt Huber's robust approach and identify sufficient conditions under which change detection procedures designed for certain *least-favorable distributions* are robust to uncertainties in a minimax sense. These robust tests are simple to implement and give significant performance improvement over some benchmark procedures that are known to be optimal in an asymptotic sense.

To my mother and father

ACKNOWLEDGMENTS

First and foremost I thank my advisers Professors Venugopal Veeravalli and Sean Meyn for their guidance, patience and support through my doctoral work. I am particularly grateful to Prof. Veeravalli for introducing me to the art of research and for the freedom he gave me in choosing my research topics. I consider myself very fortunate to have had the opportunity to work with Prof. Meyn and to be inspired by his genius and his enthusiasm for research.

I thank Professors Bruce Hajek and Pramod Viswanath for serving on my doctoral committee, and Professor Mark Hasegawa-Johnson for serving on my preliminary exam committee. I am grateful to them for their valuable suggestions. Special thanks to Prof. Hajek for being an inspirational teacher. I also thank my collaborator Dayu Huang for his enthusiasm and his contribution to the results of Chapter 2.

My research experience in graduate school was greatly enriched by the intellectual environment in CSL. I owe much to several current and former members of CSL who helped me in my research. In particular, I thank Akshay Kashyap and Jason Fuemmeler for valuable advice on research in my early days at CSL. I enjoyed stimulating conversations on numerous topics with Sreeram Kannan, Hemant Kowshik, Sreekanth Annapureddy, Vasanthan Raghavan and Kunal Srivastava among others.

I would also like to thank my close friends, both in Champaign-Urbana and outside, who have been a constant source of encouragement and companionship. Finally, I wish to thank my family for supporting me through every step in this endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Statistical decision-making refers to the process of making statistical decisions between a number of alternatives based on observations that are modeled by random variables. It is a well-studied problem with the earliest results [1] dating over seventy years ago. However, most of the research in this area has focussed on problems where it is assumed that the decision-makers have perfect knowledge about the distributions of the observations under the different hypotheses. The focus of this dissertation is on problems in which perfect statistical knowledge about the observations is not available. We study methods to cope with uncertain statistics in two specific problems - universal hypothesis testing (also called goodness-of-fit testing) and quickest change detection.

1.1 Motivation

The problem of hypothesis testing under statistical uncertainty arises naturally in several practical contexts. A simple example is the problem of anomaly detection in which one is interested in testing whether or not the system is in a normal state based on observations of certain key parameters. One may have a good model for the normal behavior of a system but little or no knowledge about the anomalous behavior of the system. Such problems are usually posed as universal hypothesis testing problems, where one assumes that the system behavior in the anomalous state is completely unknown.

Another natural example where one has to deal with uncertain statistics is the problem of fault-onset detection. Just as in the problem of anomaly detection, one often has a good model for the normal behavior of the system and little information of the system behavior in faulty state. A similar

situation arises in intrusion detection, where again one has a good model of system behavior only prior to the intrusion.

In such problems with statistical uncertainty, standard approaches to decision-making cannot be directly used as they can lead to arbitrarily poor performances as argued in the seminal work of Huber [2]. This calls for a new theory that explicitly addresses issues of statistical uncertainty in decision-making problems.

1.2 Existing Results and Known Approaches

There are several known approaches for dealing with statistical uncertainty. In some problems it is possible to obtain the same performance as the optimal test with known statistics. We refer to such schemes as *universal* schemes since they universally achieve optimal performance for all values of the unknown statistics. Examples include the Hoeffding test [3] and the generalized likelihood ratio test [4] for hypothesis testing problems involving infinite sequences of observations. These tests are known to achieve the same performance in terms of error-exponents as tests that have perfect knowledge of the distributions of the observations. We will study these examples in detail in Chapter 2. A simpler example is the following Neyman-Pearson hypothesis testing problem involving a simple null hypothesis and a composite Gaussian alternate hypothesis for the distribution of the observations variable Y:

$$\begin{aligned} \mathcal{H}_0 : Y &\sim \mathcal{N}(0,1) \\ \mathcal{H}_1 : Y &\sim \mathcal{N}(\theta,1), \theta \geq 1. \end{aligned}$$

Here $\mathcal{N}(a,b)$ denotes a Gaussian distribution with mean a and variance b. In this example one seeks to minimize the probability of error under $\mathcal{H}_1$ subject to an upper bound on the probability of error under $\mathcal{H}_0$. For this problem it can be easily shown (see e.g., [5], [6]) that the optimal test does not require any knowledge about the value of θ. Hence, the uncertainty about the parameter θ does not affect the performance of the hypothesis test in this example.

A different approach to dealing with uncertain statistics in decision-making is the robust approach introduced by Huber [2] (see also [7]). Contrary to

universal schemes that seek to achieve the same performance as the schemes with perfect knowledge, these robust schemes have a more modest aim of optimizing the worst-case performance from all possible realizations of the unknown distributions. This approach was introduced in the context of binary hypothesis testing by Huber [2] and Huber and Strassen [8, 9]. A survey of robust techniques can be found in [10] and a more recent survey in [11].

One of the drawbacks of adopting a robust approach is that the performance obtained can be much poorer than the performance with perfect statistical knowledge. In some problems it might be feasible to overcome statistical uncertainty by learning the unknown statistics online. This is especially true in dynamic problems involving optimization of rewards accrued over repeated experiments. A classical example is the problem of adaptive control [12] where one seeks to control a dynamic system with uncertain parameters. Schemes that adapt the control strategies using estimates of the unknown parameters typically perform better than robust control schemes [13] that are designed to optimize worst-case performance. An example of such a problem was the one we studied in [14]. This paper addresses the problem of uncertainties in the received signal statistics in the context of dynamic spectrum access. In [14] it was shown that a robust scheme that optimizes the worst-case performance tends to perform poorly for more favorable realizations of the signal statistics. A learning-based scheme, however, is shown to partially recover this *cost of uncertainty* by learning the parameters online.

1.3 Contribution of This Thesis

In this thesis we study approaches to deal with uncertain statistics in two different contexts. We first study the problems of universal and composite hypothesis testing and some known solutions to these problems - viz. the Hoeffding test and the generalized likelihood ratio test. We identify a new relaxation of the Kullback-Leibler divergence which we call the *mismatched divergence* that plays an important role in these tests. We obtain guidelines for setting thresholds in these tests and results on asymptotic optimality of these tests. We then study a different version of the universal hypothesis testing problem where there is added uncertainty about the null hypothesis. We obtain guidelines for setting thresholds in such problems as well.

The second problem we study is the problem of quickest change detection in the presence of uncertainty about the pre-change and post-change distributions. We obtain a robust solution to this problem under various formulations following an approach similar to Huber's approach to robust hypothesis testing [2].

1.4 Outline

In Chapter 2 we study the problems of universal and composite hypothesis testing for finite alphabet distributions. We then study a robust version of the universal hypothesis testing problem in Chapter 3 which is relevant when there is uncertainty in the observation statistics under the null hypothesis. In Chapter 4 we provide robust solutions to the quickest change detection problem under uncertainty in the observation statistics. In order to improve readability, we have relegated the proofs of many of the results to the appendix. We conclude in Chapter 5.

CHAPTER 2

UNIVERSAL AND COMPOSITE HYPOTHESIS TESTING VIA MISMATCHED DIVERGENCE

2.1 Introduction and Background

This chapter is concerned with the following hypothesis testing problem: Suppose that the observations $\boldsymbol{Z} = \{Z_t : t = 1, \ldots\}$ form an i.i.d. sequence evolving on a set of cardinality N, denoted by $\mathsf{Z} = \{z_1, z_2, \ldots, z_N\}$. Based on observations of this sequence we wish to decide if the marginal distribution of the observations is a given distribution π^0, or some other distribution π^1 that is either unknown or known only to belong to a certain class of distributions. When the observations have distribution π^0 we say that the *null hypothesis* is true, and when the observations have some other distribution π^1 we say that the *alternate hypothesis* is true.

A decision rule is characterized by a *sequence* of tests $\boldsymbol{\phi} := \{\phi_n : n \geq 1\}$, where $\phi_n : \mathsf{Z}^n \mapsto \{0, 1\}$ with Z^n representing the n-th order Cartesian-product of Z. The decision based on the first n elements of the observation sequence is given by $\phi_n(Z_1, Z_2, \ldots, Z_n)$, where $\phi_n = 0$ represents a decision in favor of accepting π^0 as the true marginal distribution.

The set of probability measures on Z is denoted $\mathcal{P}(\mathsf{Z})$. The relative entropy (or Kullback-Leibler divergence) between two distributions $\nu^1, \nu^2 \in \mathcal{P}(\mathsf{Z})$ is denoted $D(\nu^1 \| \nu^2)$, and for a given $\mu \in \mathcal{P}(\mathsf{Z})$ and $\eta > 0$ the *divergence ball* of *radius* η around μ is defined as

$$\mathcal{Q}_\eta(\mu) := \{\nu \in \mathcal{P}(\mathsf{Z}) : D(\nu \| \mu) < \eta\}. \tag{2.1}$$

The empirical distribution or *type* of the first n observations from $\boldsymbol{Z}$ is a random variable Γ^n taking values in $\mathcal{P}(\mathsf{Z})$:

$$\Gamma^n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Z_i = z\}, \qquad z \in \mathsf{Z} \tag{2.2}$$

where $\mathbb{I}$ denotes the indicator function.

In the general universal hypothesis testing problem, the null distribution π^0 is known exactly, but no prior information is available regarding the alternate distribution π^1. Hoeffding proposed in [3] a generalized likelihood-ratio test (GLRT) for the universal hypothesis testing problem, in which the alternate distribution π^1 is unrestricted — it is an arbitrary distribution in $\mathcal{P}(\mathsf{Z})$, the set of probability distributions on Z. Hoeffding's test sequence is given by

$$\phi_n^{\mathrm{H}} = \mathbb{I}\{\sup_{\pi^1 \in \mathcal{P}(\mathsf{Z})} \frac{1}{n} \sum_{i=1}^{n} \log \frac{\pi^1(Z_i)}{\pi^0(Z_i)} \geq \eta\}. \tag{2.3}$$

It is easy to see that the test (2.3) can be rewritten as follows:

$$\begin{aligned} \phi_n^{\mathrm{H}} &= \mathbb{I}\{\frac{1}{n} \sum_{i=1}^{n} \log \frac{\Gamma^n(Z_i)}{\pi^0(Z_i)} \geq \eta\} \\ &= \mathbb{I}\{\sum_{z \in \mathsf{Z}} \Gamma^n(z) \log \frac{\Gamma^n(z)}{\pi^0(z)} \geq \eta\} \\ &= \mathbb{I}\{D(\Gamma^n \| \pi^0) \geq \eta\} \\ &= \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_\eta(\pi^0)\}. \end{aligned} \tag{2.4}$$

We refer to the above test as the Hoeffding test.

If we have some prior information on the alternate distribution π^1, a different version of the GLRT is used. In particular, suppose it is known that the alternate distribution lies in a parametric family of distributions of the following form:

$$\mathcal{E}_{\pi^0} := \{\check{\pi}^r : r \in \mathbb{R}^d\}$$

where $\check{\pi}^r \in \mathcal{P}(\mathsf{Z})$ are probability distributions on Z parameterized by a parameter $r \in \mathbb{R}^d$. The specific form of $\check{\pi}^r$ is defined later in the chapter. In this case, the resulting composite hypothesis testing problem is typically solved using a GLRT (see [4] for results related to the present problem, and [15] for a more recent account) of the following form:

$$\phi_n^{\mathrm{MM}} = \mathbb{I}\{\sup_{\pi^1 \in \mathcal{E}_{\pi^0}} \langle \Gamma^n, \log \frac{\pi^1}{\pi^0} \rangle \geq \eta\} \tag{2.5}$$

where $\langle \Gamma^n, \log \frac{\pi^1}{\pi^0} \rangle = \sum_{z \in \mathsf{Z}} \Gamma^n(z) \log \frac{\pi^1(z)}{\pi^0(z)}$. We show that this test can be

interpreted as a relaxation of the Hoeffding test of (2.4). In particular we show that

$$\phi_n^{\text{MM}} = \mathbb{I}\{D^{\text{MM}}(\Gamma^n \| \pi^0) \geq \eta\} \tag{2.6}$$

where D^{MM} is a relaxation of the K-L divergence. We refer to this quantity as the *mismatched divergence* and the test (2.6) as the *mismatched test*. The mismatched divergence is a lower bound based on a relaxation of the K-L divergence in the sense that $D^{\text{MM}}(\mu\|\pi) \leq D(\mu\|\pi)$ for any $\mu, \pi \in \mathcal{P}(\mathsf{Z})$. We illustrate various properties of the mismatched divergence later in the chapter. Most of the results in this chapter were published in [16] and [17].

The terminology is borrowed from the *mismatched channel* (see Lapidoth [18] for a bibliography). The mismatched divergence described here is a generalization of the relaxation introduced in [19]. In this way we embed the analysis of the resulting universal test within the framework of Csiszár and Shields [20]. The mismatched test statistic can also be viewed as a generalization of the robust hypothesis testing statistic introduced in [21, 22].

When the alternate distribution satisfies $\pi^1 \in \mathcal{E}_{\pi^0}$, we show that, under some regularity conditions on $\mathcal{E}_{\pi^0}$, the mismatched test of (2.6) and Hoeffding's test of (2.4) have identical asymptotic performance in terms of error exponents. A consequence of this result is that the GLRT is optimal in differentiating a particular distribution from others in an exponential family of distributions. We also establish that the proposed mismatched test has a significant advantage over the Hoeffding test in terms of finite sample size performance. This advantage is due to the difference in the asymptotic variances of the two test statistics under the null hypothesis. In particular, we show that the variance of the K-L divergence grows linearly with the alphabet size, making the test impractical for applications involving large alphabet distributions. We also show that the variance of the mismatched divergence grows linearly with the dimension d of the parameter space, and can hence be controlled through a educated choice of the function class defining the mismatched divergence.

The remainder of the chapter is organized as follows. We begin in Section 2.2 with a description of mismatched divergence and the mismatched test, and describe their relation to other concepts including robust hypothesis testing, composite hypothesis testing, reverse I-projection, and maximum likelihood (ML) estimation. Formulae for the asymptotic mean and variance

of the test statistics are presented in Section 2.3. Section 2.3 also contains a discussion interpreting these asymptotic results in terms of the performance of the detection rule. Proofs of the main results are provided in Appendix A. Conclusions and directions for future research are contained in Section 2.5.

2.2 Mismatched Divergence

We adopt the following compact notation in the chapter: For any function $f\colon \mathsf{Z} \to \mathbb{R}$ and $\pi \in \mathcal{P}(\mathsf{Z})$ we denote the mean $\sum_{z\in\mathsf{Z}} f(z)\pi(z)$ by $\pi(f)$, or by $\langle \pi, f\rangle$ when we wish to emphasize the convex-analytic setting. At times we will extend these definitions to allow functions f taking values in a vector space. For $z \in \mathsf{Z}$ and $\pi \in \mathcal{P}(\mathsf{Z})$, we still use $\pi(z)$ to denote the probability assigned to element z under measure π. The meaning of such notation will be clear from context.

The logarithmic moment generating function (log-MGF) is denoted

$$\Lambda_\pi(f) = \log(\pi(\exp(f)))$$

where $\pi(\exp(f)) = \sum_{z\in\mathsf{Z}} \pi(z)\exp(f(z))$ by the notation we introduced in the previous paragraph. For any two probability measures $\nu^1, \nu^2 \in \mathcal{P}(\mathsf{Z})$ the relative entropy is expressed,

$$D(\nu^1\|\nu^2) = \begin{cases} \langle \nu^1, \log(\nu^1/\nu^2)\rangle & \text{if } \nu^1 \prec \nu^2 \\ \infty & \text{else} \end{cases}$$

where $\nu^1 \prec \nu^2$ denotes absolute continuity. The following proposition recalls a well-known variational representation. This can be obtained, for instance, by specializing the representation in [23] to an i.i.d. setting. An alternate variational representation of the divergence is utilized in [24].

Proposition 2.2.1. *The relative entropy can be expressed as the convex dual of the log moment generating function: For any two probability measures $\nu^1, \nu^2 \in \mathcal{P}(\mathsf{Z})$,*

$$D(\nu^1\|\nu^2) = \sup_f \Big(\nu^1(f) - \Lambda_{\nu^2}(f)\Big) \tag{2.7}$$

where the supremum is taken over the space of all real-valued functions on Z. Furthermore, if ν^1 and ν^2 have equal supports, then the supremum is

achieved by the log likelihood ratio function $f^* = \log(\nu^1/\nu^2)$.

Outline of proof. Although the result is well known, we provide a simple proof here since similar arguments will be reused later in the chapter.

For any function f we have,

$$\begin{aligned} D(\nu^1\|\nu^2) &= \langle \nu^1, \log(\nu^1/\nu^2)\rangle \\ &= \langle \nu^1, \log(\nu/\nu^2)\rangle + \langle \nu^1, \log(\nu^1/\nu)\rangle \end{aligned}$$

where $\nu = \nu^2 \exp(f - \Lambda_{\nu^2}(f))$. That is,

$$D(\nu^1\|\nu^2) = \nu^1(f) - \Lambda_{\nu^2}(f) + D(\nu^1\|\nu) \geq \nu^1(f) - \Lambda_{\nu^2}(f).$$

If ν^1 and ν^2 have equal supports, then the above inequality holds with equality for $f = \log(\nu^1/\nu^2)$ which would lead to $\nu = \nu^1$. This proves that (2.7) holds whenever ν^1 and ν^2 have equal supports. The proof for general distributions is similar and is omitted here. □

The representation (2.7) is the basis of the mismatched divergence. We fix a set of functions denoted by $\mathcal{F}$, and obtain a lower bound on the relative entropy by taking the supremum over the smaller set as follows:

$$D^{\text{MM}}(\nu^1\|\nu^2) := \sup_{f\in\mathcal{F}} \Big\{ \nu^1(f) - \Lambda_{\nu^2}(f) \Big\}. \tag{2.8}$$

If ν^1 and ν^2 have full support, and if the function class $\mathcal{F}$ contains the log-likelihood ratio function $f^* = \log(\nu^1/\nu^2)$, then it is immediate from Proposition 2.2.1 that the supremum in (2.8) is achieved by f^*, and in this case $D^{\text{MM}}(\nu^1\|\nu^2) = D(\nu^1\|\nu^2)$. Moreover, since the objective function in (2.8) is invariant to shifts of f, it follows that even if a constant scalar is added to the function f^*, it still achieves the supremum in (2.8).

In this chapter the function class is assumed to be defined through a finite-dimensional parametrization of the form

$$\mathcal{F} = \{f_r : r \in \mathbb{R}^d\}. \tag{2.9}$$

Further assumptions will be imposed in our main results. In particular, we will assume that $f_r(z)$ is differentiable as a function of r for each z.

We fix a distribution $\pi \in \mathcal{P}(\mathsf{Z})$ and a function class of the form (2.9). For each $r \in \mathbb{R}^d$ the *twisted distribution* $\check{\pi}^r \in \mathcal{P}(\mathsf{Z})$ is defined as

$$\check{\pi}^r := \pi \exp(f_r - \Lambda_\pi(f_r)). \tag{2.10}$$

The collection of all such distributions parameterized by r is denoted

$$\mathcal{E}_\pi := \{\check{\pi}^r : r \in \mathbb{R}^d\}. \tag{2.11}$$

2.2.1 Applications

The applications of mismatched divergence include those applications surveyed in Section 3 of [15] in their treatment of generalized likelihood ratio tests. Here we list potential applications in three domains: Hypothesis testing, source coding, and nonlinear filtering. Other applications include channel coding and signal detection, following [15].

Hypothesis testing

The primary motivation for our research is to improve the finite sample size performance of Hoeffding's universal test (2.3). The difficulty we address is the large variance of this test statistic when the alphabet size is large. Theorem 2.2.2 makes this precise:

Theorem 2.2.2. *Let $\pi^0, \pi^1 \in \mathcal{P}(\mathsf{Z})$ have full supports over Z.*

(i) *Suppose that the observation sequence $\boldsymbol{Z}$ is i.i.d. with marginal π^0. Then the normalized Hoeffding test statistic sequence $\{nD(\Gamma^n\|\pi^0) : n \geq 1\}$ has the following asymptotic bias and variance:*

$$\lim_{n\to\infty} \mathsf{E}[nD(\Gamma^n\|\pi^0)] = \tfrac{1}{2}(N-1) \tag{2.12}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD(\Gamma^n\|\pi^0)] = \tfrac{1}{2}(N-1) \tag{2.13}$$

where $N = |\mathsf{Z}|$ denotes the size (cardinality) of Z. Furthermore, the following weak convergence result holds:

$$2nD(\Gamma^n\|\pi^0) \xrightarrow[n\to\infty]{d.} \chi^2_{N-1} \tag{2.14}$$

where the right hand side denotes the chi-squared distribution with $N-1$ degrees of freedom.

(ii) *Suppose the sequence $\boldsymbol{Z}$ is drawn i.i.d. under $\pi^1 \neq \pi^0$. We then have*

$$\lim_{n\to\infty} \mathsf{E}\Big[n\Big(D(\Gamma^n\|\pi^0) - D(\pi^1\|\pi^0)\Big)\Big] = \tfrac{1}{2}(N-1).$$

□

The bias result of (2.12) follows from the unpublished report [25] and the weak convergence result of (2.14) follows from the result of [26]. All the results of the theorem, including (2.13), also follow from Theorem 2.3.2. We elaborate on this in Section 2.3.

We see from Theorem 2.2.2 that the bias of the divergence statistic $D(\Gamma^n\|\pi^0)$ decays as $\frac{N-1}{2n}$, irrespective of whether the observations are drawn from distribution π^0 or π^1. One could argue that the problem of high bias in the Hoeffding test statistic can be addressed by setting a higher threshold. However, we also notice that when the observations are drawn under π^0, the variance of the divergence statistic decays as $\frac{N-1}{2n^2}$, which can be significant when N is of the order of n^2. This is a more serious flaw of the Hoeffding test for large alphabet sizes, since it cannot be addressed as easily. The high variance indicates that the Hoeffding test is not reliable in situations where the alphabet size is of the order of the square of the sequence length.

The weak convergence result in (2.14) and other such results established later in this chapter can be used to set thresholds for a finite sample test, subject to a constraint on the probability of false alarm (see for example, [20, p. 457]). As an application of (2.12) we propose the following approximation for the false alarm probability in the Hoeffding test defined in (2.4):

$$p_{\text{FA}} := \mathsf{P}_{\pi^0}\Big\{\phi_n^{\text{H}} = 1\Big\} \approx \mathsf{P}\Big\{\sum_{i=1}^{N-1} W_i^2 \geq 2n\eta\Big\} \tag{2.15}$$

where $\{W_i\}$ are i.i.d. $N(0,1)$ random variables. In this way we can obtain a simple formula for the threshold to approximately achieve a given constraint on p_{FA}. For moderate values of the sequence length n, the χ^2 approximation gives a more accurate prediction of the false alarm probabilities for the Hoeffding test compared to those predicted using Sanov's theorem as we demonstrate below.

Consider the application of (2.15) in the following example. We used Monte Carlo simulations to approximate the false alarm probability of the Hoeffding test described in (2.4), with π^0 the uniform distribution on an alphabet of size 20. Shown in Figure 2.1 is a semi-log plot comparing three quantities: the probability of false alarm p_{FA}, estimated via simulation; the approximation (2.15) obtained from the Central Limit Theorem; and the approximation obtained from Sanov's Theorem, $\log(p_{\text{FA}}) \approx -n\eta$. It is clearly seen that the approximation based on the weak convergence result of (2.15) is *far more accurate* than the approximation based on Sanov's theorem.

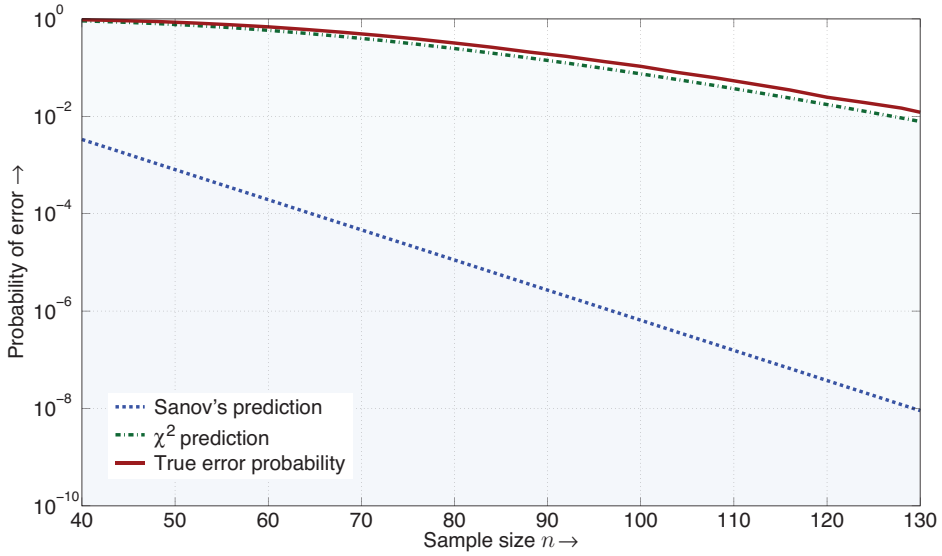

Figure 2.1: *Approximations for the false alarm probability in universal hypothesis testing.* The false alarm probability of the Hoeffding test is closely approximated by the approximation (2.15).

One approach to addressing the implementation issues of the universal test is through clustering (or partitioning) the alphabet as in [27], or smoothing in the space of probability measures as in [28, 29] to extend the Hoeffding test to the case of continuous alphabets. The mismatched test proposed here is a generalization of a partition in the following sense. Suppose that $\{A_i : 1 \le i \le N_a\}$ are disjoint sets satisfying $\cup A_i = \mathsf{X}$, and let $Y(t) = i$ if $X(t) \in A_i$. Applying (2.13), we conclude that the Hoeffding test using $\boldsymbol{Y}$ instead of $\boldsymbol{X}$ will have asymptotic variance equal to $\frac{1}{2}(N_a - 1)$, where $N_a < N$ for a non-trivial partition. We have:

Proposition 2.2.3. *Suppose that the mismatched divergence is defined with respect to the linear function class* (2.26) *using* $\psi_i = \mathbb{I}_{A_i}$, $1 \le i \le N_a$. *In this case the mismatched test* (2.5) *coincides with the Hoeffding test using observations* $\boldsymbol{Y}$. □

The advantage of the mismatched test (2.5) over a partition is that we can incorporate prior knowledge regarding alternate statistics, and we can include non-standard 'priors' such as continuity of the log-likelihood ratio function between the null and alternate distributions.

Source coding with training

Let π denote a source distribution on a finite alphabet Z. Suppose we do not know π exactly and we design optimal codelengths assuming that the distribution is μ: For letter $z \in \mathsf{Z}$ we let $\ell(z) = -\log(\mu(z))$ denote Shannon's codeword length. The expected codelength is thus

$$\mathsf{E}[\ell] = \sum_{z \in \mathsf{Z}} \ell(z)\pi(z) = H(\pi) + D(\pi \| \mu)$$

where H denotes the entropy, $-\sum_{z \in \mathsf{Z}} \pi(z) \log(\pi(z))$. Let $\ell^* := H(\pi)$ denote the optimal (minimal) expected codelength.

Now suppose it is known that under π the probability of each letter $z \in \mathsf{Z}$ is bounded away from zero. That is, we assume that for some $\epsilon > 0$,

$$\pi \in \mathbb{P}_\epsilon := \{\mu \in \mathcal{P}(\mathsf{Z}) : \mu(z) > \epsilon, \text{ for all } z \in \mathsf{Z}\}.$$

Further suppose that a training sequence of length n is given, drawn under π. We are interested in constructing a source code for encoding symbols from the source π based on these training symbols. Let Γ^n denote the type of the observations based on these n training symbols. We assign codeword lengths to each symbol z according to the following rule:

$$\ell(z) = \begin{cases} \log \frac{1}{\Gamma^n(z)} & \text{if } \Gamma^n \in \mathbb{P}_{\epsilon/2} \\ \log \frac{1}{\pi^{\mathrm{u}}(z)} & \text{else} \end{cases}$$

where π^{u} is the uniform distribution on Z.

For such a system, the conditional expected codelength conditioned on the

training symbols denoted by $\mathcal{T}$ satisfies

$$\mathsf{E}[\ell^n|\mathcal{T}] = \begin{cases} \ell^* + D(\pi\|\Gamma^n) & \text{if } \Gamma^n \in \mathbb{P}_{\epsilon/2} \\ \ell^* + D(\pi\|\pi^{\mathrm{u}}) & \text{else.} \end{cases}$$

We study the behavior of $\mathsf{E}[\ell^n - \ell^*|\mathcal{T}]$ as a function of n. We argue in Appendix A that a modification of our results from Theorem 2.3.2 can be used to establish the following relations:

$$\begin{aligned} 2n(\mathsf{E}[\ell^n|\mathcal{T}] - \ell^*) &\xrightarrow[n\to\infty]{d.} \chi^2_{N-1} \\ \mathsf{E}[n(\ell^n - \ell^*)] &\xrightarrow[n\to\infty]{} \tfrac{1}{2}(N-1) \\ \mathsf{Var}\left[n\mathsf{E}[\ell^n|\mathcal{T}]\right] &\xrightarrow[n\to\infty]{} \tfrac{1}{2}(N-1) \end{aligned} \tag{2.16}$$

where N is the cardinality of the alphabet Z. Comparing with Theorem 2.2.2 we conclude that the asymptotic behavior of the excess codelength is identical to the asymptotic behavior of the Hoeffding test statistic $D(\Gamma^n\|\pi)$ under π. Methods such as those proposed in this chapter can be used to reduce high variance, just as in the hypothesis testing problem emphasized in this chapter.

Filtering

The recent paper [30] considers approximations for the nonlinear filtering problem. Suppose that $\boldsymbol{X}$ is a Markov chain on $\mathbb{R}^n$, and $\boldsymbol{Y}$ is an associated observation process on $\mathbb{R}^p$ of the form $Y(t) = \gamma(X(t), W(t))$, where $\boldsymbol{W}$ is an i.i.d. sequence. The conditional distribution of $X(t)$ given $\{Y(0), \ldots, Y(t)\}$ is denoted B_t — it is known as the *belief state* in this literature. The evolution of the belief state can expressed in the recursive form,

$$B_{t+1} = \phi(B_t, Y_{t+1}), \qquad t \geq 0.$$

For some mapping $\phi\colon \mathcal{B}(\mathbb{R}^n) \times \mathbb{R}^p \to \mathcal{B}(\mathbb{R}^n)$.

The approximation proposed in [30] is based on a projection of B_t onto an exponential family of densities over $\mathbb{R}^n$, of the form

$$p_\theta(x) = p_0(x)\exp(\theta^{\mathrm{T}}\psi(x) - \Lambda(\theta)), \quad \theta \in \mathbb{R}^d.$$

They consider the *reverse I-projection*,

$$B^{\varrho} = \arg\min_{\mu\in\mathcal{E}} D(B\|\mu)$$

where the minimum is over $\mathcal{E} = \{p_\theta\}$. From the definition of divergence this is equivalently expressed,

$$B^{\varrho} = \arg\max_{\theta} \int \Big(\theta^{\scriptscriptstyle T}\psi(x) - \Lambda(\theta)\Big)\, B(dx). \tag{2.17}$$

A projected filter is defined by the recursion,

$$\widehat{B}_{t+1} = [\phi(\widehat{B}_t, Y_{t+1})]^{\varrho}, \qquad t \geq 0. \tag{2.18}$$

The techniques in the current chapter provide algorithms for computation of this projection, and suggest alternative projection schemes, such as the robust approach described in Section 2.2.6.

2.2.2 Basic structure of mismatched divergence

The mismatched test is defined to be a relaxation of the Hoeffding test described in (2.4). We replace the divergence functional with the mismatched divergence $D^{\text{MM}}(\Gamma^n\|\pi^0)$. Thus the mismatched test sequence is given by

$$\phi_n^{\text{MM}} = \mathbb{I}\{D^{\text{MM}}(\Gamma^n\|\pi^0) \geq \eta\} = \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_\eta^{\text{MM}}(\pi^0)\} \tag{2.19}$$

where $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ is the mismatched divergence ball of radius η around π^0 defined analogously to (2.1):

$$\mathcal{Q}_\eta^{\text{MM}}(\mu) = \{\nu \in \mathcal{P}(\mathsf{Z}) : D^{\text{MM}}(\nu\|\mu) < \eta\}. \tag{2.20}$$

The next proposition establishes some basic geometry of the mismatched divergence balls. For any function g we define the following hyperplane and half-space:

$$\begin{aligned} \mathcal{H}_g &:= \{\nu : \nu(g) = 0\} \\ \mathcal{H}_g^- &:= \{\nu : \nu(g) < 0\}. \end{aligned} \tag{2.21}$$

Proposition 2.2.4. *The following hold for any $\nu, \pi \in \mathcal{P}(\mathsf{Z})$, and any collection of functions $\mathcal{F}$:*

(i) *For each* $\eta > 0$ *we have* $\mathcal{Q}_\eta^{\text{MM}}(\pi) \subset \bigcap \mathcal{H}_g^-$*, where the intersection is over all functions* g *of the form,*

$$g = f - \Lambda_\pi(f) - \eta \tag{2.22}$$

with $f \in \mathcal{F}$*.*

(ii) *Suppose that* $\eta = D^{\text{MM}}(\nu\|\pi)$ *is finite and non-zero. Further suppose that for* $\nu^1 = \nu$ *and* $\nu^2 = \pi$*, the supremum in* (2.8) *is achieved by* $f^* \in \mathcal{F}$*. Then* $\mathcal{H}_{g^*}$ *is a supporting hyperplane to* $\mathcal{Q}_\eta^{\text{MM}}(\pi)$*, where* g^* *is given in* (2.22) *with* $f = f^*$*.*

Proof. (i) Suppose $\mu \in \mathcal{Q}_\eta^{\text{MM}}(\pi)$. Then, for any $f \in \mathcal{F}$,

$$\mu(f) - \Lambda_\pi(f) - \eta \le D^{\text{MM}}(\mu\|\pi) - \eta < 0.$$

That is, for any $f \in \mathcal{F}$, on defining g by (2.22) we obtain the desired inclusion $\mathcal{Q}_\eta^{\text{MM}}(\pi) \subset \mathcal{H}_g^-$.

(ii) Let $\mu \in \mathcal{H}_{g^*}$ be arbitrary. Then we have:

$$\begin{aligned} D^{\text{MM}}(\mu\|\pi) &= \sup_r \Big(\mu(f_r) - \Lambda_\pi(f_r)\Big) \\ &\ge \mu(f^*) - \Lambda_\pi(f^*) = \Lambda_\pi(f^*) + \eta - \Lambda_\pi(f^*) = \eta. \end{aligned}$$

Hence it follows that $\mathcal{H}_{g^*}$ supports $\mathcal{Q}_\eta^{\text{MM}}(\pi)$ at ν.

□

2.2.3 Asymptotic optimality of the mismatched test

The asymptotic performance of a binary hypothesis testing problem is typically characterized in terms of error exponents. We adopt the following criterion for performance evaluation, following Hoeffding [3] (and others, notably [28, 29]). Suppose that the observations $\boldsymbol{Z} = \{Z_t : t = 1, \ldots\}$ form an i.i.d. sequence evolving on Z. For a given π^0, and a given alternate distribution π^1, the type I and type II error exponents are denoted respectively

by

$$J_\phi^0 := \liminf_{n\to\infty} -\frac{1}{n}\log(\mathsf{P}_{\pi^0}\{\phi_n = 1\}),$$
$$J_\phi^1 := \liminf_{n\to\infty} -\frac{1}{n}\log(\mathsf{P}_{\pi^1}\{\phi_n = 0\}) \tag{2.23}$$

where in the first limit the marginal distribution of Z_t is π^0, and in the second it is π^1. The limit J_ϕ^0 is also called the false-alarm error exponent, and J_ϕ^1 the missed-detection error exponent.

For a given constraint $\eta > 0$ on the false-alarm exponent J_ϕ^0, an optimal test is the solution to the asymptotic Neyman-Pearson hypothesis testing problem,

$$\beta^*(\eta) = \sup\{J_\phi^1 : \ \textit{subject to } J_\phi^0 \geq \eta\} \tag{2.24}$$

where the supremum is over all allowed test sequences $\boldsymbol{\phi}$. While the exponent $\beta^*(\eta) = \beta^*(\eta, \pi^1)$ depends upon π^1, Hoeffding's test we described in (2.4) does not require knowledge of π^1, yet achieves the optimal exponent $\beta^*(\eta, \pi^1)$ for any π^1. The optimality of Hoeffding's test established in [3] easily follows from Sanov's theorem.

While the mismatched test described in (2.6) is not always optimal for (2.24) for a general choice of π^1, it is optimal for some specific choices of the alternate distributions. The following corollary to Proposition 2.2.4 captures this idea.

Corollary 2.2.5. *Suppose $\pi^0, \pi^1 \in \mathcal{P}(\mathsf{Z})$ have equal supports. Let $\varrho \in (0,1)$ be chosen so as to guarantee $D(\check{\pi}\|\pi^0) = \eta$ where $\check{\pi}$ is the twisted distribution defined by $\check{\pi} = \kappa(\pi^0)^{1-\varrho}(\pi^1)^{\varrho}$ with κ a normalizing constant. Further suppose that there exists $\tau \in \mathbb{R}$ and $r \in \mathbb{R}^d$ such that*

$$\varrho L(z) + \tau = f_r(z) \quad a.e. \quad [\pi^0],$$

where L is the log likelihood-ratio function $L := \log(\pi^1/\pi^0)$. Then the mismatched test is optimal in the sense that the constraint $J_{\phi^{\text{MM}}}^0 \geq \eta$ is satisfied with equality, and under π^1 the optimal error exponent $J_{\phi^{\text{MM}}}^1 = \beta^*(\eta)$ is achieved.

Proof. Suppose that the conditions stated in the corollary hold. Consider the twisted distribution $\check{\pi}$ It is known that the hyperplane $\mathcal{H}^{\text{LLR}} := \{\nu : \nu(L) = \check{\pi}(L)\}$ separates the divergence balls $\mathcal{Q}_\eta(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$ at $\check{\pi}$. This geometry, which is implicit in [28], is illustrated in Figure 2.2.

From the form of $\check{\pi}$ it is also clear that

$$\log \frac{\check{\pi}}{\pi^0} = \varrho L - \Lambda_{\pi^0}(\varrho L).$$

Hence it follows by Proposition 2.2.1 that the supremum in the variational representation of $D(\check{\pi}\|\pi^0)$ is achieved by ϱL. Furthermore, since $\varrho L + \tau \in \mathcal{F}$ for some $\tau \in \mathbb{R}$ we have

$$D^{\text{MM}}(\check{\pi}\|\pi^0) = D(\check{\pi}\|\pi^0) = \eta = \check{\pi}(\varrho L + \tau) - \Lambda_{\pi^0}(\varrho L + \tau) = \check{\pi}(\varrho L) - \Lambda_{\pi^0}(\varrho L).$$

This means that $\mathcal{H}^{\text{LLR}} = \{\nu : \nu(\varrho L - \Lambda_{\pi^0}(\varrho L) - \eta) = 0\}$. Hence, by applying Proposition 2.2.4 (ii) it follows that the hyperplane $\mathcal{H}^{\text{LLR}}$ separates $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$. This in particular means that the sets $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$ are disjoint. This fact, together with Sanov's theorem proves the corollary. □

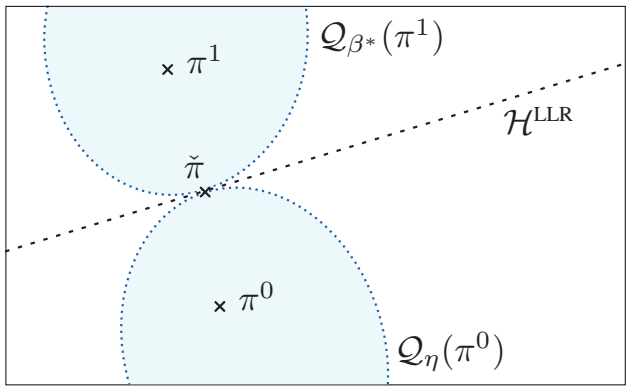

Figure 2.2: *Geometric interpretation of the log likelihood ratio test.* The exponent $\beta^* = \beta^*(\eta)$ is the largest constant satisfying $\mathcal{Q}_\eta(\pi^0) \cap \mathcal{Q}_{\beta^*}(\pi^1) = \emptyset$. The hyperplane $\mathcal{H}^{\text{LLR}} := \{\nu : \nu(L) = \check{\pi}(L)\}$ separates the convex sets $\mathcal{Q}_\eta(\pi^0)$ and $\mathcal{Q}_{\beta^*}(\pi^1)$.

The corollary indicates that while using the mismatched test in practice, the function class might be chosen to include approximations to scaled versions of the log-likelihood ratio functions of the anticipated alternate distributions $\{\pi^1\}$ with respect to π^0.

The mismatched divergence has several equivalent characterizations. We first relate it to an ML estimate from a parametric family of distributions.

2.2.4 Mismatched divergence and ML estimation

On interpreting $f_r - \Lambda_\pi(f_r)$ as a log-likelihood ratio function we obtain in Proposition 2.2.6 the following representation of mismatched divergence:

$$D^{\text{MM}}(\mu\|\pi) = \sup_{r\in\mathbb{R}^d} \Big(\mu(f_r) - \Lambda_\pi(f_r)\Big) = D(\mu\|\pi) - \inf_{\nu\in\mathcal{E}_\pi} D(\mu\|\nu). \tag{2.25}$$

The infimum on the RHS of (2.25) is known as *reverse I-projection* [20]. Proposition 2.2.7 that follows uses this representation to obtain other interpretations of the mismatched test.

Proposition 2.2.6. *The identity* (2.25) *holds for any function class $\mathcal{F}$. The supremum is achieved by some $r^* \in \mathbb{R}^d$ if and only if the infimum is attained at $\nu^* = \check{\pi}^{r^*} \in \mathcal{E}_\pi$. If a minimizer ν^* exists, we obtain the* generalized Pythagorean identity*,*

$$D(\mu\|\pi) = D^{\text{MM}}(\mu\|\pi) + D(\mu\|\nu^*).$$

Proof. For any r we have $\mu(f_r) - \Lambda_\pi(f_r) = \mu(\log(\check{\pi}^r/\pi))$. Consequently,

$$\begin{aligned} D^{\text{MM}}(\mu\|\pi) &= \sup_r \Big(\mu(f_r) - \Lambda_\pi(f_r)\Big) \\ &= \sup_r \mu\left(\log\left(\frac{\mu}{\pi}\frac{\check{\pi}^r}{\mu}\right)\right) \\ &= \sup_r \left\{D(\mu\|\pi) - D(\mu\|\check{\pi}^r)\right\}. \end{aligned}$$

This proves the identity (2.25), and the remaining conclusions follow directly. □

The representation of Proposition 2.2.6 invites the interpretation of the optimizer in the definition of the mismatched test statistic in terms of an ML estimate. Given the well-known correspondence between maximum-likelihood estimation and the generalized likelihood ratio test (GLRT), Proposition 2.2.7 implies that the mismatched test is a special case of the GLRT analyzed in [4].

Proposition 2.2.7. *Suppose that the observations $\boldsymbol{Z}$ are modeled as an i.i.d. sequence, with marginal in the family $\mathcal{E}_\pi$. Let $\hat{r}^n$ denote the ML estimate of*

r based on the first n samples,

$$\hat{r}^n \in \arg\max_{r\in\mathbb{R}^d} \mathsf{P}_{\check{\pi}^r}\{Z_1 = a_1, Z_2 = a_2, \ldots, Z_n = a_n\} = \arg\max_{r\in\mathbb{R}^d} \Pi_{i=1}^n \check{\pi}^r(a_i)$$

where a_i indicates the observed value of the i-th symbol. Assuming the maximum is attained we have the following interpretations:

(i) *The distribution $\check{\pi}^{\hat{r}^n}$ solves the reverse I-projection problem,*

$$\check{\pi}^{\hat{r}^n} \in \arg\min_{\nu\in\mathcal{E}_\pi} D(\Gamma^n\|\nu).$$

(ii) *The function $f^* = f_{\hat{r}^n}$ achieves the supremum that defines the mismatched divergence, $D^{\text{MM}}(\Gamma^n\|\pi) = \Gamma^n(f^*) - \Lambda_\pi(f^*)$.*

Proof. The ML estimate can be expressed $\hat{r}^n = \arg\max_{r\in\mathbb{R}^d}\langle\Gamma^n, \log\check{\pi}^r\rangle$, and hence (i) follows by the identity

$$\arg\min_{\nu\in\mathcal{E}_\pi} D(\Gamma^n\|\nu) = \arg\max_{\nu\in\mathcal{E}_\pi}\langle\Gamma^n, \log\nu\rangle, \qquad \nu\in\mathcal{P}.$$

Combining the result of part (i) with Proposition 2.2.6 we get the result of part (ii). ☐

From conclusions of Proposition 2.2.6 and Proposition 2.2.7 we have

$$\begin{aligned} D^{\text{MM}}(\Gamma^n\|\pi) &= \langle\Gamma^n, \log\frac{\check{\pi}^{\hat{r}^n}}{\pi}\rangle \\ &= \max_{\nu\in\mathcal{E}_\pi}\langle\Gamma^n, \log\frac{\nu}{\pi}\rangle \\ &= \max_{\nu\in\mathcal{E}_\pi}\frac{1}{n}\sum_{i=1}^n \log\frac{\nu(Z_i)}{\pi(Z_i)}. \end{aligned}$$

In general when the supremum in the definition of $D^{\text{MM}}(\Gamma^n\|\pi)$ may not be achieved, the maxima in the above equations are replaced with suprema and we have the following identity:

$$D^{\text{MM}}(\Gamma^n\|\pi) = \sup_{\nu\in\mathcal{E}_\pi}\frac{1}{n}\sum_{i=1}^n \log\frac{\nu(Z_i)}{\pi(Z_i)}.$$

Thus the test statistic used in the mismatched test of (2.6) is exactly the generalized likelihood ratio between the family of distributions $\mathcal{E}_{\pi^0}$ and π^0

where

$$\mathcal{E}_{\pi^0} = \{\pi^0 \exp(f_r - \Lambda_{\pi^0}(f_r)) : r \in \mathbb{R}^d\}.$$

As an immediate consequence of this result and Corollary 2.2.5 we obtain the following sufficient condition for optimality of the GLRT.

Theorem 2.2.8. *Let π^0 be some probability distribution over a finite set Z and η a positive constant. Let $\mathcal{F}$ be a function class and $\mathcal{E}_{\pi^0}$ be the associated parameterized family of distributions defined in (2.11). Suppose that for every $\pi^1 \in \mathcal{E}_{\pi^0}$, we have $\check{\pi} \in \mathcal{E}_{\pi^0}$ where $\check{\pi}$ is the twisted probability distribution defined by $\check{\pi} = \kappa(\pi^0)^{1-\varrho}(\pi^1)^{\varrho}$ with κ a normalizing constant.*

Consider the generalized likelihood ratio test (GLRT) between π^0 and $\mathcal{E}_{\pi^0}$ defined by the following sequence of decision rules:

$$\phi_n^{\text{GLRT}} = \mathbb{I}\{\sup_{\nu \in \mathcal{E}_{\pi^0}} \frac{1}{n} \sum_{i=1}^{n} \log \frac{\nu(Z_i)}{\pi^0(Z_i)} \geq \eta\}.$$

The GLRT solves the composite hypothesis testing problem (2.24) for all $\pi^1 \in \mathcal{E}_{\pi^0}$ in the sense that the constraint $J^0_{\phi^{\text{GLRT}}} \geq \eta$ is satisfied with equality, and under π^1 the optimal error exponent $\beta^*(\eta)$ is achieved for all $\pi^1 \in \mathcal{E}_{\pi^0}$; i.e., $J^1_{\phi^{\text{GLRT}}} = \beta^*(\eta)$.

Proof. From the earlier discussion, we know that the GLRT between π^0 and $\mathcal{E}_{\pi^0}$ is identical to the mismatched test based on $\mathcal{F}$. The conclusion of the theorem then follows directly from Corollary 2.2.5 from the straightforward equivalence relation:

$$\begin{gathered} \check{\pi} \in \mathcal{E}_{\pi^0} \\ \Updownarrow \\ \exists \tau \in \mathbb{R}, r \in \mathbb{R}^d \quad \text{such that} \quad \varrho L(z) + \tau = f_r(z) \quad a.e. \quad [\pi^0] \end{gathered}$$

where $L(z) = \log \frac{\pi^1(z)}{\pi^0(z)}$. ∎

The sufficient condition established above is a restatement of [4, Thm 2, p. 1600] and our approach provides an alternate proof for this result.

More structure can be established when the function class is linear.

2.2.5 Linear function class and I-projection

The mismatched divergence introduced in [19] was restricted to a linear function class. Let $\{\psi_i : 1 \le i \le d\}$ denote d functions on Z. Let $\psi = (\psi_1, \dots, \psi_d)^{\scriptscriptstyle T}$ and let $f_r = r^{\scriptscriptstyle T}\psi$ in the definition (2.9):

$$\mathcal{F} = \Big\{ f_r = \sum_{i=1}^{d} r_i \psi_i : r \in \mathbb{R}^d \Big\}. \tag{2.26}$$

A linear function class is particularly appealing because the optimization problem 2.8 used to define the mismatched divergence becomes a convex program and hence is easy to evaluate in practice. Furthermore, for such a linear function class, the collection of twisted distributions $\mathcal{E}_\pi$ defined in (2.11) forms an *exponential family* of distributions.

$$\mathcal{E}_\pi = \{\pi \exp(r^{\scriptscriptstyle T}\psi - \Lambda_\pi(r^{\scriptscriptstyle T}\psi)) : r \in \mathbb{R}^d\}. \tag{2.27}$$

A special case of such an exponential family is a graphical model for binary vector distributions considered in [31].

Proposition 2.2.6 expresses $D^{\text{MM}}(\mu\|\pi)$ as a difference between the ordinary divergence and the value of a reverse I-projection $\inf_{\nu\in\mathcal{E}_\pi} D(\mu\|\nu)$. The next result establishes a characterization in terms of a (forward) I-projection. For a given vector $c \in \mathbb{R}^d$ we let $\mathbb{P}$ denote the *moment class*

$$\mathbb{P} = \{\nu \in \mathcal{P}(\mathsf{Z}) : \nu(\psi) = c\} \tag{2.28}$$

where $\nu(\psi) = (\nu(\psi_1), \nu(\psi_2), \dots, \nu(\psi_d))^{\scriptscriptstyle T}$.

Proposition 2.2.9. *Suppose that the supremum in the definition of $D^{\text{MM}}(\mu\|\pi)$ is achieved at some $r^* \in \mathbb{R}^d$. Then,*

(i) *The distribution $\nu^* := \check{\pi}^{r^*} \in \mathcal{E}_\pi$ satisfies*

$$D^{\text{MM}}(\mu\|\pi) = D(\nu^*\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\}$$

where $\mathbb{P}$ is defined using $c = \mu(\psi)$ in (2.28).

(ii) *$D^{\text{MM}}(\mu\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathcal{H}_{g^*}\}$, where g^* is given in (2.22) with $f = r^{*\,{\scriptscriptstyle T}}\psi$, and $\eta = D^{\text{MM}}(\mu\|\pi)$.*

Proof. Since the supremum is achieved, the gradient must vanish by the first order condition for optimality:

$$\nabla\Big(\mu(f_r) - \Lambda_\pi(f_r)\Big)\Big|_{r=r^*} = 0.$$

The gradient is computable, and the identity above can thus be expressed $\mu(\psi) - \check{\pi}^{r^*}(\psi) = 0$. That is, the first order condition for optimality is equivalent to the constraint $\check{\pi}^{r^*} \in \mathbb{P}$. Consequently,

$$\begin{aligned} D(\nu^*\|\pi) &= \langle \check{\pi}^{r^*}, \log \frac{\check{\pi}^{r^*}}{\pi} \rangle \\ &= \check{\pi}^{r^*}(r^{*T}\psi) - \Lambda_\pi(r^{*T}\psi) \\ &= \mu(r^{*T}\psi) - \Lambda_\pi(r^{*T}\psi) = D^{\text{MM}}(\mu\|\pi). \end{aligned}$$

Furthermore, by the convexity of $\Lambda_\pi(f_r)$ in r, it follows that the optimal r^* in the definition of $D^{\text{MM}}(\nu\|\pi)$ is the same for all $\nu \in \mathbb{P}$. Hence, it follows by the Pythagorean equality of Proposition 2.2.6 that

$$D(\nu\|\pi) = D(\nu\|\nu^*) + D(\nu^*\|\pi), \text{ for all } \nu \in \mathbb{P}.$$

Minimizing over $\nu \in \mathbb{P}$ it follows that ν^* is the I-projection of π onto $\mathbb{P}$:

$$D(\nu^*\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\}$$

which gives (i).

To establish (ii), note first that by (i) and the inclusion $\mathbb{P} \subset \mathcal{H}_{g^*}$ we have

$$D^{\text{MM}}(\mu\|\pi) = \min\{D(\nu\|\pi) : \nu \in \mathbb{P}\} \geq \inf\{D(\nu\|\pi) : \nu \in \mathcal{H}_{g^*}\}.$$

The reverse inequality follows from Proposition 2.2.4 (i), and moreover the infimum is achieved with ν^*. □

The geometry underlying mismatched divergence for a linear function class is illustrated in Figure 2.3. Suppose that the assumptions of Proposition 2.2.9 hold, so that the supremum in (2.25) is achieved at r^*. Let $\eta = D^{\text{MM}}(\mu\|\pi) = \mu(f_{r^*}) - \Lambda_\pi(f_{r^*})$, and $g^* = f_{r^*} - \Big(\eta + \Lambda_\pi(f_{r^*})\Big)$. Proposition 2.2.4 implies that $\mathcal{H}_{g^*}$ defines a hyperplane passing through μ, with $\mathcal{Q}_\eta(\pi) \subset \mathcal{Q}_\eta^{\text{MM}}(\pi) \subset \mathcal{H}_{g^*}^-$. This is strengthened in the linear case by Proposition 2.2.9, which states that

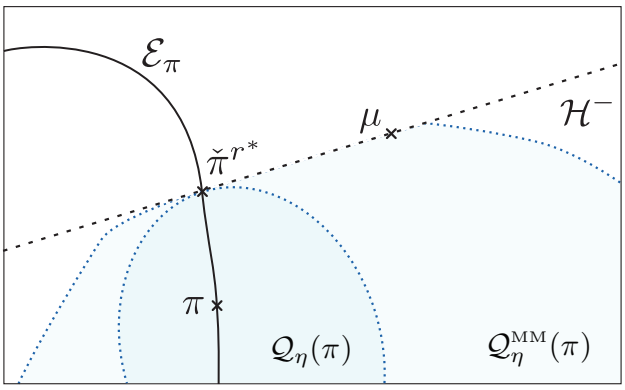

Figure 2.3: *Interpretations of the mismatched divergence for a linear function class.* The distribution $\check{\pi}^{r^*}$ is the I-projection of π onto a hyperplane $\mathcal{H}_{g^*}$. It is also the reverse I-projection of μ onto the exponential family $\mathcal{E}_\pi$.

$\mathcal{H}_{g^*}$ supports $\mathcal{Q}_\eta(\pi)$ at the distribution $\check{\pi}^{r^*}$. Furthermore Proposition 2.2.6 asserts that the distribution $\check{\pi}^{r^*}$ minimizes $D(\mu\|\check{\pi})$ over all $\check{\pi} \in \mathcal{E}_\pi$.

As a special case of Theorem 2.2.8, it can be shown that the GLRT is asymptotically optimal for an exponential family of distributions. For $\pi^1 \in \mathcal{E}_{\pi^0}$, where $\mathcal{E}_{\pi^0}$ is an exponential family of the form in (2.27), it is clear that any probability distribution of the form $\check{\pi} = \kappa(\pi^0)^{1-\varrho}(\pi^1)^\varrho$ lies in the exponential family $\mathcal{E}_{\pi^0}$. Hence the requirement of Theorem 2.2.8 is satisfied and the conclusion follows.

2.2.6 Log-linear function class and robust hypothesis testing

In the prior work [21, 22] the following relaxation of entropy is considered:

$$D^{\text{ROB}}(\mu\|\pi) := \inf_{\nu\in\mathbb{P}} D(\mu\|\nu) \tag{2.29}$$

where the moment class $\mathbb{P}$ is defined in (2.28) with $c = \pi(\psi)$, for a given collection of functions $\{\psi_i : 1 \le i \le d\}$. The associated universal test solves a min-max robust hypothesis testing problem.

We show here that D^{ROB} coincides with D^{MM} for a particular function class.

It is described as (2.9) in which each function f_r is of the log-linear form,

$$f_r = \log(1 + r^{\scriptscriptstyle T}\psi) \tag{2.30}$$

subject to the constraint that $1 + r^{\scriptscriptstyle T}\psi(z)$ is strictly positive for each z. We further require that the functions ψ have zero mean under distribution π; i.e., we require $\pi(\psi) = 0$.

Proposition 2.2.10. *For a given $\pi \in \mathcal{P}(\mathsf{Z})$, suppose that the log-linear function class $\mathcal{F}$ is chosen with functions $\{\psi_i\}$ satisfying $\pi(\psi) = 0$. Suppose that the moment class used in the definition of D^{ROB} is chosen consistently, with $c = 0$ in (2.28). We then have for each $\mu \in \mathcal{P}(\mathsf{Z})$,*

$$D^{\text{MM}}(\mu\|\pi) = D^{\text{ROB}}(\mu\|\pi).$$

Proof. For each $\mu \in \mathcal{P}(\mathsf{Z})$, we obtain the following identity by applying Theorem 1.4 in [22]:

$$\inf_{\nu\in\mathbb{P}} D(\mu\|\nu) = \sup\{\mu(\log(1 + r^{\scriptscriptstyle T}\psi)) : 1 + r^{\scriptscriptstyle T}\psi(z) > 0 \text{ for all } z \in \mathsf{Z}\}.$$

Moreover, under the assumption that $\pi(\psi) = 0$ we obtain

$$\Lambda_\pi(\log(1 + r^{\scriptscriptstyle T}\psi)) = \log(\pi(1 + r^{\scriptscriptstyle T}\psi)) = 0.$$

Combining these identities gives

$$\begin{aligned} D^{\text{ROB}}(\mu\|\pi) &:= \inf_{\nu\in\mathbb{P}} D(\mu\|\nu) \\ &= \sup\Big\{\mu(\log(1 + r^{\scriptscriptstyle T}\psi)) - \Lambda_\pi(\log(1 + r^{\scriptscriptstyle T}\psi)) : \\ &\qquad\qquad 1 + r^{\scriptscriptstyle T}\psi(z) > 0 \text{ for all } z \in \mathsf{Z}\Big\} \\ &= \sup_{f\in\mathcal{F}}\Big\{\mu(f) - \Lambda_\pi(f)\Big\} = D^{\text{MM}}(\mu\|\pi). \end{aligned}$$

□

More properties of the mismatched divergence, including a generalization of the Pinsker's inequality can be found in [31].

2.3 Asymptotic Statistics

In this section, we analyze the asymptotic statistics of the mismatched test. We require some assumptions regarding the function class $\mathcal{F} = \{f_r : r \in \mathbb{R}^d\}$ to establish these results. Note that the second and third assumptions given below involve a distribution $\mu^0 \in \mathcal{P}(\mathsf{Z})$, and a vector $s \in \mathbb{R}^d$. We will make specialized versions of these assumptions in establishing our results, based on specific values of μ^0 and s. We use $\mathsf{Z}_{\mu^0} \subset \mathsf{Z}$ to denote the support of μ^0 and $\mathcal{P}(\mathsf{Z}_{\mu^0})$ to denote the space of probability measures supported on Z_{μ^0}, viewed as a subset of $\mathcal{P}(\mathsf{Z})$.

Assumptions

(A1) $f_r(z)$ is C^2 in r for each $z \in \mathsf{Z}$.

(A2) There exists an open neighborhood $B \subset \mathcal{P}(\mathsf{Z}_{\mu^0})$ of μ^0 such that for each $\mu \in B$, the supremum in the definition of $D^{\text{MM}}(\mu\|\mu^0)$ in (2.8) is achieved at a unique point $r(\mu)$.

(A3) The vectors $\{\psi_0, \dots, \psi_d\}$ are linearly independent over the support of μ^0, where $\psi_0 \equiv 1$, and for each $i \geq 1$

$$\psi_i(z) = \frac{\partial}{\partial r_i} f_r(z)\Big|_{r=s}, \qquad z \in \mathsf{Z}. \tag{2.31}$$

The linear-independence assumption in (A3) is defined as follows: If there are constants $\{a_0, \dots, a_d\}$ satisfying $\sum_{i=1}^d a_i\psi_i(z) = 0$ a.e. $[\mu^0]$, then $a_i = 0$ for each i. In the case of a linear function class, the functions $\{\psi_i, i \geq 1\}$ defined in (2.31) are just the basis functions in (2.26). Lemma 2.3.1 provides an alternate characterization of Assumption (A3).

For any $\mu \in \mathcal{P}(\mathsf{Z})$ define the covariance matrix Σ_μ via

$$\Sigma_\mu(i,j) = \mu(\psi_i\psi_j) - \mu(\psi_i)\mu(\psi_j), \qquad 1 \leq i,j \leq d. \tag{2.32}$$

We use $\mathsf{Cov}_\mu(g)$ to denote the covariance of an arbitrary real-valued function g under μ:

$$\mathsf{Cov}_\mu(g) := \mu(g^2) - \mu(g)^2. \tag{2.33}$$

Lemma 2.3.1. *Assumption (A3) holds if and only if* $\Sigma_{\mu^0} > 0$.

Proof. We evidently have $v^T \Sigma_{\mu^0} v = \mathsf{Cov}_{\mu^0}(v^T \psi) \geq 0$ for any vector $v \in \mathbb{R}^d$. Hence, we have the following equivalence: For any $v \in \mathbb{R}^d$, on denoting $c_v = \mu^0(v^T \psi)$,

$$v^T \Sigma_{\mu^0} v = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{d} v_i \psi_i(z) = c_v \quad \text{a.e. } [\mu^0].$$

The conclusion of the lemma follows. □

We now present our main asymptotic results. Theorem 2.3.2 identifies the asymptotic bias and variance of the mismatched test statistic under the null hypothesis, and also under the alternate hypothesis. A key observation is that the asymptotic bias and variance does not depend on N, the cardinality of Z.

Theorem 2.3.2. *Suppose that the observation sequence $\boldsymbol{Z}$ is i.i.d. with marginal π. Suppose that there exists r^* satisfying $f_{r^*} = \log(\pi/\pi^0)$. Further, suppose that Assumptions (A1), (A2), (A3) hold with $\mu^0 = \pi$ and $s = r^*$. Then,*

(i) *When $\pi = \pi^0$,*

$$\lim_{n\to\infty} \mathsf{E}[nD^{\text{MM}}(\Gamma^n \| \pi^0)] = \tfrac{1}{2}d \tag{2.34}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD^{\text{MM}}(\Gamma^n \| \pi^0)] = \tfrac{1}{2}d \tag{2.35}$$

$$2nD^{\text{MM}}(\Gamma^n \| \pi^0) \xrightarrow[n\to\infty]{d.} \chi^2_d$$

(ii) *When $\pi = \pi^1 \neq \pi^0$, we have with $\sigma_1^2 := \mathsf{Cov}_{\pi^1}(f_{r^*})$,*

$$\lim_{n\to\infty} \mathsf{E}[n(D^{\text{MM}}(\Gamma^n \| \pi^0) - D(\pi^1 \| \pi^0))] = \tfrac{1}{2}d \tag{2.36}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[n^{\frac{1}{2}} D^{\text{MM}}(\Gamma^n \| \pi^0)] = \sigma_1^2 \tag{2.37}$$

$$n^{\frac{1}{2}}(D^{\text{MM}}(\Gamma^n \| \pi^0) - D(\pi^1 \| \pi^0)) \xrightarrow[n\to\infty]{d.} \mathcal{N}(0, \sigma_1^2). \tag{2.38}$$

□

In part (ii) of Theorem 2.3.2, the assumption that r^* exists implies that π^1 and π^0 have equal supports. Furthermore, if Assumption (A3) holds in part (ii), then a sufficient condition for Assumption (A2) is that the function $V(r) := (-\pi^1(f_r) + \Lambda_{\pi^0}(f_r))$ be coercive in r. And, under (A3), the function

V is strictly convex and coercive in the following settings: (i) If the function class is linear, or (ii) the function class is log-linear, and the two distributions π^1 and π^0 have common support. We use this fact in Theorem 2.3.3 for the linear function class. The assumption of the existence of r^* satisfying $f_{r^*} = \log(\pi^1/\pi^0)$ in part (ii) of Theorem 2.3.2 can be relaxed. In the case of a linear function class we have the following extension of part (ii).

Theorem 2.3.3. *Suppose that the observation sequence $\boldsymbol{Z}$ is drawn i.i.d. with marginal π^1 satisfying $\pi^1 \prec \pi^0$. Let $\mathcal{F}$ be the linear function class defined in (2.26). Suppose the supremum in the definition of $D^{\text{MM}}(\pi^1\|\pi^0)$ is achieved at some $r^1 \in \mathbb{R}^d$. Further, suppose that the functions $\{\psi_i\}$ satisfy the linear independence condition of Assumption (A3) with $\mu^0 = \pi^1$. Then we have*

$$\begin{aligned}
\lim_{n\to\infty} \mathsf{E}[n(D^{\text{MM}}(\Gamma^n\|\pi^0) - D^{\text{MM}}(\pi^1\|\pi^0))] &= \tfrac{1}{2}\mathsf{trace}(\Sigma_{\pi^1}\Sigma_{\check{\pi}}^{-1}) \\
\lim_{n\to\infty} \mathsf{Var}\,[n^{\frac{1}{2}} D^{\text{MM}}(\Gamma^n\|\pi^0)] &= \sigma_1^2 \\
n^{\frac{1}{2}}(D^{\text{MM}}(\Gamma^n\|\pi^0) - D^{\text{MM}}(\pi^1\|\pi^0)) &\xrightarrow[n\to\infty]{d.} \mathcal{N}(0,\sigma_1^2)
\end{aligned}$$

where in the first limit $\check{\pi} = \pi^0 \exp(f_{r^1} - \Lambda_{\pi^0}(f_{r^1}))$, and Σ_{π^1} and $\Sigma_{\check{\pi}}$ are defined as in (2.32). In the second two limits $\sigma_1^2 = \mathsf{Cov}_{\pi^1}(f_{r^1})$. □

Although we have not explicitly imposed Assumption (A2) in Theorem 2.3.3, the argument we presented following Theorem 2.3.2 ensures that when $\pi^1 \prec \pi^0$, Assumption (A2) is satisfied whenever Assumption (A3) holds. Furthermore, it can be shown that the achievement of the supremum required in Theorem 2.3.3 is guaranteed if π^1 and π^0 have equal supports. We also note that the vector s appearing in Eq. (2.31) of Assumption (A3) is arbitrary when the parametrization of the function class is linear.

The weak convergence results in Theorem 2.3.2 (i) can be derived from Clarke and Barron [25, 32] (see also [20, Theorem 4.2]), following the maximum-likelihood estimation interpretation of the mismatched test obtained in Proposition 2.2.7. In the statistics literature, such results are called *Wilks phenomena* after the initial work by Wilks [26]. These results can be used to set thresholds for a target false alarm probability in the mismatched test, just like we did for the Hoeffding test in (2.15).

Implications for Hoeffding test The divergence can be interpreted as a special case of mismatched divergence defined with respect to a linear function class. Using this interpretation, the results of Theorem 2.3.2 can also be specialized to obtain results on the Hoeffding test statistic. To satisfy the uniqueness condition of Assumption (A2), we require that the function class should not contain any constant functions. Now suppose that the span of the linear function class $\mathcal{F}$ together with the constant function $f^0 \equiv 1$ spans the set of all functions on Z. This together with Assumption (A3) would imply that $d = N - 1$, where N is the size of the alphabet Z. It follows from Proposition 2.2.1 that for such a function class the mismatched divergence coincides with the divergence. Thus, an application of Theorem 2.3.2 (i) gives rise to the results stated in Theorem 2.2.2.

To prove Theorem 2.3.2 and Theorem 2.3.3 we need some lemmas, whose proofs are given in the Appendix.

The following lemma will be used to deduce part (ii) of Theorem 2.3.2 from part (i).

Lemma 2.3.4. *Let $D_{\mathcal{F}}^{\text{MM}}$ denote the mismatched divergence defined using function class $\mathcal{F}$. Suppose $\pi^1 \prec \pi^0$ and the supremum in the definition of $D_{\mathcal{F}}^{\text{MM}}(\pi^1 \| \pi^0)$ is achieved at some $f_{r^*} \in \mathcal{F}$. Let $\check{\pi} = \pi^0 \exp(f_{r^*} - \Lambda_{\pi^0}(f_{r^*}))$ and $\mathcal{G} = \mathcal{F} - f_{r^*} := \{f_r - f_{r^*} : r \in \mathbb{R}^d\}$. Then for any μ satisfying $\mu \prec \pi^0$, we have*

$$D_{\mathcal{F}}^{\text{MM}}(\mu \| \pi^0) = D_{\mathcal{F}}^{\text{MM}}(\pi^1 \| \pi^0) + D_{\mathcal{G}}^{\text{MM}}(\mu \| \check{\pi}) + \langle \mu - \pi^1, \log(\frac{\check{\pi}}{\pi^0}) \rangle. \tag{2.39}$$

□

Suppose we apply the decomposition result from Lemma 2.3.4 to the type of the observation sequence $\boldsymbol{Z}$, assumed to be drawn i.i.d. with marginal π^1. If there exists r^* satisfying $f_{r^*} = \log(\pi^1/\pi^0)$, then we have $\check{\pi} = \pi^1$. The decomposition becomes

$$D_{\mathcal{F}}^{\text{MM}}(\Gamma^n \| \pi^0) = D_{\mathcal{F}}^{\text{MM}}(\pi^1 \| \pi^0) + D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1) + \langle \Gamma^n - \pi^1, f_{r^*} \rangle. \tag{2.40}$$

For large n, the second term in the decomposition (2.40) has a mean of order n^{-1} and variance of order n^{-2}, as shown in part (i) of Theorem 2.3.2. The third term has zero mean and variance of order n^{-1}, since by the Central

Limit Theorem,

$$n^{\frac{1}{2}}\langle \Gamma^n - \pi^1, f_{r^*} \rangle \xrightarrow[n\to\infty]{d.} \mathcal{N}(0, \mathsf{Cov}_{\pi^1}(f_{r^*})). \tag{2.41}$$

Thus, the asymptotic variance of $D_{\mathcal{F}}^{\text{MM}}(\Gamma^n \| \pi^0)$ is dominated by that of the third term and the asymptotic bias is dominated by that of the second term. Thus we see that part (ii) of Theorem 2.3.2 can be deduced from part (i).

Lemma 2.3.5. *Let $\boldsymbol{X} = \{X^i : i = 1, 2, \ldots\}$ be an i.i.d. sequence with mean $\bar{x}$ taking values in a compact convex set $\mathsf{X} \subset \mathbb{R}^m$, containing $\bar{x}$ as a relative interior point. Define $S^n = \frac{1}{n}\sum_{i=1}^n X^i$. Suppose we are given a function $h : \mathbb{R}^m \mapsto \mathbb{R}$, that is continuous over X, and a compact set K containing $\bar{x}$ as a relative interior point such that*

1. *The gradient $\nabla h(x)$ and the Hessian $\nabla^2 h(x)$ are continuous over a neighborhood of K.*
2. $\lim_{n\to\infty} -\dfrac{1}{n}\log \mathsf{P}\{S^n \notin K\} > 0.$

Let $M = \nabla^2 h(\bar{x})$ and $\Xi = \mathsf{Cov}(X^1)$. Then,

(i) *The normalized asymptotic bias of $\{h(S^n) : n \geq 1\}$ is obtained via*

$$\lim_{n\to\infty} n\mathsf{E}[h(S^n) - h(\bar{x})] = \tfrac{1}{2}\mathsf{trace}(M\Xi).$$

(ii) *If in addition to the above conditions, the directional derivative satisfies $\nabla h(\bar{x})^{\intercal}(X^1 - \bar{x}) = 0$ almost surely, then the asymptotic variance decays as n^{-2}, with*

$$\lim_{n\to\infty} \mathsf{Var}\,[nh(S^n)] = \tfrac{1}{2}\mathsf{trace}(M\Xi M\Xi).$$

□

Lemma 2.3.6. *Suppose that the observation sequence $\boldsymbol{Z}$ is drawn i.i.d. with marginal $\mu \in \mathcal{P}(\mathsf{Z})$. Let $h : \mathcal{P}(\mathsf{Z}) \mapsto \mathbb{R}$ be a continuous real-valued function whose gradient and Hessian are continuous in a neighborhood of μ. If the directional derivative satisfies $\nabla h(\mu)^{\intercal}(\nu - \mu) \equiv 0$ for all $\nu \in \mathcal{P}(\mathsf{Z})$, then*

$$2n(h(\Gamma^n) - h(\mu)) \xrightarrow[n\to\infty]{d.} W^{\intercal} M W \tag{2.42}$$

where $M = \nabla^2 h(\mu)$ and $W \sim \mathcal{N}(0, \Sigma_W)$ with $\Sigma_W = diag(\mu) - \mu\mu^{\intercal}$. □

Lemma 2.3.7. *Suppose that V is an m-dimensional, $\mathcal{N}(0, I_m)$ random variable, and $D\colon \mathbb{R}^m \to \mathbb{R}^m$ is a projection matrix. Then $\xi := \|DV\|^2$ is a chi-squared random variable with K degrees of freedom, where K denotes the rank of D.* □

Before we proceed to the proofs of Theorem 2.3.2 and Theorem 2.3.3, we recall the optimization problem (2.25) defining the mismatched divergence:

$$D^{\text{MM}}(\mu\|\pi^0) = \sup_{r\in\mathbb{R}^d} \Big(\mu(f_r) - \Lambda_{\pi^0}(f_r)\Big). \tag{2.43}$$

The first order condition for optimality is given by

$$g(\mu, r) = 0 \tag{2.44}$$

where g is the vector valued function that defines the gradient of the objective function in (2.43):

$$\begin{aligned} g(\mu, r) &:= \nabla_r \Big(\mu(f_r) - \Lambda_{\pi^0}(f_r)\Big) \\ &= \mu(\nabla_r f_r) - \frac{\pi^0(e^{f_r}\nabla_r f_r)}{\pi^0(e^{f_r})}. \end{aligned} \tag{2.45}$$

On letting $\psi^r = \nabla_r f_r$ we obtain

$$g(\mu, r) = \mu(\psi^r) - \check{\pi}^r(\psi^r). \tag{2.46}$$

The gradient $\nabla_r g(\mu, r)$ of $g(\mu, r)$ with respect to r is given by

$$\nabla_r g(\mu, r) = \mu(\nabla_r^2 f_r) - \check{\pi}^r(\nabla_r^2 f_r) - [\check{\pi}^r(\psi^r \psi^{r\,T}) - \check{\pi}^r(\psi^r)\check{\pi}^r(\psi^{r\,T})] \tag{2.47}$$

where the definition of the twisted distribution is as given in (2.10):

$$\check{\pi}^r := \pi^0 \exp(f_r - \Lambda_{\pi^0}(f_r)).$$

In these formulae we have extended the definition of $\mu(M)$ for matrix-valued functions M on Z via $[\mu(M)]_{ij} := \mu(M_{ij}) = \sum_z M_{ij}(z)\mu(z)$.

Proof of Theorem 2.3.2. Without loss of generality, we assume that π^0 has full support over Z. Suppose that the observation sequence $\boldsymbol{Z}$ is drawn

i.i.d. with marginal distribution $\pi \in \mathcal{P}(\mathsf{Z})$. We have $D^{\text{MM}}(\Gamma^n \| \pi^0) \xrightarrow[n\to\infty]{a.s.} D^{\text{MM}}(\pi \| \pi^0)$ by the law of large numbers.

1) *Proof of part (i):* We first prove the results concerning the bias and variance of the mismatched test statistic. We apply Lemma 2.3.5 to the function $h(\mu) := D^{\text{MM}}(\mu \| \pi^0)$. The other terms appearing in the lemma are taken to be $X^i = (\mathbb{I}_{z_1}(Z_i), \mathbb{I}_{z_2}(Z_i), \ldots, \mathbb{I}_{z_N}(Z_i))^T$, $\mathsf{X} = \mathcal{P}(\mathsf{Z})$, $\bar{x} = \pi^0$, and $S^n = \Gamma^n$. Let $\Xi = \mathsf{Cov}(X^1)$. It is easy to see that $\Xi = \text{diag}(\pi^0) - \pi^0 \pi^{0T}$ and $\Sigma_{\pi^0} = \Psi \Xi \Psi^T$, where Σ_{π^0} is defined in (2.32), and Ψ is a $d \times N$ matrix defined by

$$\Psi(i,j) = \psi_i(z_j). \tag{2.48}$$

This can be expressed as the concatenation of column vectors via $\Psi = [\psi(z_1), \psi(z_2), \ldots, \psi(z_N)]$.

We first demonstrate that

$$M = \nabla^2 h(\pi_0) = \Psi^T (\Sigma_{\pi^0})^{-1} \Psi \tag{2.49}$$

and then check to make sure that the other requirements of Lemma 2.3.5 are satisfied. The first two conclusions of Theorem 2.3.2 (i) will then follow from Lemma 2.3.5, since

$$\mathsf{trace}(M\Xi) = \mathsf{trace}((\Sigma_{\pi^0})^{-1} \Psi \Xi \Psi^T) = \mathsf{trace}(I_d) = d$$

and similarly $\mathsf{trace}(M\Xi M\Xi) = \mathsf{trace}(I_d) = d$.

We first prove that under the assumptions of Theorem 2.3.2 (i), there is a function $r : \mathcal{P}(\mathsf{Z}) \mapsto \mathbb{R}$ that is C^1 in a neighborhood of π^0 such that $r(\mu)$ solves (2.43) for μ in this neighborhood. Under the uniqueness assumption (A2), the function $r(\mu)$ coincides with the function given in (A2).

By the assumptions, we know that when $\mu = \pi^0$, (2.44) is satisfied by r^* with $f_{r^*} \equiv 0$. It follows that $\pi^0 = \breve{\pi}^{r^*}$. Substituting this into (2.47), we obtain $\nabla_r g(\mu, r)\Big|_{\substack{\mu=\pi^0 \\ r=r^*}} = -\Sigma_{\pi^0}$, which is negative-definite by Assumption (A3) and Lemma 2.3.1. Therefore, by the Implicit Function Theorem, there is an open neighborhood U around $\mu = \pi^0$, an open neighborhood V of r^*, and a continuously differentiable function $r : U \to V$ that satisfies $g(\mu, r(\mu)) = 0$, for $\mu \in U$. This fact together with Assumptions (A2) and (A3) ensures that when $\mu \in U \cap B$, the vector $r(\mu)$ uniquely achieves the supremum in (2.43).

Taking the total derivative of (2.44) with respect to $\mu(z)$ we get

$$\frac{\partial r(\mu)}{\partial \mu(z)} = -\Big[\nabla_r g(\mu, r(\mu))\Big]^{-1} \frac{\partial g(\mu, r(\mu))}{\partial \mu(z)}. \tag{2.50}$$

Consequently, when $\mu = \pi^0$,

$$\left.\frac{\partial r(\mu)}{\partial \mu(z)}\right|_{\mu=\pi^0} = \Sigma_{\pi^0}^{-1}\psi(z). \tag{2.51}$$

These results enable us to identify the first and second order derivative of $h(\mu) = D^{\text{MM}}(\mu\|\pi^0)$. Applying $g(\mu, r(\mu)) = 0$, we obtain the derivatives of h as follows:

$$\frac{\partial}{\partial \mu(z)} h(\mu) = f_{r(\mu)}(z). \tag{2.52}$$

$$\frac{\partial^2}{\partial \mu(z)\partial \mu(\bar{z})} h(\mu) = (\nabla_r f_{r(\mu)}(z))^T \frac{\partial r(\mu)}{\partial \mu(\bar{z})}. \tag{2.53}$$

When $\mu = \pi^0$, substituting (2.51) in (2.53), we obtain (2.49).

We now verify the remaining conditions required for applying Lemma 2.3.5:

(a) It is straightforward to see that $h(\pi^0) = 0$.

(b) The function h is uniformly bounded since $h(\mu) = D^{\text{MM}}(\mu\|\pi^0) \leq D(\mu\|\pi^0) \leq \max_z \log(\frac{1}{\pi^0(z)})$ and π^0 has full support.

(c) Since $f_{r(\mu)} = 0$ when $\mu = \pi^0$, it follows by (2.52) that $\frac{\partial}{\partial \mu(z)} h(\mu)\Big|_{\mu=\pi^0} = 0$.

(d) Pick a compact $K \subset U \cap B$ so that K contains π^0 as a relative interior point, and $K \subset \{\mu \in \mathcal{P}(\mathsf{Z}) : \max_u |\mu(u) - \pi^0(u)| < \frac{1}{2}\min_u |\pi^0(u)|\}$. This choice of K ensures that $\lim_{n\to\infty} -\frac{1}{n}\log \mathsf{P}\{S^n \notin K\} > 0$. Note that since $r(\mu)$ is continuously differentiable on $U \cap B$, it follows by (2.52) and (2.53) that h is C^2 on K.

Thus the results on convergence of the bias and variance follow from Lemma 2.3.5.

The weak convergence result is proved using Lemma 2.3.6 and Lemma 2.3.7. We observe that the covariance matrix of the Gaussian vector W given in

Lemma 2.3.6 is $\Sigma_W = \Xi = \text{diag}(\pi^0) - \pi^0\pi^{0T}$. This does not have full rank since $\Xi\mathbf{1} = 0$, where $\mathbf{1}$ is the $N \times 1$ vector of ones. Hence we can write

$$\Xi = GG^T$$

where G is an $N \times k$ matrix for some $k < N$. In fact, since the support of π^0 is full, we have $k = N - 1$ (see Lemma 2.3.1). Based on this representation we can write $W = GV$, where $V \sim \mathcal{N}(0, I_k)$.

Now, by Lemma 2.3.6, the limiting random variable is given by $U := W^TMW = V^TG^TMGV$, where $M = \nabla^2_\mu D^{\text{MM}}(\mu\|\pi^0)\Big|_{\pi^0} = \Psi^T(\Psi\Xi\Psi^T)^{-1}\Psi$. We observe that the matrix $D = G^TMG$ satisfies $D^2 = D$. Moreover, since $\Psi\Xi\Psi^T$ has rank d under Assumption (A3), matrix D also has rank d. Applying Lemma 2.3.7 to matrix D, we conclude that $U \sim \chi^2_d$.

2) *Proof of part (ii):* The conclusion of part (ii) is derived using part (i) and the decomposition in (2.40). We will study the bias, variance, and limiting distribution of each term in the decomposition.

For the second term, note that the dimensionality of the function class $\mathcal{G}$ is also d. Applying part (i) of this theorem to $D^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)$, we conclude that its asymptotic bias and variance are given by

$$\lim_{n\to\infty} \mathsf{E}[nD^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)] = \tfrac{1}{2}d, \tag{2.54}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD^{\text{MM}}_{\mathcal{G}}(\Gamma^n\|\pi^1)] = \tfrac{1}{2}d. \tag{2.55}$$

For the third term, since $\boldsymbol{Z}$ is i.i.d. with marginal π^1, we have

$$\mathsf{E}[\langle\Gamma^n - \pi^1, f_{r^*}\rangle] = 0, \tag{2.56}$$

$$\mathsf{Var}\,[n^{\frac{1}{2}}\langle\Gamma^n - \pi^1, f_{r^*}\rangle] = \mathsf{Cov}_{\pi^1}(f_{r^*}). \tag{2.57}$$

The bias result (2.36) follows by combining (2.54), (2.56) and using the decomposition (2.40). To prove the variance result (2.37), we again apply the

decomposition (2.40) to obtain

$$\begin{aligned}
&\lim_{n\to\infty} \mathsf{Var}\left[n^{\frac{1}{2}} D_{\mathcal{F}}^{\text{MM}}(\Gamma^n \| \pi^0)\right] \\
&= \lim_{n\to\infty} \Big\{ \mathsf{Var}\left[n^{\frac{1}{2}} D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1)\right] + \mathsf{Var}\left[n^{\frac{1}{2}} \langle \Gamma^n - \pi^1, f_{r^*} \rangle\right] \\
&\quad + 2E\Big[n^{\frac{1}{2}} \Big(D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1) - \mathsf{E}[D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1)] \Big) n^{\frac{1}{2}} \langle \Gamma^n - \pi^1, f_{r^*} \rangle \Big] \Big\}.
\end{aligned} \tag{2.58}$$

From (2.55) it follows that the limiting value of the first term on the right hand side of (2.58) is 0. The limiting value of the third term is also 0 by the Cauchy-Bunyakovsky–Schwarz inequality. Thus, (2.58) together with (2.57) gives (2.37).

Finally, we prove the weak convergence result (2.38) by again applying the decomposition (2.40). By (2.54) and (2.55), we conclude that the second term $n^{\frac{1}{2}} D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \pi^1)$ converges in mean square to 0 as $n \to \infty$. The weak convergence of the third term is given in (2.41). Applying Slutsky's theorem, we obtain (2.38). □

Proof of Theorem 2.3.3. The proof of this result is very similar to that of Theorem 2.3.2 (ii) except that we use the decomposition in (2.39) with $\mu = \Gamma^n$. We first prove the following generalization of (2.54) and (2.55) that characterizes the asymptotic mean and variance of the second term in (2.39) with $\mu = \Gamma^n$:

$$\lim_{n\to\infty} \mathsf{E}[n D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \check{\pi})] = \tfrac{1}{2} \mathsf{trace}\Big(\Sigma_{\pi^1} (\Sigma_{\check{\pi}})^{-1} \Big) \tag{2.59}$$

$$\lim_{n\to\infty} \mathsf{Var}\left[n D_{\mathcal{G}}^{\text{MM}}(\Gamma^n \| \check{\pi})\right] = \tfrac{1}{2} \mathsf{trace}\Big(\Sigma_{\pi^1} (\Sigma_{\check{\pi}})^{-1} \Sigma_{\pi^1} (\Sigma_{\check{\pi}})^{-1} \Big) \tag{2.60}$$

where $\mathcal{G} = \mathcal{F} - f_{r^1}$, and $\check{\pi}$ is defined in the statement of the proposition. The argument is similar to that of Theorem 2.3.2 (i): We denote $\tilde{f}_r := f_r - f_{r^1}$, and define $h(\mu) := D_{\mathcal{G}}^{\text{MM}}(\mu \| \check{\pi}) = \sup_{r \in \mathbb{R}^d} \Big(\mu(\tilde{f}_r) - \Lambda_{\check{\pi}}(\tilde{f}_r) \Big)$. To apply Lemma 2.3.5, we prove the following:

$$h(\pi^1) = 0, \tag{2.61}$$

$$\nabla_\mu h(\pi^1) = 0, \tag{2.62}$$

$$\textit{and} \qquad M = \nabla_\mu^2 h(\pi^1) = \Psi^T (\Sigma_{\check{\pi}})^{-1} \Psi. \tag{2.63}$$

The last two inequalities (2.62) and (2.63) are analogous to (2.52) and (2.53).

We can also verify that the rest of the conditions of Lemma 2.3.5 hold. This establishes (2.59) and (2.60).

To prove (2.61), first note that the supremum in the optimization problem defining $D^{\text{MM}}(\pi^1\|\check{\pi})$ is achieved by $\tilde{f}_{r^1}$, and we know by definition that $\tilde{f}_{r^1} = 0$. Together with the definition $D^{\text{MM}}(\pi^1\|\check{\pi}) = \pi^1(\tilde{f}_{r^1}) - \Lambda_{\check{\pi}}(\tilde{f}_r)$, we obtain (2.61).

Redefine $g(\mu, r) := \nabla_r\Big(\mu(\tilde{f}_r) - \Lambda_{\check{\pi}}(\tilde{f}_r)\Big)$. The first order optimality condition of the optimization problem defining $D^{\text{MM}}(\mu\|\check{\pi})$ gives $g(\mu, r) = 0$. The assumption that $\mathcal{F}$ is a linear function class implies that $\tilde{f}_r$ is linear in r. Consequently $\nabla_r^2\tilde{f}_r = 0$. By the same argument that leads to (2.47), we can show that

$$\nabla_r g(\mu, r) = -\left[\frac{\check{\pi}\Big(e^{\tilde{f}_r}\nabla_r\tilde{f}_r\nabla_r\tilde{f}_r^T\Big)}{\check{\pi}\Big(e^{\tilde{f}_r}\Big)} - \frac{\check{\pi}\Big(e^{\tilde{f}_r}\nabla_r\tilde{f}_r\Big)\check{\pi}\Big(e^{\tilde{f}_r}\nabla_r\tilde{f}_r^T\Big)}{(\check{\pi}\Big(e^{\tilde{f}_r}\Big))^2}\right]. \tag{2.64}$$

Together with the fact that $\tilde{f}_{r^1} = 0$ and $\nabla_r\tilde{f}_r = \nabla_r f_r$, we obtain

$$\nabla_r g(\mu, r)\Big|_{\substack{\mu=\pi^1\\ r=r^1}} = -\Sigma_{\check{\pi}}. \tag{2.65}$$

Proceeding as in the proof of Theorem 2.3.2 (i), we obtain (2.62) and (2.63).

Now using similar steps as in the proof of Theorem 2.3.2 (ii), and noticing that $\log(\frac{\check{\pi}}{\pi^0}) = f_{r^1}$, we can establish the following results on the third term of (2.39):

$$\begin{aligned}
\mathsf{E}[\langle\Gamma^n - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle] &= 0\\
\mathsf{Var}\,[n^{\frac{1}{2}}\langle\Gamma^n - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle] &= \mathsf{Cov}_{\pi^1}(f_{r^1})\\
n^{\frac{1}{2}}\langle\Gamma^n - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle &\xrightarrow[n\to\infty]{d.} \mathcal{N}(0, \mathsf{Cov}_{\pi^1}(f_{r^1})).
\end{aligned}$$

Continuing the same arguments as in Theorem 2.3.2 (i), we obtain the result of Theorem 2.3.3. □

2.3.1 Interpretation of the asymptotic results and performance comparison

The asymptotic results established above can be used to study the finite sample performance of the mismatched test and Hoeffding test. Recall that in the discussion surrounding Figure 2.1 we concluded that the approximation obtained from a Central Limit Theorem gives much better estimates of error probabilities as compared to those suggested by Sanov's theorem.

Suppose the log-likelihood ratio function $\log(\pi^1/\pi^0)$ lies in the function class $\mathcal{F}$. In this case, the results of Theorem 2.3.2 and Lemma 2.3.4 are informally summarized in the following approximations: With Γ^n denoting the empirical distributions of the i.i.d. process $\boldsymbol{Z}$,

$$D^{\text{MM}}(\Gamma^n \| \pi^0) \approx \begin{cases} D(\pi^0 \| \pi^0) + \frac{1}{2}\frac{1}{n}\sum_{k=1}^d W_k^2\,, & Z_i \sim \pi^0 \\ D(\pi^1 \| \pi^0) + \frac{1}{2}\frac{1}{n}\sum_{k=1}^d W_k^2 + \frac{1}{\sqrt{n}}\sigma_1 U\,, & Z_i \sim \pi^1 \end{cases} \tag{2.66}$$

where $\{W_k\}$ is i.i.d., $N(0,1)$, and U is also $N(0,1)$ but not independent of the W_k's. The standard deviation σ_1 is given in Theorem 2.3.2. These distributional approximations are valid for large n, and are subject to assumptions on the function class used in the theorem.

We observe from (2.66) that, for large enough n, when the observations are drawn under π^0, the mismatched divergence is well approximated by $\frac{1}{2n}$ times a chi-squared random variable with d degrees of freedom. We also observe that when the observations are drawn under π^1, the mismatched divergence is well approximated by a Gaussian random variable with mean $D(\pi^1\|\pi^0)$ and with a variance proportional to $\frac{1}{n}$ and independent of d. Thus we expect to see a better receiver operating characteristic (ROC) for a lower value of d provided the log-likelihood ratio function $\log(\pi^1/\pi^0)$ lies in the function class $\mathcal{F}$. Since the mismatched test can be interpreted as a GLRT, these results capture the rate of degradation of the finite sample performance of a GLRT as the dimensionality of the parameterized family of alternate hypotheses increases. We corroborate this intuitive reasoning through Monte Carlo simulation experiments.

We estimated via simulation the performances of the Hoeffding test and mismatched tests designed using a linear function class. We compared the error probabilities of these tests for an alphabet size of $N = 19$ and sequence

length of $n = 40$. We chose π^0 to be the uniform distribution, and π^1 to be the distribution obtained by convolving two uniform distributions on sets of size $(N+1)/2$. We chose the basis function ψ_1 appearing in (2.26) to be the log-likelihood ratio between π^1 and π^0, viz.,

$$\psi_1(z_i) = \log \frac{\pi^1(z_i)}{\pi^0(z_i)}, \quad 1 \le i \le N$$

and the other basis functions $\psi_2, \psi_3, \ldots, \psi_d$ were chosen uniformly at random. Figure 2.4 shows a comparison of the ROCs of the Hoeffding test and mismatched tests for different values of dimension d. Plotted on the x-axis is the probability of false alarm, i.e., the probability of misclassification under π^0; shown on the y-axis is the probability of detection, i.e., the probability of correct classification under π^1. The various points on each ROC curve are obtained by varying the threshold η used in the Hoeffding test of (2.4) and mismatched test of (2.19).

From Figure 2.4 we see that as d increases the performance of the mismatched tests degrades. This is consistent with the approximation (2.66) which suggests that the variance of the mismatched divergence increases with d. Furthermore, as we saw earlier, the Hoeffding test can be interpreted as a special case of the mismatched test for a specific choice of the function class with $d = N - 1$ and hence the performance of the mismatched test matches the performance of the Hoeffding test when $d = N - 1$.

To summarize, the above results suggest that although the Hoeffding test is optimal in an error-exponent sense, it is disadvantageous in terms of finite sample error probabilities to blindly use the Hoeffding test if it is known a priori that the alternate distribution belongs to some parameterized family of distributions.

2.4 Approximate Implementation

Although the mismatched test has performance advantages over the Hoeffding test, it can be computationally complex to implement. The optimization problem in (2.8) that needs to be solved to evaluate the mismatched divergence $D^{\text{MM}}(\Gamma^n \| \pi^0)$ can be complex, especially for non-linear function classes. In this section, we propose an approximation to the mismatched divergence

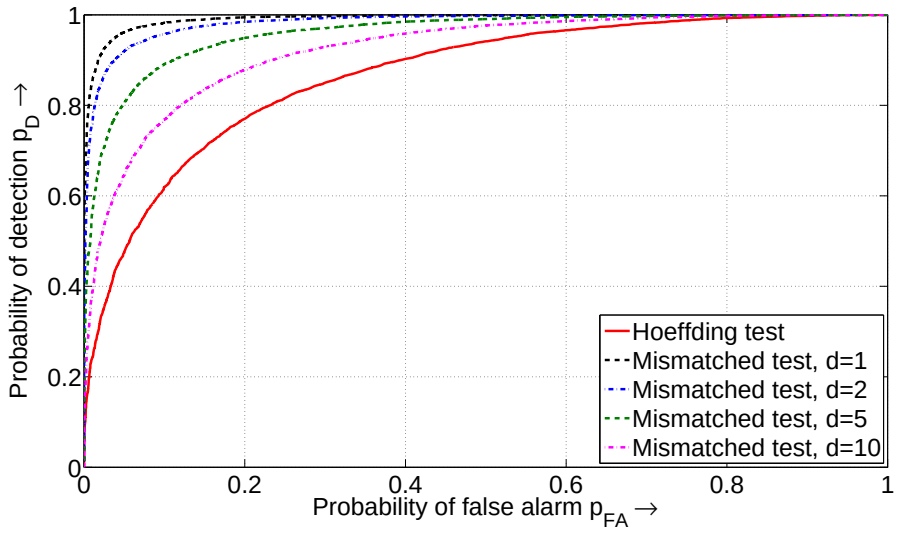

Figure 2.4: *Comparisons of ROCs of Hoeffding and mismatched tests.*

that can be computed easily and also gives good performance in terms of error probabilities of the mismatched test that uses the approximation.

The approximation which we derive in Appendix A is based on a Taylor's approximation of the objective function in (2.8). Assuming that Assumption (A1) holds, we let Φ denote the Hessian $\nabla^2 f_r$ evaluated at $r = r^0$ where r^0 satisfies $f_{r^0} = 0$. The approximate mismatched divergence between any distribution μ and π^0 is then given by

$$\hat{D}^{\text{MM}}(\mu \| \pi^0) = \tfrac{1}{2}(\mu - \pi^0)^T \Psi^T M_\mu^{-1} \Psi (\mu - \pi^0) \tag{2.67}$$

where Ψ is defined in (2.48) and $M_\mu = \Sigma_{\pi^0} + \pi(\Phi) - \mu(\Phi)$ is assumed to be invertible.

The approximate mismatched test uses the following test statistic:

$$\hat{D}^{\text{MM}}(\Gamma^n \| \pi^0) = \tfrac{1}{2}(\Gamma^n - \pi^0)^T \Psi^T M_{\Gamma^n}^{-1} \Psi (\Gamma^n - \pi^0).$$

For the linear function class of (2.26), Φ is a null matrix and hence M_{Γ^n} is independent of Γ^n and always invertible provided Assumption (A3) holds. The test statistic in this case is a quadratic function,

$$\hat{D}^{\text{MM}}(\Gamma^n \| \pi^0) = \tfrac{1}{2}(\Gamma^n - \pi^0)^T \Psi^T \Sigma_{\pi^0}^{-1} \Psi (\Gamma^n - \pi^0).$$

Similarly, for the log-linear function class of (2.30), M_{Γ^n} is given by

$$M_{\Gamma^n} = \Sigma_{\pi^0} + \pi(\psi\psi^T) - \Gamma^n(\psi\psi^T).$$

From the expression for the approximate mismatched divergence, it follows that

$$\begin{aligned}
\hat{D}^{\text{MM}}(\pi^0\|\pi^0) &= 0 \\
\nabla_\mu \hat{D}^{\text{MM}}(\mu\|\pi^0)\Big|_{\mu=\pi^0} &= 0 \\
\nabla^2_\mu \hat{D}^{\text{MM}}(\mu\|\pi^0)\Big|_{\mu=\pi^0} &= \Psi^T M^{-1}_{\pi^0}\Psi = \Psi^T\Sigma^{-1}_{\pi^0}\Psi.
\end{aligned}$$

We see that $\hat{D}^{\text{MM}}(\mu\|\pi^0)$ and $D^{\text{MM}}(\mu\|\pi^0)$ have the same gradient and Hessian at $\mu = \pi^0$. Therefore, by following the same steps used in proving the first part of Theorem 2.3.2 we conclude that the asymptotic behavior of $\hat{D}^{\text{MM}}(\Gamma^n\|\pi^0)$ and $D^{\text{MM}}(\Gamma^n\|\pi^0)$ are identical when the observation sequence $\boldsymbol{Z}$ is drawn i.i.d. with marginal π^0. We have

$$\lim_{n\to\infty} \mathsf{E}[n\hat{D}^{\text{MM}}(\Gamma^n\|\pi)] = \tfrac{1}{2}d \tag{2.68}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[n\hat{D}^{\text{MM}}(\Gamma^n\|\pi)] = \tfrac{1}{2}d \tag{2.69}$$

$$2n\hat{D}^{\text{MM}}(\Gamma^n\|\pi) \xrightarrow[n\to\infty]{d.} \chi^2_d. \tag{2.70}$$

The weak convergence result above can be used to set thresholds for achieving a target false alarm probability in the approximate mismatched test, just like was done for the Hoeffding test in (2.15).

In order to justify the use of the approximate mismatched test, we simulated its performance for the same parameters considered in the simulations of Section 2.3.1. Figure 2.5 shows a comparison of the ROCs of the mismatched tests and the approximate mismatched tests. We see that the performance of the approximate test is quite close to that of the mismatched test and is significantly better than the performance of the Hoeffding test, thus justifying the practical usefulness of the approximation.

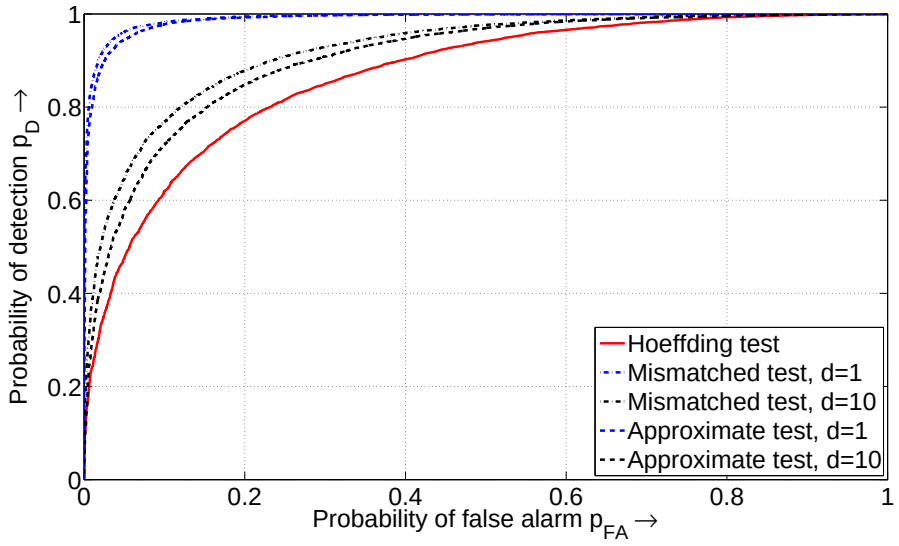

Figure 2.5: *Comparisons of ROCs of approximate mismatched tests and mismatched tests.*

2.5 Summary

The mismatched test studied in this chapter provides a solution to the universal hypothesis testing problem that can incorporate prior knowledge in order to reduce variance. The main results of Section 2.3 show that the variance reduction over Hoeffding's optimal test is substantial when the state space is large.

The dimensionality of the function class can be chosen by the designer to ensure that the the bias and variance are within tolerable limits. It is in this phase of design that prior knowledge is required to ensure that the error-exponent remains sufficiently large under the alternate hypothesis (see e.g. Corollary 2.2.5). In this way the designer can make effective tradeoffs between the power of the test and the variance of the test statistic.

The mismatched divergence provides a unification of several approaches to robust and universal hypothesis testing. Although constructed in an i.i.d. setting, the mismatched tests are applicable in very general settings, and the performance analysis presented here can be easily generalized to any stationary processes satisfying the Central Limit Theorem.

CHAPTER 3

UNIVERSAL HYPOTHESIS TESTING UNDER MODEL UNCERTAINTY

In the universal hypothesis testing problem studied in Chapter 2, we assumed that the statistics of the observations under the null hypothesis are known exactly. In this chapter we study the universal hypothesis testing problem where there is uncertainty about the distribution under the null hypothesis. We study adaptations to the popular Hoeffding test and Kolmogorov-Smirnov (KS) test that ensure robustness to uncertainties in the model. The KS test is a standard solution to the universal hypothesis testing problem for infinite alphabets.

We first consider a robust version of the Hoeffding test that was studied in [22]. Following our interpretation of the robust test statistic as a mismatched divergence in Section 2.2.6 of Chapter 2, we show that our results on weak convergence of the mismatched divergence can be used to set thresholds for the robust test. Later in the chapter we propose a robust version of the KS test and obtain new weak convergence results on the robust test statistic that are also useful for setting thresholds in robust universal hypothesis testing involving continuous distributions.

3.1 Robust Hoeffding Test

In the general universal hypothesis testing problem, we know the null distribution π^0 exactly but we do not have any prior information about the alternate distribution π^1. In Chapter 2 we studied the Hoeffding test (2.4) which optimizes the error exponents in the sense of (2.24). We now study a robust version of the Hoeffding test that can be used when the null distribution π^0 is not known exactly. We continue to use the notation we introduced in Chapter 2. As before, we let $\boldsymbol{Z} = \{Z_1, Z_2, \ldots\}$ denote the sequence of observations based on which we have to make the decision about the hypoth-

esis.

We saw in (2.14) that the test-statistic satisfies the following weak convergence result under π^0:

$$2nD(\Gamma^n\|\pi^0) \xrightarrow[n\to\infty]{d.} \chi^2_{N-1} \tag{3.1}$$

where Γ^n is the empirical distribution of the first n observations, N is the size of the support of distribution π^0, and χ^2_γ denotes a chi-square random variable with γ degrees of freedom. This result enables us to set approximate thresholds for large n using tables of the chi-square distribution.

In this section we study a robust version of the Hoeffding test that can be used when the distribution of the observations under the null hypothesis is not known exactly but is known to belong to an uncertainty class of distributions $\mathbb{P}$. The robust Hoeffding test proposed in [22] is based on a robust version of the divergence defined as

$$D^{\text{ROB}}(\mu\|\mathbb{P}) := \inf_{\pi\in\mathbb{P}} D(\mu\|\pi).$$

The robust test statistic is given by $D^{\text{ROB}}(\Gamma^n\|\mathbb{P})$. Thus the proposed test is represented by the binary decision,

$$\phi^{\text{ROB}}_{\tau,n} = \mathbb{I}\{D^{\text{ROB}}(\Gamma^n\|\mathbb{P}) > \tau\} \tag{3.2}$$

where τ is a threshold that must be chosen to meet the constraint on the false alarm probability.

In the rest of this section we address the following question: Can we obtain any convergence results for the worst case probability of error of the robust Hoeffding test like those implied by the weak convergence result of (3.1)? For the uncertainty class $\mathbb{P}$ considered in [22], and for any $\pi \in \mathbb{P}$, we evaluate the limiting value of

$$\lim_{n\to\infty} \mathsf{P}_\pi\{2nD^{\text{ROB}}(\Gamma^n\|\mathbb{P}) > \delta\} \tag{3.3}$$

where the subscript of π in P_π indicates that the observations are drawn i.i.d. according to law π. Using such results we provide guidelines for setting thresholds for the robust tests that guarantee a uniform false constraint over all $\pi \in \mathbb{P}$.

As we saw in Section 2.2.6, let $\mathbb{P}$ be an uncertainty class defined through

the following moment constraints:

$$\mathbb{P} := \{\pi \in \mathcal{P}(\mathsf{Z}) : \pi(\psi_i) = 0, \quad 1 \le i \le d\}$$

where $\pi(\psi_i)$ denotes the expected value $\sum_{z\in\mathsf{Z}} \pi(z)\psi_i(z)$ as in Chapter 2. Let ψ denote the vector of functions $(\psi_1, \psi_2, \ldots, \psi_d)^T$ and Z_π denote the support of distribution π. We make the following assumptions about the functions $\{\psi_i : 1 \le i \le d\}$:

Assumptions

(B1) There is some distribution $\pi^0 \in \mathbb{P}$ such that the functions $\{\psi_i : 0 \le i \le d\}$ are linearly independent over Z_{π^0}, where $\psi_0 \equiv 1$.

(B2) The origin $0 \in \mathbb{R}^d$ is an interior point of the set of feasible moment vectors, defined as

$$\Delta := \{x \in \mathbb{R}^d : x_i = \nu(\psi_i), i = 1, \ldots, d, \text{ for some } \nu \in \mathcal{P}(\mathsf{Z})\}.$$

The linear independence assumption of (B1) is identical to Assumption (A3) introduced in Chapter 2.

We know from the results of [22] that the robust divergence with respect to $\mathbb{P}$ can be expressed as the solution to the following optimization problem:

$$D^{\text{ROB}}(\mu\|\mathbb{P}) = \sup_{r\in\mathcal{R}} \mu(\log(1 + r^T\psi)) \tag{3.4}$$

where the supremum is taken over

$$\mathcal{R} := \{r \in \mathbb{R}^d : 1 + r^T\psi(z) \ge 0 \text{ for all } z \in \mathsf{Z}\}.$$

We also saw in Section 2.2.6 that the robust divergence can be expressed as a mismatched divergence defined with respect to a log-linear function class. In the rest of this section, we argue that the results on weak convergence of the mismatched divergence statistic from Theorem 2.3.2 can be used to set thresholds for the robust hypothesis testing procedure.

The main result of this section is the following theorem. For any $\pi \in \mathbb{P}$, let d_π denote the number that is one lower than the maximal number of functions in $\{\psi_i : 0 \le i \le d\}$ that are linearly independent over Z_π. In other

words, $d_\pi + 1$ is the dimension of the span of the functions $\{\psi_i : 0 \leq i \leq d\}$ when restricted to Z_π.

Theorem 3.1.1. *Suppose assumptions (B1) and (B2) hold. Then, the following weak convergence result holds under π:*

$$2nD^{\mathrm{ROB}}(\Gamma^n \| \mathbb{P}) \xrightarrow[n\to\infty]{d.} \chi^2_{d_\pi}. \tag{3.5}$$

Hence we have

$$\sup_{\pi \in \mathbb{P}} \lim_{n\to\infty} \mathsf{P}_\pi\{2nD^{\mathrm{ROB}}(\Gamma^n \| \mathbb{P}) > \delta\} = 1 - F(d, \delta) \tag{3.6}$$

where $F(d, \delta)$ is the cumulative distribution function of a chi-square distribution with d degrees of freedom evaluated at δ. □

The second conclusion in the theorem above follows directly from the first, using the fact that chi-square distributions are stochastically ordered according to the number of degrees of freedom. Assumption (B1) ensures that the supremum is achieved by π^0 with $d_{\pi^0} = d$. The proof of the main result is included in the Appendix B.

The result of (3.6) can be used to set thresholds for the robust Hoeffding test. For a given n, suppose we choose τ in (3.2) such that $\tau = \frac{\delta}{2n}$ where δ satisfies

$$1 - F(d, \delta) = \alpha.$$

By (3.6) this choice of threshold ensures that for large enough n, and for all $\pi \in \mathbb{P}$,

$$\mathsf{P}_\pi\{\phi^{\mathrm{ROB}}_{\tau,n}(\boldsymbol{Z}) = 1\} \leq \alpha \tag{3.7}$$

thus guaranteeing a uniform bound on the false alarm probability over all distributions from $\mathbb{P}$.

3.2 Robust Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is a universal hypothesis test for testing the null hypothesis that a sequence of observations are drawn according to a probability law with distribution function F^0. The observations are assumed

to be drawn from a closed interval $\mathsf{Z} \subset \mathbb{R}$. The distribution function F^0 takes the role of π^0 in Chapter 2. The test statistic used in the KS test is

$$D_n = \sup_{x \in \mathsf{Z}} |F_n(x) - F^0(x)| \tag{3.8}$$

where F_n represents the empirical distribution function of the observations defined by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Z_i \leq x\}$$

where $\mathbb{I}$ is the indicator function. In the KS test based on $\boldsymbol{Z}$, the test statistic D_n is compared to a threshold τ chosen so that the probability of error under hypothesis $\mathcal{H}_0$ is less than some desired level. It is generally difficult to compute the exact distribution of the statistic D_n for large sequence lengths n. The typical practice is to use the following result by Kolmogorov for approximating the distribution of D_n for large n. When the observations $\boldsymbol{Z}$ are drawn i.i.d. from F^0, the following weak convergence result holds:

$$\mathsf{P}_{F^0}\{\sqrt{n} D_n > \delta\} \xrightarrow[n \to \infty]{d.} \mathsf{P}\{\sup_{t \in [0,1]} |K(t)| > \delta\} \tag{3.9}$$

where K is the Brownian bridge [33, p. 335]. The subscript of F^0 in (3.9) indicates that the observations were drawn from distribution F^0. Following this result thresholds are usually set using look-up tables containing values of the distribution function of $\sup_{t \in [0,1]} K(t)$, just as we did in (2.15).

While using the KS test in practice, one often faces the problem of overfitting. If the true underlying distribution is not exactly F^0 but some other distribution *close* to F^0, then the KS-test will eventually reject hypothesis $\mathcal{H}_0$ for n large enough.

In this work we propose a new robust version of the KS test that could potentially address this issue. The idea is to enlarge the set of null hypotheses beyond the singleton $\{F^0\}$. We define a new uncertainty class of distribution functions as follows:

$$\mathcal{F} = \{G \in \mathcal{P}(\mathsf{Z}) : F^-(x) \leq G(x) \leq F^+(x), \quad \forall x \in \mathsf{Z}\},$$

where $\mathcal{P}(\mathsf{Z})$ is the space of all probability distributions on Z and F^- and F^+ are continuous probability distribution functions such that the nominal

distribution $F^0 \in \mathcal{F}$. The advantage of using such an uncertainty class, as will become clear later, is that the resulting robust test is a simple modification of the standard KS test, and also admits a straightforward confidence approximation in the asymptotic setting. The distributions F^+ and F^- can be chosen to control the size of the class $\mathcal{F}$ that are acceptable as being close to F^0.

The proposed robust test uses the following robust test statistic:

$$E_n = \min_{F \in \mathcal{F}} \sup_{x \in \mathsf{Z}} |F_n(x) - F(x)| \tag{3.10}$$

where F_n is the empirical distribution function. The binary decision is given by

$$\phi^{\text{ROB}}_{\tau,n}(\boldsymbol{Z}) = \mathbb{I}\{E_n > \tau\}. \tag{3.11}$$

Let F^* denote the minimizer in (3.10) and let x^* denote the point at which the supremum in (3.10) is achieved for $F = F^*$. It is clear that if $E_n > 0$, then F^* is either F^+ or F^-. By definition it is obvious that for any $\delta > 0$, the event $\{\sqrt{n}E_n > \delta\}$ can be written as the union of two events as follows:

$$\begin{aligned} \{\sqrt{n}E_n > \delta\} \;\; = \;\; & \{\sup_{x \in \mathsf{Z}} \sqrt{n}(F_n(x) - F^+(x)) > \delta\} \cup \\ & \{\sup_{x \in \mathsf{Z}} \sqrt{n}[-(F_n(x) - F^-(x))] > \delta\}. \end{aligned} \tag{3.12}$$

The threshold used in (3.11) can be set using the following theorem that studies the asymptotic behavior of

$$p_n(\delta) := \max_{G \in \mathcal{F}} \mathsf{P}_G\{\sqrt{n}E_n > \delta\}.$$

Theorem 3.2.1. *For n large enough, $p_n(\delta)$ satisfies the following inequalities:*

$$p(\delta) \leq p_n(\delta) \leq 2p(\delta)$$

with $p(\delta)$ defined by $p(\delta) := \mathsf{P}\{\sup_{t \in [0,1]} K(t) > \delta\}$ where K is the Brownian bridge. □

We prove the theorem in Appendix B.

Theorem 3.2.1 suggests that if δ is chosen such that $2p(\delta) = \alpha$, and the threshold τ in the robust test (3.11) is chosen such that $\tau = \frac{\delta}{\sqrt{n}}$, then for

a large enough n, the test satisfies a uniform bound on the probability of probability of false alarm over all distributions in $\mathcal{F}$; i.e., for n large enough,

$$\max_{G\in\mathcal{F}} \mathsf{P}_G\{\phi^{\text{ROB}}_{\tau,n}(\boldsymbol{Z}) = 1\} \leq \alpha. \tag{3.13}$$

Furthermore, for n large enough we are also guaranteed that

$$\max_{G\in\mathcal{F}} \mathsf{P}_G\{\phi^{\text{ROB}}_{\tau,n}(\boldsymbol{Z}) = 1\} \geq \tfrac{1}{2}\alpha$$

which means that the maximum false alarm probability is within a factor of 0.5 from the desired level.

3.3 Summary

In this chapter we studied the problem of universal hypothesis testing when there is uncertainty about the observation statistics under the null hypothesis. We studied the robust Hoeffding test for finite alphabets that was proposed in [22] and proposed a robust version of the Kolmogorov-Smirnov test for infinite alphabets. We obtained weak-convergence results that enable us to set thresholds for these robust universal hypothesis tests that are similar to the approaches for setting thresholds in ordinary universal hypothesis tests.

CHAPTER 4

MINIMAX ROBUST QUICKEST CHANGE DETECTION

4.1 Introduction

The problem of detecting an abrupt change in a system based on observations is a dynamic hypothesis testing problem with a rich set of applications. Such problems of change detection were first studied by Page over fifty years ago in the context of quality control [34]. In its standard formulation there is a sequence of observations whose distribution changes at some unknown point in time, referred to as the 'change-point'. The goal is to detect this change as soon as possible, subject to a false alarm constraint. Some applications of change detection are intrusion detection in computer networks and security systems, detecting faults in infrastructure of various kinds, and spectrum monitoring for opportunistic access to wireless networks.

Most of the past work in the area of change detection has been restricted to the setting where the distributions of the observations prior to the change and after the change are known exactly (see, e.g., [35], [36], [37], [38]; for an overview of the work in this area, see [39], [40] and [41]). The three most popular criteria for optimizing the tradeoff between detection delay and false alarm rate are the Lorden criterion [36] and the Pollak criterion, in which the change-point is a deterministic quantity, and Shiryaev's Bayesian formulation [42], in which the change-point is modeled as a random variable with a known prior distribution. In this chapter we study all these three versions of change detection, under the setting where the pre-change and post-change distributions are not known exactly but belong to known uncertainty classes. We pose a minimax robust version of the standard quickest change detection problem wherein the objective is to identify the change detection rule that minimizes the maximum delay over all possible distributions. This minimization should be performed while meeting the false alarm constraint for

all possible values of the unknown distributions. We obtain a solution to this problem when the uncertainty classes satisfy some specific conditions. Under these conditions we can identify least favorable distributions (LFDs) from the uncertainty classes, and the optimal robust change detection rule is then the optimal (non-robust) change detection rule for the LFDs. These conditions are similar to those given by Huber [43] for robust hypothesis testing problems. We also discuss related results on robust sequential detection [43] [44] later in the chapter. The results of this chapter were also published in [45] and [46].

Although there has been some prior work on robust change detection, these approaches are distinctly different from ours. The maximin approach of [47] is similar in that they also identify LFDs for the robust problem. However, their result is restricted to asymptotic optimality (as the false alarm constraint goes to zero) under the Lorden criterion. A similar formulation is also discussed in [48, Sec.7.3.1]. Some other approaches to this problem (e.g. [49], [50]) are aimed at developing algorithms for quickest change detection with unknown distributions. These works study the asymptotic performance of the proposed tests under different distributions but do not seek to guarantee minimax robustness over a given class of distributions.

A closely related problem is the composite quickest change detection problem. In general, these problems also address the setting where the pre-change and post-change distributions are unknown. However, unlike the robust problem, in composite problems one seeks to identify a change detection procedure that is simultaneously optimal under all possible values of the unknown distributions. Exact solutions to these problems are often intractable and hence most results are restricted to asymptotic optimality. One such solution to a composite change detection problem is discussed in [36] when only the post-change distribution is unknown. In [36] a test is given that is asymptotically optimal under the Lorden criterion for all possible values of the unknown post-change distribution in a one-dimensional exponential family of distributions. This test is also referred to as the Generalized Likelihood Ratio Test (GLR Test), and was also studied in [51] and [52]. An alternate asymptotically optimal solution for the setting in which both pre-change and post-change distributions are unknown was studied in [53].

We provide a performance comparison of our proposed robust test with the GLR test. Although the GLR test asymptotically performs as well as

the optimal test with known distributions, we show via simulations that our robust test can give improved performance over the GLR test for moderate values of the false alarm constraint. The GLR test is also often prohibitively complex to implement in practice, while the proposed robust CUSUM test admits a simple recursive implementation.

For the asymptotic version of the problem, we also provide an analytical upper bound on the delay incurred by our robust test and use it to provide an upper bound on the drop in performance of our test relative to the optimal non-robust test.

The rest of the chapter is organized as follows. We first state the problem that we are studying in Section 4.2. In Section 4.3 we describe the robust solution and present some analysis. We discuss some examples in Section 4.4 and summarize the results of this chapter in Section 4.5.

4.2 Problem Statement

In the online quickest change detection problem we are given observations from a sequence $\{X_n : i = 1, 2, \ldots\}$ taking values in a set $\mathcal{X}$. There are two known distributions $\nu_0, \nu_1 \in \mathcal{P}(\mathcal{X})$ where $\mathcal{P}(\mathcal{X})$ is the set of probability distributions on $\mathcal{X}$. Initially, the observations are drawn i.i.d. under distribution ν_0. Their distribution switches abruptly to ν_1 at some unknown time λ so that $X_n \sim \nu_0$ for $n \leq \lambda - 1$ and $X_n \sim \nu_1$ for $n \geq \lambda$. This is illustrated in Figure 4.1. The observations are stochastically independent conditioned on the change-point. The objective is to identify the occurrence of change with minimum delay subject to false alarm constraints. We use E_m^ν to denote the expectation operator and P_m^ν to denote the probability law when the change happens at m and the pre-change and post-change distributions are ν_0 and ν_1 respectively. The symbols are replaced with E_∞^ν and P_∞^ν when the change does not happen. Similarly, if the pre-change and post-change distributions are some μ and γ, respectively, and the change happens at time m, we use $\mathsf{E}_m^{\mu,\gamma}$ to denote the expectation operator and $\mathsf{P}_m^{\mu,\gamma}$ the probability law. We further use $\mathcal{F}_m$ to denote the sigma algebra generated by $(X_1, X_2, \ldots, X_m)$.

A sequential change detection procedure is characterized by a stopping time τ with respect to the observation sequence. The design of the quickest change detection procedure involves optimizing the tradeoff between two per-

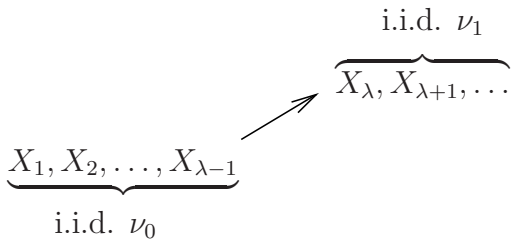

Figure 4.1: *Illustration of the change-point problem.* Initial observations X_1 through $X_{\lambda-1}$ have distribution ν_0. Later observations have distribution ν_1.

formance measures: detection delay and frequency of false alarms. There are various standard mathematical formulations for the optimal tradeoff. In the minimax formulation of [36] the change-point is assumed to be an unknown deterministic quantity. The worst-case detection delay is defined as

$$\mathrm{WDD}(\tau) = \sup_{\lambda \geq 1} \ \text{ess sup} \ \mathsf{E}_\lambda^\nu[(\tau - \lambda + 1)^+ | \mathcal{F}_{\lambda-1}]$$

where $x^+ = \max(x, 0)$. This quantity captures the worst-case value of the expected detection delay over all possible locations of the change-point and all possible realizations of the pre-change observations. The false alarm rate is defined as

$$\mathrm{FAR}(\tau) = \frac{1}{\mathsf{E}_\infty^\nu[\tau]}.$$

Here $\mathsf{E}_\infty^\nu[\tau]$ can be interpreted as the mean time to false alarm. Under the Lorden criterion, the objective is to find the stopping rule that minimizes the worst-case delay subject to an upper bound on the false alarm rate:

$$\text{Minimize } \mathrm{WDD}(\tau) \text{ subject to } \mathrm{FAR}(\tau) \leq \alpha \tag{4.1}$$

It was shown by Moustakides [35] that the optimal solution to (4.1) is given by a slightly modified version of the cumulative sum (CUSUM) test proposed by Page [34]. We describe this test later in the chapter.

An alternate formulation of the change detection problem was studied by Pollak [37]. Even here the change point is modeled as a deterministic quantity. However the delay to be minimized is no longer the worst-case delay but a worst-case average delay (also referred to as supremum average

detection delay by some authors) defined by

$$J_{\text{SRP}}(\tau) = \sup_{\lambda \geq 1} \mathsf{E}_{\lambda}^{\nu}[\tau - \lambda | \tau \geq \lambda].$$

The Pollak criterion of optimality of a stopping rule τ for change detection is given by

$$\text{Minimize } J_{\text{SRP}}(\tau) \text{ subject to } \text{FAR}(\tau) \leq \alpha \tag{4.2}$$

where the minimization is over all stopping times τ such that $J_{\text{SRP}}(\tau)$ is well-defined. Pollak [37] established the asymptotic optimality of the Shiryaev-Roberts-Pollak (SRP) stopping rule for (4.2).

Another approach to change detection is the Bayesian formulation of [38, 42]. Here the change-point is modeled as a random variable Λ with prior probability distribution, $\pi_k = \mathsf{P}(\Lambda = k), k = 1, 2, \ldots$. The performance measures are the average detection delay (ADD) and probability of false alarm (PFA) defined by

$$\text{ADD}(\tau) = \mathsf{E}^{\nu}[(\tau - \Lambda)^{+}], \qquad \text{PFA}(\tau) = \mathsf{P}^{\nu}(\tau < \Lambda)$$

where E^{ν} represents the expectation operator and P^{ν} the probability law when the pre-change and post-change distributions are ν_0 and ν_1 respectively. For a given $\alpha \in (0, 1)$, the optimization problem under the Bayesian criterion is:

$$\text{Minimize ADD}(\tau) \text{ subject to PFA}(\tau) \leq \alpha. \tag{4.3}$$

When the prior distribution on the change-point follows a geometric distribution, the optimal solution to the above problem is given by the Shiryaev test [42].

The robust versions of (4.1), (4.2) and (4.3) are intended to capture situations in which one or both of the distributions ν_0 and ν_1 are not known exactly, but are known to belong to uncertainty classes of distributions, $\mathcal{P}_0, \mathcal{P}_1 \subset \mathcal{P}(\mathcal{X})$. The objective is to minimize the worst-case delay amongst all possible values of the unknown distributions, while satisfying the false-alarm constraint for all possible values of the unknown distributions. Thus the robust version of the Lorden criterion is to identify the stopping rule that

solves the following optimization problem:

$$\begin{aligned} \min \quad & \sup_{\nu_0 \in \mathcal{P}_0, \nu_1 \in \mathcal{P}_1} \mathrm{WDD}(\tau) \\ \text{s.t.} \quad & \sup_{\nu_0 \in \mathcal{P}_0} \mathrm{FAR}(\tau) \leq \alpha. \end{aligned} \tag{4.4}$$

Similarly, the robust version of the Pollak criterion is:

$$\begin{aligned} \min \quad & \sup_{\nu_0 \in \mathcal{P}_0, \nu_1 \in \mathcal{P}_1} J_{\mathrm{SRP}}(\tau) \\ \text{s.t.} \quad & \sup_{\nu_0 \in \mathcal{P}_0} \mathrm{FAR}(\tau) \leq \alpha \end{aligned} \tag{4.5}$$

and the robust version of the Bayesian criterion is:

$$\begin{aligned} \min \quad & \sup_{\nu_0 \in \mathcal{P}_0, \nu_1 \in \mathcal{P}_1} \mathrm{ADD}(\tau) \\ \text{s.t.} \quad & \sup_{\nu_0 \in \mathcal{P}_0} \mathrm{PFA}(\tau) \leq \alpha. \end{aligned} \tag{4.6}$$

The optimal stopping rule τ under each of the robust criteria described above has the following minimax interpretation. For any other stopping rule τ' that guarantees the false alarm constraint for all values of unknown distributions from the uncertainty classes, there is at least one pair of distributions such that the delay obtained under τ' will be at least as high as the maximum delay obtained with τ over all pairs of distributions from the uncertainty classes. In the rest of this chapter we provide solutions to the robust problems (4.4), (4.5) and (4.6) when the uncertainty classes satisfy some specific conditions.

4.3 Robust Change Detection

4.3.1 Least favorable distributions

The solution to the robust problem is simplified greatly if we can identify least favorable distributions (LFDs) from the uncertainty classes such that the solution to the robust problem is given by the solution to the non-robust problem designed with respect to the LFDs. LFDs were first identified for a simpler problem - the robust hypothesis testing problem - by Huber and

Strassen in [43] and [54]. It was later shown in [55] that one can identify these LFDs if the uncertainty classes satisfy a joint stochastic boundedness condition. Before we introduce this condition, we need the following notation. If X and X' are two real-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ such that

$$\mathsf{P}(X \geq t) \geq \mathsf{P}(X' \geq t), \text{ for all } t \in \mathbb{R},$$

then we say that the random variable X is *stochastically larger than* [55] the random variable X'. We denote this relation via the notation $X \succ X'$. Equivalently if $X \sim \mu$ and $X' \sim \mu'$, we also denote $\mu \succ \mu'$.

Definition 1 (Joint Stochastic Boundedness) [55]: Consider the pair $(\mathcal{P}_0, \mathcal{P}_1)$ of classes of distributions defined on a measurable space $(\mathcal{X}, \mathcal{F})$. Let $(\overline{\nu}_0, \underline{\nu}_1) \in \mathcal{P}_0 \times \mathcal{P}_1$ be some pair of distributions from this pair of classes such that $\underline{\nu}_1$ is absolutely continuous with respect to $\overline{\nu}_0$. Let L^* denote the log-likelihood ratio between $\underline{\nu}_1$ and $\overline{\nu}_0$ defined as the logarithm of the Radon-Nikodym derivative $\log \frac{d\underline{\nu}_1}{d\overline{\nu}_0}$. Corresponding to each $\nu_j \in \mathcal{P}_j$, we use μ_j to denote the distribution of $L^*(X)$ when $X \sim \nu_j, j = 0, 1$. Similarly we use $\overline{\mu}_0$ (respectively $\underline{\mu}_1$) to denote the distribution of $L^*(X)$ when $X \sim \overline{\nu}_0$ (respectively $\underline{\nu}_1$). The pair $(\mathcal{P}_0, \mathcal{P}_1)$ is said to be jointly stochastically bounded by $(\overline{\nu}_0, \underline{\nu}_1)$ if for all $(\nu_0, \nu_1) \in \mathcal{P}_0 \times \mathcal{P}_1$,

$$\overline{\mu}_0 \succ \mu_0 \text{ and } \mu_1 \succ \underline{\mu}_1. \qquad \blacksquare$$

Loosely speaking, under the joint stochastic boundedness (JSB) condition, the LFD from one uncertainty class is the distribution that is *nearest* to the other uncertainty class. This notion can be made rigorous in terms of Kullback-Leibler divergence and other Ali-Silvey distances between distributions in the uncertainty classes, as shown in [56, Corollary 1].

Huber and Strassen [54] have established a procedure to obtain robust solutions to the Neyman-Pearson hypothesis testing problem provided the uncertainty classes can be described in terms of 2-alternating capacities. As pointed out in [55], any pair of uncertainty classes that can be described in terms of 2-alternating capacities also satisfy the JSB condition (see [54, Theorem 4.1]). This observation suggests that we can identify examples of uncertainty classes which satisfy the joint stochastic boundedness condition using

the results in [54], [55], and [6]. These include ϵ-contamination classes, total variation neighborhoods, Prohorov distance neighborhoods, band classes, and p-point classes. In general it is difficult to identify the distributions $\overline{\nu}_0$ and $\underline{\nu}_1$. However, for ϵ-contamination classes, total variation neighborhoods, and Lévy metric neighborhoods, the method suggested in [6, pp. 241-248] can be used to identify these distributions.

We show that under certain assumptions on $\mathcal{P}_0$ and $\mathcal{P}_1$, the pair of distributions $(\overline{\nu}_0, \underline{\nu}_1)$ are LFDs for the robust change detection problem in (4.4), (4.5) and (4.6). Thus the optimal stopping rules designed *assuming* known pre-change and post-change distributions $\overline{\nu}_0$ and $\underline{\nu}_1$, respectively, are optimal for the robust problems (4.4), (4.5) and (4.6). We use E_m^* to denote the expectation operator and P_m^* to denote the probability law when the change happens at m and the pre-change and post-change distributions are $\overline{\nu}_0$ and $\underline{\nu}_1$, respectively.

We need the following straightforward result. For completeness we provide a proof in Appendix C.

Lemma 4.3.1. *Suppose $\{U_i : 1 \leq i \leq n\}$ is a set of mutually independent random variables, and $\{V_i : 1 \leq i \leq n\}$ is another set of mutually independent random variables such that $U_i \succ V_i, 1 \leq i \leq n$. Now let $h : \mathbb{R}^n \mapsto \mathbb{R}$ be a continuous real-valued function defined on $\mathbb{R}^n$ that satisfies*

$$\begin{aligned} & h(x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_n) \\ & \qquad \geq \quad h(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \end{aligned}$$

for all $x_1^n \in \mathbb{R}^n, a > x_i$, and $i \in \{1, \ldots, n\}$. Then we have,

$$h(U_1, U_2, \ldots, U_n) \succ h(V_1, V_2, \ldots, V_n).$$

□

4.3.2 Lorden criterion

When the distributions ν_0 and ν_1 are known, the solution to (4.1) is given by the CUSUM test [35]. The optimal stopping time is given by

$$\tau_{\mathrm{C}} = \inf\{n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} L^{\nu}(X_i) \geq \eta\} \tag{4.7}$$

where L^{ν} is the log-likelihood ratio between ν_1 and ν_0, and the threshold η is chosen so that, $\mathsf{E}_{\infty}^{\nu}(\tau_{\mathrm{C}}) = \frac{1}{\alpha}$. The following theorem provides a solution to the robust Lorden problem when the distributions are unknown.

Theorem 4.3.2. *Suppose the following conditions hold:*

(i) *The uncertainty classes $\mathcal{P}_0, \mathcal{P}_1$ are jointly stochastically bounded by $(\overline{\nu}_0, \underline{\nu}_1)$.*

(ii) *All distributions $\nu_0 \in \mathcal{P}_0$ are absolutely continuous with respect to $\overline{\nu}_0$; i.e.,*

$$\nu_0 \ll \overline{\nu}_0, \quad \nu_0 \in \mathcal{P}_0. \tag{4.8}$$

(iii) *The function $L^*(.)$, representing the log-likelihood ratio between $\underline{\nu}_1$ and $\overline{\nu}_0$ is continuous over the support of $\overline{\nu}_0$.*

Then the optimal stopping rule that solves (4.4) is given by the following CUSUM test:

$$\tau_{\mathrm{C}}^{*} = \inf\left\{n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} L^{*}(X_i) \geq \eta\right\} \tag{4.9}$$

where the threshold η is chosen so that, $\mathsf{E}_{\infty}^{*}(\tau_{\mathrm{C}}^{*}) = \frac{1}{\alpha}$. □

We prove the theorem in Appendix C. Two brief remarks are in order. Firstly, the discussion in [48, p. 198] suggests that when LFDs exist under our formulation, they also solve the asymptotic problem, as expected. Secondly, the robust CUSUM test admits a simple recursive implementation similar to the ordinary CUSUM test. Clearly,

$$S_{n+1} = S_n^{+} + L^{*}(X_{n+1}) \tag{4.10}$$

where $S_n = \max_{1 \le k \le n} \sum_{i=k}^{n} L^*(X_i)$ is the test statistic appearing in (4.9). Thus it is easy to compute the test statistic recursively.

Asymptotic analysis of the robust CUSUM

In general, for any pair of pre-change and post-change distributions (ν_0, ν_1) from the uncertainty classes, we expect the performance of the robust CUSUM test to be poorer than that of the optimal CUSUM test designed with respect to the correct distributions. The drop in performance can be interpreted as the *cost of robustness*. Although it is not easy to characterize this cost in general, some insight can be obtained by performing an asymptotic analysis in the setting where the false alarm constraint α goes to zero. Our analysis uses the result of [36, Theorem 2] (also see [48, Theorem 6.16]). We use $\mathrm{WDD}^{\nu}(\tau_{\mathrm{C}}^*)$ to denote the worst-case delay obtained by employing the stopping rule τ_{C}^* when the pre-change and post-change distributions are given by ν_0 and ν_1. Similarly, $\mathrm{WDD}^{*}(\tau_{\mathrm{C}}^*)$ is used to denote the same quantity when the pre-change and post-change distributions are the LFDs.

As mentioned in the remark following Theorem 2 in [36], we can interpret the robust CUSUM test as a repeated one-sided sequential probability ratio test (SPRT) between $\underline{\nu}_1$ and $\overline{\nu}_0$. Let τ_{SPRT} denote the stopping rule of the SPRT. We apply [36, Theorem 2] to τ_{SPRT} when the true distributions are the LFDs. It follows that

$$\mathsf{E}_{\infty}^{*}(\tau_{\mathrm{C}}^*) \ge \frac{1}{\alpha}$$

where $B = \frac{1}{\alpha}$ is used as the upper threshold in the SPRT given by τ_{SPRT}. From (C.8), we know that

$$\mathsf{E}_{\infty}^{\nu}(\tau_{\mathrm{C}}^*) \ge \mathsf{E}_{\infty}^{*}(\tau_{\mathrm{C}}^*) \ge \frac{1}{\alpha}.$$

We again apply the theorem to τ_{SPRT}, but with the true distributions given by any $\nu_0 \in \mathcal{P}_0$ and $\nu_1 \in \mathcal{P}_1$. We now have

$$\mathrm{WDD}^{\nu}(\tau_{\mathrm{C}}^*) \le \mathsf{E}(\tau_{\mathrm{SPRT}})$$

where the expression on the right-hand side denotes the expected stopping time of the SPRT when the observations follow distribution ν_1. Now, by applying the well-known Wald's identity [57] as suggested in the remark

following [36, Theorem 2], we obtain

$$\mathsf{E}(\tau_{\text{SPRT}}) = \frac{|\log \alpha|}{I_{\nu_1}}(1 + o(1)), \quad \text{as } \alpha \to 0$$

where $o(1) \to 0$ as $\alpha \to 0$ and

$$I_{\nu_1} = \int L^*(x) d\nu_1(x) = D(\nu_1 \| \overline{\nu}_0) - D(\nu_1 \| \underline{\nu}_1).$$

Thus

$$\text{WDD}^{\nu}(\tau_{\text{C}}^*) \leq \frac{|\log(\alpha)|(1 + o(1))}{D(\nu_1 \| \overline{\nu}_0) - D(\nu_1 \| \underline{\nu}_1)}.$$

It is also known from [36, Theorem 3] that any stopping rule τ that satisfies the false alarm constraint $\text{FAR}(\tau) \leq \alpha$ must satisfy the lower bound

$$\text{WDD}^{\nu}(\tau) \geq \frac{|\log(\alpha)|(1 + o(1))}{D(\nu_1 \| \nu_0)}$$

and that this lower bound is achieved by the optimal CUSUM test between ν_1 and ν_0. Thus, the worst-case delay of the robust test is asymptotically larger by a factor no more than

$$\frac{D(\nu_1 \| \nu_0)}{D(\nu_1 \| \overline{\nu}_0) - D(\nu_1 \| \underline{\nu}_1)}$$

when compared with the delay incurred by the optimal test. This factor is thus an upper bound on the asymptotic cost of robustness.

4.3.3 Pollak criterion

The SRP stopping rule is asymptotically optimal for (4.2). Let R_0^{ν} be a random variable with distribution ψ supported on $\mathbb{R}_+$ and define

$$R_n^{\nu} = L^{\nu}(X_n)(1 + R_{n-1}^{\nu}), \quad n \geq 1. \tag{4.11}$$

When the distributions ν_0 and ν_1 are known, the SRP stopping rule is given by

$$\tau_{\text{SRP}}^{\nu,\eta,\psi} = \inf\left\{n \geq 0 : R_n^{\nu} \geq \eta\right\}. \tag{4.12}$$

Asymptotic optimality property: The SRP test of (4.12) is asymptotically optimal for (4.2) in the following sense [37]: For every $0 < \alpha < 1$ there exists threshold η and probability measure ψ_η such that the stopping rule $\tau_{\text{SRP}} := \tau_{\text{SRP}}^{\nu,\eta,\psi_\eta}$ satisfies $\text{FAR}(\tau_{\text{SRP}}) = \alpha$ and for any other stopping rule τ that satisfies the false alarm constraint $\text{FAR}(\tau) \leq \alpha$, we have

$$J_{\text{SRP}}(\tau) \geq J_{\text{SRP}}(\tau_{\text{SRP}}) + o(1) \tag{4.13}$$

where $o(1) \to 0$ as $\alpha \to 0$.

The following theorem identifies a stopping rule that extends the above asymptotic optimality property to the setting where the post-change distribution is unknown.

Theorem 4.3.3. *Suppose the following conditions hold:*

(i) *The uncertainty class $\mathcal{P}_0$ is a singleton $\mathcal{P}_0 = \{\nu_0\}$ and the pair $(\mathcal{P}_0, \mathcal{P}_1)$ is jointly stochastically bounded by $(\nu_0, \underline{\nu}_1)$.*

(ii) *The function $L^*(.)$, representing the log-likelihood ratio between $\underline{\nu}_1$ and ν_0 is continuous over the support of ν_0.*

Let $\tau^*_{\text{SRP}} := \tau_{\text{SRP}}^{\nu^*,\eta,\psi_\eta}$ denote the SRP stopping rule defined with respect to the LFDs $(\nu_0, \underline{\nu}_1)$, with parameters η and ψ_η chosen such that the asymptotic optimality property of (4.13) is satisfied. Then the stopping rule τ^*_{SRP} is also asymptotically optimal for (4.5) in the following sense: For every $0 < \alpha < 1$ and for any stopping rule τ that satisfies the false alarm constraint FAR$(\tau) \leq \alpha$, we have

$$\sup_{\nu^1 \in \mathcal{P}_1} J^\nu_{\text{SRP}}(\tau) \geq \sup_{\nu^1 \in \mathcal{P}_1} J^\nu_{\text{SRP}}(\tau^*_{\text{SRP}}) + o(1) \tag{4.14}$$

where $o(1) \to 0$ as $\alpha \to 0$. □

The result of (4.14) can be interpreted as follows: The difference between the worst-case values of the delays incurred by the stopping rule τ^*_{SRP} and any other stopping rule τ approaches zero as the false alarm constraint α approaches zero.

Our proof, provided in Appendix C, is useful only when $\mathcal{P}_0$ is a singleton. It is possible that the asymptotic optimality result may still hold even for general $\mathcal{P}_0$, although the current proof is not applicable. We elaborate on

this further in the discussion in the next section on the Bayesian criterion, and also in Appendix C following the proof of the theorem.

We also note that in some cases our proof can be adapted to obtain tests that are exactly optimal for the robust Pollak criterion of (4.5). Polunchenko and Tartakovsky [58] study the Shiryaev-Roberts procedure (SR-r) which is identical to the SRP procedure described earlier, except for the fact that R_0 is not random but fixed at some constant r. Theorem 2 of [58] shows the exact non-asymptotic optimality of the SR-r procedure for detecting a change in distribution from Exp(1) to Exp(2) where Exp(θ) refers to an exponential distribution with mean θ^{-1}. Using that result, the proof of Theorem 4.3.3 can be adapted to obtain the exact robust solution to the optimization problem in (4.5). In particular it can be shown that the SR-r procedure for detecting change from Exp(1) to Exp(2) given in [58, Theorem 2] is also optimal for (4.5) when $\mathcal{P}_0 = \{\text{Exp}(1)\}$ and $\mathcal{P}_1 = \{\text{Exp}(\theta) : \theta \geq 2\}$.

4.3.4 Bayesian criterion

When the distributions ν_0 and ν_1 are known and the prior distribution of the change-point is geometric, the solution to (4.3) is given by the Shiryaev test [42]. Denoting the parameter of the geometric distribution by ρ, we have

$$\pi_k = \rho(1-\rho)^{k-1}, \quad k \geq 1.$$

The Shiryaev stopping rule is based on comparing the posterior probability of change to a threshold η'

$$\tau_{\text{S}} = \inf\left\{n \geq 1 : \mathsf{P}^{\nu}(\Lambda \leq n | \mathcal{F}_n) \geq \eta'\right\}.$$

It can be equivalently expressed as

$$\tau_{\text{S}} = \inf\left\{n \geq 1 : \log(\sum_{k=1}^{n} \pi_k \exp(\sum_{i=k}^{n} L^{\nu}(X_i))) \geq \eta\right\} \tag{4.15}$$

where the threshold η is chosen such that $\text{PFA}(\tau_{\text{S}}) = \mathsf{P}^{\nu}(\tau_{\text{S}} < \Lambda) = \alpha$. The following theorem, proved in Appendix C, identifies a solution to the robust Shiryaev problem (4.6).

Theorem 4.3.4. *Suppose the following conditions hold:*

(i) *The uncertainty class $\mathcal{P}_0$ is a singleton $\mathcal{P}_0 = \{\nu_0\}$ and the pair $(\mathcal{P}_0, \mathcal{P}_1)$ is jointly stochastically bounded by $(\nu_0, \underline{\nu}_1)$.*

(ii) *The prior distribution of the change-point is a geometric distribution.*

(iii) *The function $L^*(.)$, representing the log-likelihood ratio between $\underline{\nu}_1$ and ν_0, is continuous over the support of ν_0.*

Then the optimal stopping rule that solves (4.6) is given by the following Shiryaev test:

$$\tau_{\mathrm{S}}^* = \inf\left\{n \geq 1 : \log(\sum_{k=1}^{n} \pi_k \exp(\sum_{i=k}^{n} L^*(X_i))) \geq \eta\right\} \tag{4.16}$$

where the threshold η is chosen so that $\mathsf{P}^*(\tau_{\mathrm{S}}^* < \Lambda) = \alpha$. □

We note that our results under the Bayesian and Pollak criteria are applicable only when the pre-change distribution is known exactly and hence these results are weaker than our result under the Lorden criterion. Suppose $\mathcal{P}_0$ is not a singleton and $(\mathcal{P}_0, \mathcal{P}_1)$ is jointly stochastically bounded by $(\overline{\nu}_0, \underline{\nu}_1)$. In this case, the stopping rule τ_{S}^* defined with respect to $(\overline{\nu}_0, \underline{\nu}_1)$ is not optimal for the robust Bayesian criterion (4.6). In particular, when the pre-change distribution is $\nu_0 \neq \overline{\nu}_0$ and the post-change distribution is $\nu_1 = \underline{\nu}_1$, it can be shown that the average detection delay $\mathrm{ADD}^{\nu}(\tau_{\mathrm{S}}^*)$ of the stopping rule τ_{S}^* is in general higher than the average detection delay $\mathrm{ADD}^*(\tau_{\mathrm{S}}^*)$ when the pre-change and post-change distributions are $(\overline{\nu}_0, \underline{\nu}_1)$. This is because the likelihood ratios of the pre-change observations appearing in (4.16) are stochastically larger under $\overline{\nu}_0$ than under ν_0. This leads to a stopping time that is stochastically smaller under $(\overline{\nu}_0, \underline{\nu}_1)$ than under $(\nu_0, \underline{\nu}_1)$. Hence there is no reason to believe that τ_{S}^* solves the robust problem (4.6).

Even in the case of the Pollak criterion studied in Section 4.3.3, our robust result holds only when $\mathcal{P}_0$ is a singleton and the JSB condition holds. However, unlike in the Bayesian case, we do not have a simple explanation for why the result cannot be extended to the setting where the pre-change distribution is not known exactly. It is possible that for some specific choices of the uncertainty classes, the stopping rule designed with respect to $(\overline{\nu}_0, \underline{\nu}_1)$ may be asymptotically optimal for the robust problem of (4.5), although we do not expect this to be true in general.

However, such a problem does not arise for the robust CUSUM test we studied in Section 4.3.2, since the worst-case detection delay $\mathrm{WDD}^{\nu}(\tau_{\mathrm{C}}^{*})$ of the robust CUSUM depends only on the support of the pre-change distribution when post-change distribution is kept fixed at $\nu_1 = \underline{\nu}_1$.

Comparison with robust sequential detection It is interesting to compare our results with some known results on robust sequential detection. We have shown that provided the JSB condition and other regularity conditions hold, change detection tests designed with respect to the LFDs exactly solve the minimax robust change detection problem under the Lorden and Bayesian criteria. However, the known minimax optimality results in robust sequential detection are all for the asymptotic settings - as error probabilities go to zero [43] or as the size of the uncertainty classes diminishes [44]. Huber [43] showed that an exact minimax result does not hold for the robust sequential detection problem in general. He provided examples where the expected stopping times of the SPRT designed with respect to the LFDs are not least favorable under the LFDs. This is similar to the reason why the robust Shiryaev test is not optimal for the Bayesian problem when $\mathcal{P}_0$ is not a singleton as explained above.

4.4 Some Examples and Simulation Results

4.4.1 Gaussian mean shift

Here we consider a simple example to illustrate the results. Assume ν_0 is known to be a standard Gaussian distribution with mean zero and unit variance, so that $\mathcal{P}_0$ is a singleton. Let $\mathcal{P}_1$ be the collection of Gaussian distributions with means from the interval $[0.1, 3]$ and unit variance.

$$\begin{aligned} \mathcal{P}_0 &= \{\mathcal{N}(0,1)\} \\ \mathcal{P}_1 &= \{\mathcal{N}(\theta,1) : \theta \in [0.1,3]\} \end{aligned} \tag{4.17}$$

It is easily verified that $(\mathcal{P}_0, \mathcal{P}_1)$ is jointly stochastically bounded by $(\overline{\nu}_0, \underline{\nu}_1)$ given by

$$\overline{\nu}_0 \sim \mathcal{N}(0,1), \qquad \underline{\nu}_1 \sim \mathcal{N}(0.1,1).$$

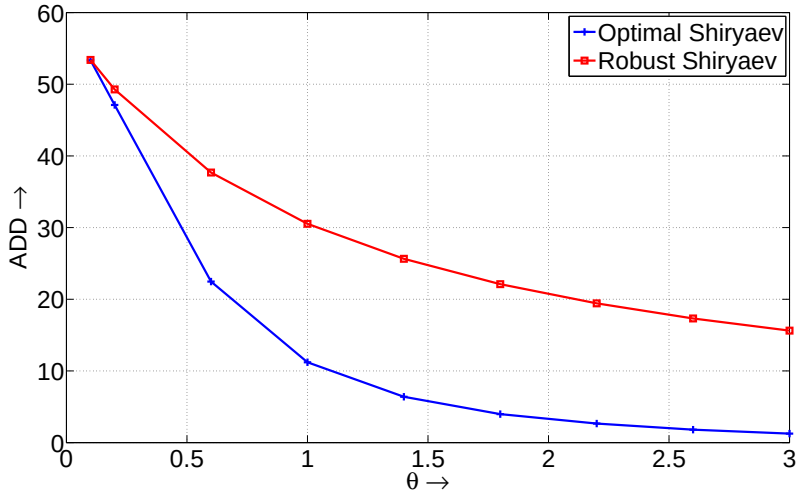

Figure 4.2: *Comparison of robust and non-robust Shiryaev tests for* $\alpha = 0.001$ *for the Gaussian mean shift example.*

Bayesian criterion

We simulated the Bayesian and robust Bayesian change detection tests for this problem assuming a geometric prior distribution for the change-point with parameter 0.1 and a false alarm constraint of $\alpha = 0.001$. From the performance curves plotted in Figure 4.2, we can see that the robust Shiryaev test gives the same average detection delay (ADD) as the optimal Shiryaev test at $\underline{\nu}_1$ which corresponds to $\theta = 0.1$ in the figure. This is expected since the robust test is identical to the optimal test at $\underline{\nu}_1$. For all other values of $\nu_1 \in \mathcal{P}_1$, the performance of the robust test is strictly better than the performance at $\underline{\nu}_1$ and hence this test is indeed minimax optimal. We also see in Figure 4.2 that the average delays obtained with the robust test are much higher than those obtained with the optimal test, especially at high values of the mean θ. The probability of false alarm and average detection were estimated via Monte Carlo simulations with a standard deviation of 0.1% for the estimates.

Lorden criterion and comparison with GLR test

Under the Lorden criterion, we compared the performances of three tests - the optimal CUSUM test with known θ, the robust CUSUM test designed with respect to the LFDs, and the CUSUM test based on the Generalized

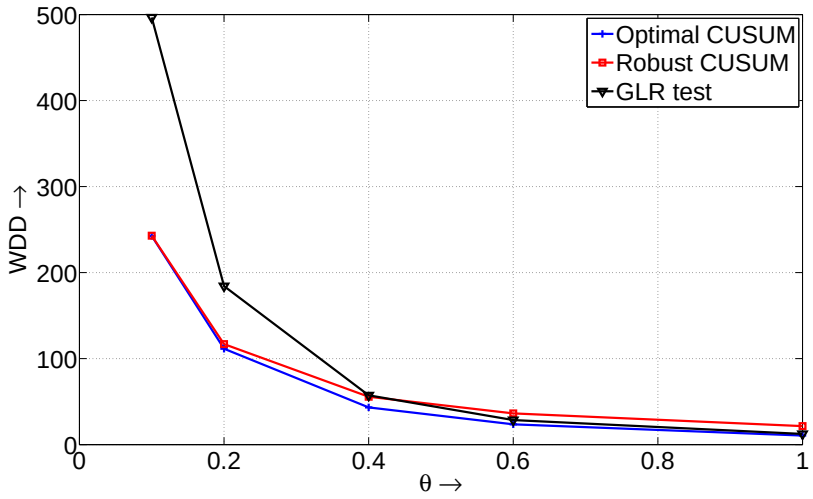

Figure 4.3: *Comparison of various tests for false alarm rate of* $\alpha = 0.001$ *for the Gaussian mean shift example.*

Likelihood Ratio (GLR test) suggested in [36]. The stopping time under the GLR test is given by

$$\tau_{\text{GLR}} = \inf\{n \geq 1 : \max_{1 \leq k \leq n} \sup_{\nu_1 \in \mathcal{P}_1} \sum_{i=k}^{n} L^{\nu}(X_i) \geq \eta\} \tag{4.18}$$

where η is chosen so that the false alarm constraint is met with equality. The GLR test does not require knowledge of θ but still achieves the same asymptotic performance as the optimal CUSUM test with known θ when the false alarm constraint goes to zero for some choices of the uncertainty classes including the example considered above.

Table 4.1: Delays obtained using various tests under the Lorden criterion for a false alarm rate of $\alpha = 0.001$.

θ	Optimal CUSUM	Robust CUSUM	GLR test
0.1	242.7	242.7	496
0.2	111.5	116.8	184
0.4	43.2	55.6	57.2
0.6	23.5	36.3	28.6
1.0	10.5	21.5	12.35

Figure 4.3 and Table 4.1 show estimates of the worst-case detection delay (WDD) obtained under the these tests designed for a false alarm constraint

of $\alpha = 0.001$, for various values of θ. These values are estimated using Monte Carlo simulations. The delay values have a standard deviation lower than 1% and the false alarm value has a standard deviation lower than 3%.

From the performance curves in Figure 4.3 and the values in Table 4.1 we see that the GLR test gives better performance than our robust solution at higher values of θ, and is close to optimal at these high values of θ. However, the robust test gives much better performance than the GLR test at the low values of θ. This is expected since the robust solution is minimax optimal and hence is expected to perform better at the unfavorable values of θ.

An important difference between the two solutions is that although the robust CUSUM test based on the LFDs admits a simple recursive implementation like we described in (4.10), the GLR test is in general very complex to implement. This is because the supremum in (4.18) may be achieved at different values of ν_1 for different n. Furthermore, the optimization in (4.18) may not be easy to solve for general uncertainty classes - particularly non-parametric classes like the ϵ-uncertainty classes considered next.

4.4.2 ϵ-contamination classes

We now discuss an example in which the uncertainty class $\mathcal{P}_0$ is no longer a singleton. For some scalar $\epsilon \in (0, 1)$, consider the following ϵ-contamination classes:

$$\mathcal{P}_0 = \{\nu_0 : \nu_0 = (1-\epsilon)\mathcal{N}(0,1) + \epsilon H_0, \quad H_0 \in \mathcal{P}(\mathbb{R})\} \tag{4.19}$$

$$\mathcal{P}_1 = \{\nu_1 : \nu_1 = (1-\epsilon)\mathcal{N}(1,1) + \epsilon H_1, \quad H_1 \in \mathcal{P}(\mathbb{R})\} \tag{4.20}$$

where $\mathcal{P}(\mathbb{R})$ is the collection of all probability measures on $\mathbb{R}$ and $\mathcal{N}(\mu, \sigma)$ denotes the probability measure corresponding to a Gaussian random variable with mean μ and variance σ^2. In other words, the distributions in uncertainty class $\mathcal{P}_i$ are mixtures of a Gaussian distribution with mean i and unit variance, and an arbitrary probability distribution on $\mathbb{R}$ with weights given by $1-\epsilon$ and ϵ respectively.

Following the method outlined in [43], we identified LFDs for these uncertainty classes and evaluated the performance of the robust test. Let p_i denote the density function of a $\mathcal{N}(\mu, 1)$ random variable and let q_i denote the density function of the least favorable distribution from $\mathcal{P}_i$. It is established in

[43] that the densities of the LFDs have the following structure:

$$q_0(x) = \begin{cases} (1-\epsilon)p_0(x) & \text{if } L(x) \leq b \\ \frac{1-\epsilon}{b}p_1(x) & \text{if } L(x) > b \end{cases} \tag{4.21}$$

$$q_1(x) = \begin{cases} (1-\epsilon)p_1(x) & \text{if } L(x) > a \\ a(1-\epsilon)p_0(x) & \text{if } L(x) \leq a \end{cases} \tag{4.22}$$

where $L(x) = \frac{p_1(x)}{p_0(x)}$. The scalars a and b are identified by the following relation:

$$\begin{aligned} (1-\epsilon)\int_{\{x:L(x)\leq b\}} p_0(x)dx + \frac{1-\epsilon}{b}\int_{\{x:L(x)>b\}} p_1(x)dx &= 1 \\ (1-\epsilon)\int_{\{x:L(x)>a\}} p_1(x)dx + a(1-\epsilon)\int_{\{x:L(x)\leq a\}} p_0(x)dx &= 1. \end{aligned}$$

In order to compare the performance of the robust test with that of the optimal test we chose the following distributions for H_0 and H_1:

$$H_0 = \mathcal{N}(0, \sigma_0), \sigma_0 \in [0.1, 10] \qquad H_1 = \mathcal{N}(1, \sigma_1), \sigma_1 \in [0.1, 10].$$

Table 4.2 shows the values of the worst-case delay (WDD) obtained when σ_0 is kept fixed at $\sigma_0 = 1$ and σ_1 is varied. Shown are the results obtained using the robust CUSUM test as well as the optimal CUSUM test for $\epsilon = 0.05$ and for $\epsilon = 0.005$. We notice that the difference in performance between the robust test and the optimal test is larger for larger values of ϵ. This matches the intuition that the cost of robustness would be higher for a larger uncertainty class of distributions. The delay values and false alarm rates were estimated to have standard deviations lower than 0.1% and 1% respectively.

Table 4.3 shows the values of worst-case delay obtained under the optimal CUSUM tests when σ_1 is kept fixed at $\sigma_1 = 1$ and σ_0 is varied. The delay values and false alarm rates were estimated to have standard deviations lower than 0.1% and 1% respectively. We have not included the delays obtained under the robust test, since the delay of the robust test is invariant with σ_0. The delay obtained under the robust test for $\epsilon = 0.05$ and $\epsilon = 0.005$ are respectively 15.09 and 11.27 as shown in the third row of Table 4.2 corresponding to $\sigma_1 = 1$.

Table 4.2: Delays obtained using various tests under the Lorden criterion for ϵ-uncertainty classes with $\alpha = 0.001$ and $\sigma_0 = 1$.

	$\epsilon = 0.05$		$\epsilon = 0.005$	
σ_1	Robust CUSUM	Optimal CUSUM	Robust CUSUM	Optimal CUSUM
0.1	14.77	9.17	11.27	10.38
0.5	14.86	9.12	11.27	10.39
1	15.09	9.08	11.27	10.35
5	15.52	8.78	11.29	10.33
10	15.59	8.65	11.29	10.34

Table 4.3: Delays obtained using the optimal CUSUM test for ϵ-uncertainty classes with $\alpha = 0.001$ and $\sigma_1 = 1$.

σ_0	Optimal CUSUM for $\epsilon = 0.05$	Optimal CUSUM for $\epsilon = 0.005$
0.1	10.56	10.55
0.5	10.50	10.52
1	10.44	10.56
5	10.02	10.58
10	9.85	10.59

4.5 Summary

In this chapter, we have shown that for uncertainty classes that satisfy some certain stochastic boundedness conditions, the optimal change detectors designed for the least favorable distributions are optimal in a minimax sense. This is shown for the Lorden criterion, the Pollak criterion, and Shiryaev's Bayesian criterion. However, robustness comes at a potential cost. The optimal stopping rule designed for the LFDs may perform quite sub-optimally for other distributions from the uncertainty class when compared with the optimal performance that can be obtained in the case where these distributions are known exactly. Using an asymptotic analysis, we have also obtained an analytic upper bound on this cost of robustness for the robust solution under the Lorden criterion. Nevertheless for some parameter ranges our robust test yields significant performance improvement over the CUSUM test designed for the Generalized Likelihood Ratio statistic, which is a benchmark for the composite quickest change detection problem. Our robust solution also has the added advantage that it can be implemented in a simple recursive man-

ner, while the GLR test does not admit a recursive solution in general, and may require the solution to a complex non-convex optimization problem at every time instant.

CHAPTER 5

CONCLUSION

In this thesis, we have studied various approaches to decision-making under statistical uncertainty. We have focussed on two specific problems, viz., the universal hypothesis testing problem and the robust quickest change detection problem.

In the universal hypothesis testing problem studied in Chapter 2 we demonstrated the improved error performance of the mismatched test over the Hoeffding test when the alternate distribution is known to lie in a parameterized set of distributions. The results suggest that it is important to make use of any available information about the alternate distribution while designing the hypothesis test. The advantage stems from the fact that the mismatched divergence statistic used in the mismatched test has reduced variance compared to the Kullback-Leibler divergence used in the Hoeffding test. We also saw that in these tests, weak convergence results on the test statistic are more accurate than large deviations results in terms of predicting the error probability for finite numbers of samples.

In Chapter 3 we considered two special cases of the universal hypothesis testing problem with uncertainty under the null hypothesis. For these problems, we obtained new weak convergence results that provide guidelines on how to set thresholds that meet a desired false alarm requirement for large sample sizes. These results are analogous to known results for universal hypothesis testing problems without uncertainty under the null hypothesis.

We studied the related problem of quickest change detection under statistical uncertainty in Chapter 4. There we adopted a robust approach of optimizing the worst-case performance. Following Huber [2], we showed that, for some uncertainty classes, it is possible to identify least favorable distributions (LFDs) such that designing the test for the LFDs is robust in a minimax sense. We also demonstrated that for some values of the unknown distributions, our robust test can give substantial performance improvement

over the generalized likelihood ratio test which is the benchmark test for change detection problems with composite hypotheses.

5.1 Extensions

The theoretical framework of mismatched divergence and mismatched tests introduced in Chapters 2 and 3 can be extended in various directions. Although we have restricted our attention to i.i.d observations from a finite alphabet, the approach of mismatched testing is applicable in very general settings, and the performance analysis presented here can be extended to any stationary process satisfying the Central Limit Theorem.

One of the questions left unanswered in Chapter 2 is how to systematically choose the function class $\mathcal{F}$ for specific applications. Some initial work on selecting basis functions for a linear function class is reported in [59]. In [59] the authors consider a finite set of alternate distributions and propose a heuristic scheme for selecting basis functions that ensure good error performance of the resulting mismatched test under all distributions from the set. They do so by optimizing a weighted linear combination of lower bounds on the error exponent corresponding to each alternate distribution from the set. However, in any instance of a universal hypothesis test, one will encounter only observations from one distribution. Hence the weighted optimization scheme might perform poorly if the resultant test gives a poor exponent for the encountered distribution. This can be addressed by adopting the robust approach. The heuristic scheme can be adapted to optimize the worst-case value of the lower bounds rather than a weighted combination of the lower bounds. This robust scheme, however, may require a more complex optimization algorithm.

An alternate direction for future work is the idea of adapting the function class. One could adapt the function class $\mathcal{F}$ based on the length n of the observation sequence. It would be interesting to see whether it is possible to gradually increase the dimensionality d of the function class with n, to obtain a test that gives good performance for all alternate distributions but does not suffer from the high bias and variance of the Hoeffding test.

Another interesting direction for further exploration is to identify the convergence rates of the asymptotic results established in Chapters 2 and 3. In

these chapters we used results on weak convergence of the various test statistics for setting test thresholds for meeting false alarm constraints. The accuracy of these schemes depends on how well the test statistic is approximated by their weak limits, which in turn can be estimated from the convergence rates and sequence length. The first step is to identify the rate of convergence of the basic weak convergence result in Lemma 2.3.6. A potential approach for accomplishing this is to combine Edgeworth expansions [60] that provide rates of convergence of the empirical distribution function in the Central Limit Theorem, with the Taylor's expansion used in the proof of the lemma.

In Chapter 4 we considered the performance evaluation of the proposed robust change detection rules only under the Lorden criterion. It would be useful to obtain similar asymptotic results under the Pollak criterion and Shiryaev's Bayesian criterion as well. A starting point for obtaining such results would be to study the asymptotic analysis of the SRP stopping rule given in [37] and that of the Shiryaev stopping rule given in [61], and see if these results can be extended to obtain performance bounds for the robust version of these stopping rules.

The minimax optimality properties of the robust change detection procedures proposed in Chapter 4 hold only when the uncertainty classes satisfy the JSB condition. There are several practically important uncertainty classes that do not satisfy this condition. An important question that needs to be addressed is whether one can design stopping rules that guarantee some form of robustness in such problems. A potential approach is to use adaptive change detection procedures that attempt to learn the unknown distributions online (see, e.g., [62]).

Most solutions to decision-making under statistical uncertainty reported in this thesis are of a theoretical nature. More work needs to be done to fine-tune the various tests and stopping rules proposed in this thesis for specific applications. The results of Chapter 2 were applied for problems in building energy surveillance in [63].

The results of Chapters 3 and 4 are restricted to uncertainty classes that satisfy strict conditions like joint stochastic boundedness. Further work needs to be done to address more general classes of uncertainty that may be relevant for specific applications.

APPENDIX A

PROOFS OF CHAPTER 2

A.1 Excess Codelength for Source Coding with Training

The results in Theorem 2.3.2 give us the asymptotic behavior of $D(\Gamma^n \| \pi)$ but what we need here is the behavior of $D(\pi \| \Gamma^n)$. Define

$$h(\mu) = \begin{cases} D(\pi\|\mu) & \text{if } \mu \in \mathbb{P}_{\epsilon/2} \\ D(\pi\|\pi^{\mathrm{u}}) & \text{else} \end{cases} .$$

It is clear that h is uniformly bounded from above by $\log \frac{2}{\epsilon}$. Although h is not continuous at the boundary of $\mathbb{P}_{\epsilon/2}$, a modified version of Lemmas 2.3.5 and 2.3.6 can be applied to the function h to establish the results of (2.16) following the same steps used in proving Theorem 2.3.2. The Hessian matrix M appearing in the statement of the lemmas is given by

$$M = \nabla^2 h(\pi) = \ \mathrm{diag}(\pi)^{-1}.$$

Hence, $\mathrm{trace}(M\Omega) = \mathrm{trace}(M\Omega M\Omega) = N - 1$.

A.2 Proof of Lemma 2.3.4

Proof. In the following chain of identities, the first, third and fifth equalities follow from relation (2.25) and Proposition 2.2.6.

$$\begin{aligned}
D_{\mathcal{F}}^{\text{MM}}(\mu\|\pi^0) &= D(\mu\|\pi^0) - \inf\{D(\mu\|\nu) : \nu = \pi^0 \exp(f - \Lambda_{\pi^0}(f)), f \in \mathcal{F}\} \\
&= D(\mu\|\check{\pi}) + \langle \mu, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&\qquad\qquad - \inf\{D(\mu\|\nu) : \nu = \check{\pi} \exp(f - \Lambda_{\check{\pi}}(f)), f \in \mathcal{G}\} \\
&= D_{\mathcal{G}}^{\text{MM}}(\mu\|\check{\pi}) + \langle \mu, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&= D_{\mathcal{G}}^{\text{MM}}(\mu\|\check{\pi}) + \langle \mu - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle + D(\pi^1\|\pi^0) - D(\pi^1\|\check{\pi}) \\
&= D_{\mathcal{G}}^{\text{MM}}(\mu\|\check{\pi}) + \langle \mu - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle + D_{\mathcal{F}}^{\text{MM}}(\pi^1\|\pi^0).
\end{aligned}$$

□

A.3 Proof of Lemma 2.3.5

The following simple lemma will be used in multiple places in the proof that follows.

Lemma A.3.1. *If a sequence of random variables* $\{A^n\}$ *satisfies* $\mathsf{E}[A^n] \xrightarrow[n\to\infty]{} a$ *and* $\{\mathsf{E}[(A^n)^2]\}$ *is a bounded sequence, and another sequence of random variables* $\{B^n\}$ *satisfies* $B^n \xrightarrow[n\to\infty]{m.s.} b$*, then* $\mathsf{E}[A^n B^n] \xrightarrow[n\to\infty]{} ab$*.* □

Proof of Lemma 2.3.5. Without loss of generality, we can assume that the mean $\bar{x}$ is the origin in $\mathbb{R}^m$ and that $h(\bar{x}) = 0$.

Since the Hessian is continuous over the set K, we have by Taylor's theorem:

$$\begin{aligned}
n(h(S^n) - \nabla h(\bar{x})^T S^n)\mathbb{I}_{\{S^n \in K\}} &= n[h(\bar{x}) + \tfrac{1}{2} S^{n\,T} \nabla^2 h(\widetilde{S}^n) S^n]\mathbb{I}_{\{S^n \in K\}} \\
&= \frac{n}{2} S^{n\,T} \nabla^2 h(\widetilde{S}^n) S^n \mathbb{I}_{\{S^n \in K\}} \qquad \text{(A.1)}
\end{aligned}$$

where $\widetilde{S}^n = \gamma S^n$ for some $\gamma = \gamma(n) \in [0, 1]$. By the strong law of large numbers we have $S^n \xrightarrow[n\to\infty]{a.s.} \bar{x}$. Hence $\widetilde{S}^n \xrightarrow[n\to\infty]{a.s.} \bar{x}$ and $\nabla^2 h(\widetilde{S}^n) \xrightarrow[n\to\infty]{a.s.} \nabla^2 h(\bar{x}) = M$

since $\nabla^2 h$ is continuous at $\bar{x}$. Now by the boundedness of the second derivative over K and the fact that

$$\mathbb{I}_{\{S^n \in K\}} \xrightarrow[n\to\infty]{a.s.} 1$$

we have $(\nabla^2 h(\widetilde{S}^n))_{i,j}\mathbb{I}_{\{S^n\in K\}} \xrightarrow[n\to\infty]{m.s.} M_{i,j}$.

Under the assumption that $\boldsymbol{X}$ is i.i.d. on the compact set X, we have

$$\mathsf{E}[nS_i^n S_j^n] = \Sigma_{i,j} \text{ for all } n$$

and $\mathsf{E}[(nS_i^n S_j^n)^2]$ converges to a finite quantity as $n \to \infty$. Hence the results of Lemma A.3.1 are applicable with $A^n = nS_i^n S_j^n$ and $B^n = \nabla^2 h(\widetilde{S}^n)_{i,j}\mathbb{I}_{\{S^n\in K\}}$, which gives:

$$\mathsf{E}[nS_i^n S_j^n \nabla^2 h(\widetilde{S}^n)_{i,j}\mathbb{I}_{\{S^n\in K\}}] \xrightarrow[n\to\infty]{} \Sigma_{i,j}M_{i,j}. \tag{A.2}$$

Thus we have

$$\begin{aligned}\mathsf{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n)\mathbb{I}_{\{S^n\in K\}}] \quad &= \quad \mathsf{E}[\frac{n}{2} S^{n\,T}\nabla^2 h(\widetilde{S}^n)S^n \mathbb{I}_{\{S^n\in K\}}] \\ &\xrightarrow[n\to\infty]{} \quad \tfrac{1}{2}\text{trace}(M\Xi). \end{aligned} \tag{A.3}$$

Since X is compact, h is continuous, and h is differentiable at $\bar{x}$, it follows that there are scalars $\overline{h}$ and $\overline{x}$ such that $\sup_{x\in\mathsf{X}} |h(x)| \le \overline{h}$ and $|\nabla h(\bar{x})^T S^n| < \overline{x}$. Hence,

$$|\mathsf{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n)\mathbb{I}_{\{S^n\notin K\}}]| \le n(\overline{h} + \overline{x})\mathsf{P}\{S^n \notin K\} \xrightarrow[n\to\infty]{} 0 \tag{A.4}$$

where we use the assumption that the $\mathsf{P}\{S^n \notin K\}$ decays exponentially in n. Combining (A.3) and (A.4) and using the fact that S^n has zero mean, we have

$$\mathsf{E}[nh(S^n)] = \mathsf{E}[n(h(S^n) - \nabla h(\bar{x})^T S^n)] \xrightarrow[n\to\infty]{} \tfrac{1}{2}\text{trace}(M\Xi).$$

This establishes the result of (i).

Under the condition that the directional derivative is zero, (A.1) can be written as

$$nf(S^n)\mathbb{I}_{\{S^n\in K\}} = \frac{n}{2} S^{n\,T}\nabla^2 h(\widetilde{S}^n)S^n\mathbb{I}_{\{S^n\in K\}}. \tag{A.5}$$

Now by squaring (A.5), we have

$$(nh(S^n)\mathbb{I}_{\{S^n\in K\}})^2 = \frac{n^2}{4}\sum_{i,j,k,\ell} S_i^n(\nabla^2 h(\widetilde{S}^n))_{i,j}S_j^n S_k^n(\nabla^2 h(\widetilde{S}^n))_{k,\ell}S_\ell^n\mathbb{I}_{\{S^n\in K\}}. \tag{A.6}$$

As before, by the boundedness of the Hessian we have:

$$(\nabla^2 h(\widetilde{S}^n))_{i,j}(\nabla^2 h(\widetilde{S}^n))_{k,\ell}\mathbb{I}_{\{S^n\in K\}} \xrightarrow[n\to\infty]{m.s.} M_{i,j}M_{k,\ell}.$$

It can also be shown that

$$\mathsf{E}[n^2 S_i^n S_j^n S_k^n S_\ell^n] = \frac{F_{i,j,k,l}}{n} + \Sigma_{i,j}\Sigma_{k,\ell} + \Sigma_{j,k}\Sigma_{i,\ell} + \Sigma_{i,k}\Sigma_{j,\ell} \text{ for all } n$$

where $F_{i,j,k,l} = \mathsf{E}[X_i^1 X_j^1 X_k^1 X_\ell^1]$. Moreover, $\mathsf{E}[(n^2 S_i^n S_j^n S_k^n S_\ell^n)^2]$ is finite for each n and converges to a finite quantity as $n\to\infty$ since the moments of X^i are finite. Thus we can again apply Lemma A.3.1 to see that

$$\begin{aligned}&\mathsf{E}[n^2 S_i^n \nabla^2 h(\widetilde{S}^n)_{i,j}S_j^n S_k^n \nabla^2 h(\widetilde{S}^n)_{k,\ell}S_\ell^n\mathbb{I}_{\{S^n\in K\}}]\\ &\xrightarrow[n\to\infty]{} (\Sigma_{i,j}\Sigma_{k,\ell} + \Sigma_{j,k}\Sigma_{i,\ell} + \Sigma_{i,k}\Sigma_{j,\ell})M_{i,j}M_{k,\ell}.\end{aligned} \tag{A.7}$$

Putting together terms and using (A.5) we obtain:

$$\mathsf{E}[(nh(S^n))^2\mathbb{I}_{\{S^n\in K\}}] \xrightarrow[n\to\infty]{} \tfrac{1}{2}\text{trace}(M\Xi M\Xi) + \tfrac{1}{4}(\text{trace}(M\Xi))^2.$$

Now similar to (A.4) we have:

$$|\mathsf{E}[(nh(S^n))^2\mathbb{I}_{\{S^n\notin K\}}]| \le n^2\overline{h}^2\mathsf{P}\{S^n\notin K\} \xrightarrow[n\to\infty]{} 0. \tag{A.8}$$

Consequently

$$\mathsf{E}[(nh(S^n))^2] \xrightarrow[n\to\infty]{} \tfrac{1}{2}\text{trace}(M\Xi M\Xi) + \tfrac{1}{4}(\text{trace}(M\Xi))^2$$

which gives (ii).

□

A.4 Proof of Lemma 2.3.6

We know from (2.2) that Γ^n can be written as an empirical average of i.i.d. vectors. Hence, it satisfies the central limit theorem which says that

$$n^{\frac{1}{2}}(\Gamma^n - \mu) \xrightarrow[n\to\infty]{d.} W \tag{A.9}$$

where the distribution of W is defined below (2.42).

Considering a second-order Taylor's expansion and using the condition on the directional derivative, we have

$$n(h(\Gamma^n) - h(\mu)) = \tfrac{1}{2}n((\Gamma^n - \mu)^T \nabla^2 h(\tilde{\Gamma}^n)(\Gamma^n - \mu))$$

where $\tilde{\Gamma}^n = \gamma\Gamma^n + (1-\gamma)\mu$ for some $\gamma = \gamma(n) \in [0,1]$. We also know by the strong law of large numbers that Γ^n and hence $\tilde{\Gamma}^n$ converge to μ almost surely. By the continuity of the Hessian, we have

$$\nabla^2 h(\tilde{\Gamma}^n) \xrightarrow[n\to\infty]{a.s.} \nabla^2 h(\mu). \tag{A.10}$$

By applying the vector-version of Slutsky's theorem [64], together with (A.9) and (A.10), we conclude that

$$2n((\Gamma^n - \mu)^T \nabla^2 h(\tilde{\Gamma}^n)(\Gamma^n - \mu)) \xrightarrow[n\to\infty]{d.} W^T \nabla^2 h(\mu) W,$$

thus establishing the lemma.

A.5 Proof of Lemma 2.3.7

Proof. The assumption that D is a projection matrix implies that $D^2 = D$. Let $\{u^1, \dots, u^m\}$ denote an orthonormal basis, chosen so that the first K vectors span the range space of D. Hence $Du^i = u^i$ for $1 \le i \le K$, and $Du^i = 0$ for all other i.

Let U denote the unitary matrix whose m columns are $\{u^1, \dots, u^m\}$. Then $\widetilde{V} = UV$ is also an $\mathcal{N}(0, I_m)$ random variable, and hence DV and $D\widetilde{V}$ have the same Gaussian distribution.

To complete the proof we demonstrate that $\|D\widetilde{V}\|^2$ has a chi-squared dis-

tribution: By construction the vector $\widetilde{Y} = D\widetilde{V}$ has components given by

$$\widetilde{Y}_i = \begin{cases} \widetilde{V}_i & 1 \leq i \leq K \\ 0 & K < i \leq m \end{cases}.$$

It follows that $\|\widetilde{Y}\|^2 = \|D\widetilde{V}\|^2 = \widetilde{V}_1^2 + \cdots + \widetilde{V}_K^2$ has a chi-squared distribution with K degrees of freedom. □

A.6 Derivation of Approximate Mismatched Divergence

We know that the mismatched divergence is given by,

$$D^{\text{MM}}(\mu\|\pi^0) = \sup_r \{\mu(f_r) - \Lambda_{\pi^0}(f_r)\}. \tag{A.11}$$

Let $f_{r^0} = 0$ be the function identically equal to zero. The log moment generating function has the following gradient and Hessian at $r = r^0$:

$$\nabla_r \Lambda_{\pi^0}(f_r)\Big|_{r=r^0} = \pi^0(\psi), \qquad \nabla_r^2 \Lambda_{\pi^0}(f_r)\Big|_{r=r^0} = \Sigma_{\pi^0} + \pi(\Phi).$$

Applying a second-order Taylor approximation to the objective function in (A.11), we have

$$\mu(f_r) - \Lambda_{\pi^0}(f_r) \approx (r - r^0)^{\text{T}}(\mu(\psi) - \pi^0(\psi)) - \tfrac{1}{2}(r - r^0)^{\text{T}} M_\mu (r - r^0) \tag{A.12}$$

where

$$M_\mu = \Sigma_{\pi^0} + (\pi(\Phi) - \mu(\Phi)).$$

The approximate mismatched divergence $\hat{D}^{\text{MM}}(\mu\|\pi^0)$ between μ and π^0 is defined as the maximum value of the expression on the right side of (A.12). Assuming that M_μ is invertible we have

$$\begin{aligned} \hat{D}^{\text{MM}}(\mu\|\pi) &= \sup_{r \in Re^d} (r - r^0)^{\text{T}}(\mu(\psi) - \pi^0(\psi)) - \tfrac{1}{2}(r - r^0)^{\text{T}} M_\mu (r - r^0) \\ &= \tfrac{1}{2}(\mu(\psi) - \pi^0(\psi))^{\text{T}} M_\mu^{-1} (\mu(\psi) - \pi^0(\psi)) \\ &= \tfrac{1}{2}(\mu - \pi^0)^{\text{T}} \Psi^{\text{T}} M_\mu^{-1} \Psi (\mu - \pi^0). \end{aligned}$$

APPENDIX B

PROOFS OF CHAPTER 3

B.1 Proof of Theorem 3.1.1

In order to prove the main result, we need the following lemma:

Lemma B.1.1. *Suppose the functions* $\{\psi_i : 0 \leq i \leq d\}$ *are linearly independent over the support of* π. *If* Z_π *denotes the support of* $\pi \in \mathbb{P}$, *then there exists an open neighborhood* $B \subset \mathcal{P}(\mathsf{Z}_\pi)$ *such that for all* $\mu \in B$, *the supremum in (3.4) is achieved at a unique point* $r(\mu)$.

Proof. We verify that the proposition in the lemma holds for $B = \{\mu \in \mathcal{P}(\mathsf{Z}) : |\mu(z) - \pi(z)| < \epsilon \text{ for all } z \in \mathsf{Z}_\pi, \quad \mu(y) = 0 \text{ for all } y \in \mathsf{Z} \setminus \mathsf{Z}_\pi\}$ where $\epsilon = \frac{1}{2}\min_{z \in \mathsf{Z}_\pi} \pi(z)$. Clearly, for all $\mu \in B$, the support of μ is equal to Z_π and hence $0 \leq D^{\text{ROB}}(\mu\|\mathbb{P}) \leq D(\mu\|\pi) < \infty$.

Now since the functions ψ_i are linearly independent over Z_π, and the value of the the optimization problem in (3.4) is finite, it follows that we can restrict the constraint set in (3.4) to a bounded subset of the closed set $\mathcal{R}$. Thus we can restrict the optimization in (3.4) to a compact set. Furthermore, the objective function is strictly concave by the linear independence assumption and hence the conclusion of the lemma follows. □

Proof of Theorem 3.1.1. Suppose $d_\pi = d$. We know by the discussion in Section 2.2.6 that the robust divergence can be expressed as a mismatched divergence defined with respect to a log-linear function class. I.e.,

$$D^{\text{ROB}}(\mu\|\mathbb{P}) = D^{\text{MM}}(\mu\|\pi) = \sup_{f \in \mathcal{F}} \{\mu(f) - \Lambda_\pi(f)\}$$

where $\mathcal{F}$ is the log-linear function class

$$\mathcal{F} = \{\log(1 + r^T\psi) : r \in \mathcal{R}\}.$$

From the conclusions of Lemma B.1.1 it follows that the conditions of Theorem 2.3.2(i) are satisfied with $r^* = 0$ and hence the conclusion of (3.5) follows when $d_\pi = d$.

Now, if $d_\pi < d$ it is clear that we can define a new set of d_π functions $\psi' = \{\psi'_1, \ldots, \psi'_{d_\pi}\}$ such that the span of the functions in ψ' over Z_π is identical to that of the functions in ψ over Z_π, and further guaranteeing the condition that $\{\psi_0, \psi'_1, \ldots, \psi'_{d_\pi}\}$ are linearly independent over Z_π. Hence for all $\mu \in \mathcal{P}(\mathsf{Z}_\pi)$ the formula for the rate function in (3.4) is unaffected by this transformation. Thus the result of (3.5) holds for a general $\pi \in \mathbb{P}$, by the same arguments as before. □

B.2 Proof of Theorem 3.2.1

Proof. We know from (3.12) that for any $G \in \mathcal{F}$,

$$\begin{aligned}
&\mathsf{P}_G\{\sup_{x\in\mathsf{Z}} \sqrt{n}(F_n(x) - F^+(x)) > \delta\} \\
&\quad\leq \mathsf{P}_G\{\sqrt{n}E_n > \delta\} \\
&\quad\leq \mathsf{P}_G\{\sup_{x\in\mathsf{Z}} \sqrt{n}(F_n(x) - F^+(x)) > \delta\} \\
&\qquad + \mathsf{P}_G\{\sup_{x\in\mathsf{Z}} \sqrt{n}[-(F_n(x) - F^-(x))] > \delta\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\max_{G\in\mathcal{F}} \mathsf{P}_G\{\sup_{x\in\mathsf{Z}} \sqrt{n}(F_n(x) - F^+(x)) > \delta\} \\
&\quad\leq \max_{G\in\mathcal{F}} \mathsf{P}_G\{\sqrt{n}E_n > \delta\} \\
&\quad\leq \max_{G\in\mathcal{F}} \mathsf{P}_G\{\sup_{x\in\mathsf{Z}} \sqrt{n}(F_n(x) - F^+(x)) > \delta\} \\
&\qquad + \max_{G\in\mathcal{F}} \mathsf{P}_G\{\sup_{x\in\mathsf{Z}} \sqrt{n}[-(F_n(x) - F^-(x))] > \delta\}
\end{aligned}$$

which simplifies to:

$$\begin{aligned}
&\mathsf{P}_{F^+}\{\sup_{x\in\mathsf{Z}}\sqrt{n}(F_n(x)-F^+(x))>\delta\}\\
&\leq \max_{G\in\mathcal{F}}\mathsf{P}_G\{\sqrt{n}E_n>\delta\}\\
&\leq \mathsf{P}_{F^+}\{\sup_{x\in\mathsf{Z}}\sqrt{n}(F_n(x)-F^+(x))>\delta\}\\
&\quad +\mathsf{P}_{F^-}\{\sup_{x\in\mathsf{Z}}\sqrt{n}[-(F_n(x)-F^-(x))]>\delta\}. \qquad \text{(B.1)}
\end{aligned}$$

It is known by another result of Kolmogorov [33, p. 335] that

$$\mathsf{P}_{F^+}\{\sqrt{n}\sup_{x\in\mathsf{Z}}(F_n(x)-F^+(x))>\delta\}\xrightarrow[n\to\infty]{} p(\delta).$$

Similarly it also follows that

$$\mathsf{P}_{F^-}\{\sqrt{n}\sup_{x\in\mathsf{Z}}(F^-(x)-F_n(x))>\delta\}\xrightarrow[n\to\infty]{} p(\delta).$$

Taking limits in (B.1) and applying the above limiting results we arrive at the claimed result. □

APPENDIX C

PROOFS OF CHAPTER 4

C.1 Proof of Lemma 4.3.1

We prove this claim by induction. For $n = 1$, the claim holds because if $h : \mathbb{R} \mapsto \mathbb{R}$ is a non-decreasing continuous function, we have

$$\begin{aligned} \mathsf{P}(h(U_1) \geq t) &= \mathsf{P}(U_1 \geq \sup\{x : h(x) < t\} \\ &\geq \mathsf{P}(V_1 \geq \sup\{x : h(x) < t\} \\ &= \mathsf{P}(h(V_1) \geq t). \end{aligned}$$

Assume the claim is true for $n = N$ and now consider $n = N + 1$. For any fixed $x_1^N \in \mathbb{R}^N$, since the function h is non-decreasing in each of its components, it follows by the proof for $n = 1$ that

$$h(x_1, x_2, \ldots, x_N, U_{N+1}) \succ h(x_1, x_2, \ldots, x_N, V_{N+1}). \tag{C.1}$$

We further have

$$\begin{aligned} &\mathsf{P}(h(U_1, U_2, \ldots, U_{N+1}) \geq t) \\ &= \int f_{U_1^N}(x_1^N)\mathsf{P}(h(x_1, x_2, \ldots, x_N, U_{N+1}) \geq t)dx_1^N \\ &\geq \int f_{U_1^N}(x_1^N)\mathsf{P}(h(x_1, x_2, \ldots, x_N, V_{N+1}) \geq t)dx_1^N \qquad \text{(C.2)} \\ &= \mathsf{P}(h(\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_N, V_{N+1}) \geq t) \qquad \text{(C.3)} \\ &= \int f_{V_{N+1}}(y)\mathsf{P}(h(\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_N, y) \geq t)dy \\ &\geq \int f_{V_{N+1}}(y)\mathsf{P}(h(V_1, V_2, \ldots, V_N, y) \geq t)dy \qquad \text{(C.4)} \\ &= \mathsf{P}(h(V_1, V_2, \ldots, V_{N+1}) \geq t) \end{aligned}$$

where (C.2) is obtained via (C.1). The variables $\tilde{U}_i$ appearing in (C.3) are

random variables with exact same statistics as U_i and independent of V_i's. The inequality of (C.4) is obtained by using the induction hypothesis for $n = N$. Thus we have shown that

$$h(U_1, U_2, \ldots, U_{N+1}) \succ h(V_1, V_2, \ldots, V_{N+1})$$

which proves the lemma by the principle of mathematical induction. □

C.2 Proof of Theorem 4.3.2

Proof. Suppose $\mathcal{P}_0$ and $\mathcal{P}_1$ satisfy the conditions of the theorem. Since the CUSUM test is optimal for known distributions, it is clear that the test given in (4.9) is optimal when the pre- and post-change distributions are $\overline{\nu}_0$ and $\underline{\nu}_1$, respectively. Hence, it suffices to show that the values of $\mathrm{WDD}(\tau_{\mathrm{C}}^*)$ and $\mathrm{FAR}(\tau_{\mathrm{C}}^*)$ obtained under any $\nu_0 \in \mathcal{P}_0$ and any $\nu_1 \in \mathcal{P}_1$, are no higher than their respective values when the pre- and post-change distributions are $\overline{\nu}_0$ and $\underline{\nu}_1$. We use Y_i^* to denote the random variable $L^*(X_i)$ when the pre-change and post-change distributions of the observations from the sequence $\{X_i : i = 1, 2, \ldots\}$ are $\overline{\nu}_0$ and $\underline{\nu}_1$, respectively, and Y_i^ν to denote the random variable $L^*(X_i)$ when the pre- and post-change distributions are ν_0 and ν_1, respectively. We first prove the theorem for a special case.

Case 1: $\mathcal{P}_0$ is a singleton given by $\mathcal{P}_0 = \{\nu_0\}$.

Clearly, in this case $\overline{\nu}_0 = \nu_0$ and (4.8) is met trivially. Furthermore, in this case, the false alarm constraint is also met trivially since the false alarm rate obtained by using the stopping rule τ_{C}^* is independent of the true value of the post-change distribution. Fix the change-point to be λ. Now, to complete the proof for the scenario where $\mathcal{P}_0$ is a singleton, we will show that for all $\lambda \geq 1$,

$$\mathsf{E}_\lambda^*[(\tau_{\mathrm{C}}^* - \lambda + 1)^+ | \mathcal{F}_{\lambda-1}] \succ \mathsf{E}_\lambda^\nu[(\tau_{\mathrm{C}}^* - \lambda + 1)^+ | \mathcal{F}_{\lambda-1}] \tag{C.5}$$

which will establish that the value of $\mathrm{WDD}(\tau_{\mathrm{C}}^*)$, obtained under any $\nu_1 \in \mathcal{P}_1$, is no higher than the value when the true post-change distribution is $\underline{\nu}_1$.

Since we now have $\overline{\nu}_0 = \nu_0$, both Y_i^* and Y_i^ν have the same distributions for $i < \lambda$ and hence we assume without loss of generality that for all $i < \lambda$,

$Y_i^* = Y_i^\nu$ with probability one. Under this assumption, we will show that for all integers $N \geq 0$, the following relation holds with probability one:

$$\begin{aligned} \mathsf{P}_\lambda^*((\tau_{\mathrm{C}}^* - \lambda + 1)^+ \leq N | \mathcal{F}_{\lambda-1}) \\ \leq \quad \mathsf{P}_\lambda^\nu((\tau_{\mathrm{C}}^* - \lambda + 1)^+ \leq N | \mathcal{F}_{\lambda-1}) \end{aligned} \tag{C.6}$$

which will then establish (C.5). Since τ_{C}^* is a stopping time, the event $\{(\tau_{\mathrm{C}}^* - \lambda + 1)^+ \leq 0\}$ is $\mathcal{F}_{\lambda-1}$-measurable. Hence, with probability one, (C.6) holds with equality for $N = 0$. Now it suffices to verify (C.6) for $N \geq 1$. We know by the stochastic ordering condition on $\mathcal{P}_1$ that

$$Y_i^\nu \succ Y_i^*, \text{ for all } i \geq \lambda \tag{C.7}$$

Now we have the following equivalence between two events:

$$\begin{aligned} \{\tau_{\mathrm{C}}^* \leq N\} &= \left\{ \max_{1 \leq n \leq N} \max_{1 \leq k \leq n} \sum_{i=k}^{n} L^*(X_i) \geq \eta \right\} \\ &= \left\{ \max_{1 \leq k \leq n \leq N} \sum_{i=k}^{n} L^*(X_i) \geq \eta \right\}. \end{aligned}$$

It is easy to see that the function

$$f(x_1, \ldots, x_N) \triangleq \max_{1 \leq k \leq n \leq N} \sum_{i=k}^{n} x_i$$

is continuous and non-decreasing in each of its components as required by Lemma 4.3.1. Hence for $N \geq 1$, the following hold with probability one:

$$\begin{aligned} \mathsf{P}_\lambda^*((\tau_{\mathrm{C}}^* - \lambda + 1)^+ \leq N | \mathcal{F}_{\lambda-1}) \\ &= \mathsf{P}_\lambda^*(\tau_{\mathrm{C}}^* \leq N + \lambda - 1 | \mathcal{F}_{\lambda-1}) \\ &= \mathsf{P}_\lambda(f(Y_1^*, \ldots, Y_{N+\lambda-1}^*) \geq \eta | \mathcal{F}_{\lambda-1}) \\ &\leq \mathsf{P}_\lambda(f(Y_1^\nu, \ldots, Y_{N+\lambda-1}^\nu) \geq \eta | \mathcal{F}_{\lambda-1}) \\ &= \mathsf{P}_\lambda^\nu(\tau_{\mathrm{C}}^* \leq N | \mathcal{F}_{\lambda-1}) \\ &= \mathsf{P}_\lambda^\nu((\tau_{\mathrm{C}}^* - \lambda + 1)^+ \leq N | \mathcal{F}_{\lambda-1}) \end{aligned}$$

where the inequality follows from Lemma 4.3.1 and (C.7), using the fact that f is a non-decreasing function with respect to its last N arguments and the fact that $Y_i^\nu = Y_i^*$ for $i < \lambda$. Thus, for all integers $N \geq 0$, (C.6) holds with

probability one and hence (C.5) is satisfied. This proves the result for the case where $\mathcal{P}_0$ is a singleton.

Case 2: $\mathcal{P}_0$ is any class of distributions satisfying (4.8).

Suppose that the change does not occur. Then we know by the stochastic ordering condition on $\mathcal{P}_0$ that $Y_i^* \succ Y_i^\nu$ for all i. It follows by Lemma 4.3.1 that

$$\begin{aligned} \mathsf{P}_\infty^*(\tau_{\mathrm{C}}^* \leq N) &= \mathsf{P}_\infty(f(Y_1^*, \ldots, Y_N^*) \geq \eta) \\ &\geq \mathsf{P}_\infty(f(Y_1^\nu, \ldots, Y_N^\nu) \geq \eta) \\ &= \mathsf{P}_\infty^\nu(\tau_{\mathrm{C}}^* \leq N). \end{aligned}$$

Since the above relation holds for all $N \geq 1$, we have

$$\mathsf{E}_\infty^\nu(\tau_{\mathrm{C}}^*) \geq \mathsf{E}_\infty^*(\tau_{\mathrm{C}}^*) = \frac{1}{\alpha} \tag{C.8}$$

and hence the value of $\mathrm{FAR}(\tau_{\mathrm{C}}^*)$ is no higher than α for all values of $\nu_0 \in \mathcal{P}_0$ and $\nu_1 \in \mathcal{P}_1$.

Now suppose the change-point is fixed at λ. A useful observation is that for any given stopping rule τ and fixed post-change distribution ν_1, the random variable $\mathsf{E}_\lambda^{\nu_0,\nu_1}[(\tau - \lambda + 1)^+ | \mathcal{F}_{\lambda-1}]$ is a fixed deterministic function of the random observations $(X_1, \ldots, X_{\lambda-1})$, irrespective of the distribution ν_0. Thus the essential supremum of this random variable depends only on the support of ν_0. Applying this observation to the stopping rule τ_{C}^*, and using the relation (4.8), we have for all $\nu_0 \in \mathcal{P}_0, \nu_1 \in \mathcal{P}_1$,

$$\begin{aligned} &\text{ess sup } \mathsf{E}_\lambda^{\nu_0,\nu_1}[(\tau_{\mathrm{C}}^* - \lambda + 1)^+ | \mathcal{F}_{\lambda-1}] \\ &\qquad \leq \quad \text{ess sup } \mathsf{E}_\lambda^{\overline{\nu}_0,\nu_1}[(\tau_{\mathrm{C}}^* - \lambda + 1)^+ | \mathcal{F}_{\lambda-1}]. \end{aligned}$$

We also know from *Case 1* above that for all $\nu_1 \in \mathcal{P}_1$,

$$\begin{aligned} &\text{ess sup } \mathsf{E}_\lambda^{\overline{\nu}_0,\nu_1}[(\tau_{\mathrm{C}}^* - \lambda + 1)^+ | \mathcal{F}_{\lambda-1}] \\ &\qquad \leq \quad \text{ess sup } \mathsf{E}_\lambda^*[(\tau_{\mathrm{C}}^* - \lambda + 1)^+ | \mathcal{F}_{\lambda-1}]. \end{aligned}$$

Taking the supremum over $\lambda \geq 1$, it follows from the above two relations that the value of $\mathrm{WDD}(\tau_{\mathrm{C}}^*)$ under any pair of distributions $(\nu_0, \nu_1) \in \mathcal{P}_0 \times \mathcal{P}_1$ is no larger than that under $(\overline{\nu}_0, \underline{\nu}_1)$. Thus τ_{C}^* solves the robust problem (4.4). $\square$

C.3 Proof of Theorem 4.3.3

Proof. Let $\tau^*_{\text{SRP}} := \tau^{\nu^*,\eta,\psi_\eta}_{\text{SRP}}$ denote the SRP stopping rule defined with respect to the LFDs $(\nu_0, \underline{\nu}_1)$ satisfying the asymptotic optimality property of (4.13) as mentioned in the statement of the theorem. It is easy to see that for any integers $\lambda \geq 1$ and $N \geq 1$, we have

$$\begin{aligned}
&\mathsf{P}^\nu_\lambda(\tau^*_{\text{SRP}} - \lambda \leq N | \tau^*_{\text{SRP}} \geq \lambda, R^*_0 = r) \\
&\qquad = \quad \frac{\mathsf{P}^\nu_\lambda(\{\tau^*_{\text{SRP}} - \lambda \leq N\} \cap \{\tau^*_{\text{SRP}} \geq \lambda\} | R^*_0 = r)}{\mathsf{P}^\nu_\lambda(\tau^*_{\text{SRP}} \geq \lambda | R^*_0 = r)}
\end{aligned}$$

where R^*_0 denotes the random variable with distribution ψ_η used for initializing the iteration in (4.11). We follow the same steps as in the proof of Theorem 4.3.2. Let Y^ν_i denote the random variable $L^*(X_i)$ when the pre-change and post-change distributions are ν_0 and ν_1 respectively. Since τ^*_{SRP} is a stopping time the event $\{\tau^*_{\text{SRP}} \geq \lambda\}$ is measurable with respect to the pre-change observations and hence we can represent this event as

$$\{\tau^*_{\text{SRP}} \geq \lambda\} = \{(Y^\nu_1, Y^\nu_2, \ldots, Y^\nu_{\lambda-1}) \in T\}$$

where T is the set of pre-change trajectories corresponding to the event $\{\tau^*_{\text{SRP}} \geq \lambda\}$. Now, for any $r \in \mathbb{R}_+$, let $f(t|r)$ denote the conditional probability density function of the random vector $(Y^\nu_1, Y^\nu_2, \ldots, Y^\nu_{\lambda-1})$ evaluated at t conditioned on $R^*_0 = r$. Then we can express the conditional distribution of the delay as

$$\begin{aligned}
&\mathsf{P}^\nu_\lambda(\tau^*_{\text{SRP}} - \lambda \leq N | \tau^*_{\text{SRP}} \geq \lambda, R^*_0 = r) \\
&\qquad = \quad \frac{\int_T \mathsf{P}_\lambda(h_t(Y^\nu_\lambda, Y^\nu_{\lambda+1}, \ldots, Y^\nu_{\lambda+N}) \geq g(t, N, \eta) | R^*_0 = r) f(t|r) dt}{\int_T f(t|r) dt} \qquad \text{(C.9)}
\end{aligned}$$

such that for all $t \in T$, the function $h_t : \mathbb{R}^{N-\lambda+1} \mapsto \mathbb{R}$ satisfies the requirements of Lemma 4.3.1, and $g(t, N, \eta)$ is some real-valued function. The exact form of function h_t can be obtained from the iterations of (4.11) used to define the SRP stopping rule of (4.12). We note that in (C.9) the post-change distribution ν_1 affects only the first term under the integral in the numerator.

Thus, it follows by applying Lemma 4.3.1 that

$$\begin{aligned}\mathsf{P}^*_\lambda(\tau^*_{\text{SRP}} - \lambda \le N | \tau^*_{\text{SRP}} \ge \lambda, R^*_0 = r) \\ \le \mathsf{P}^\nu_\lambda(\tau^*_{\text{SRP}} - \lambda \le N | \tau^*_{\text{SRP}} \ge \lambda, R^*_0 = r)\end{aligned} \tag{C.10}$$

for all $\nu_1 \in \mathcal{P}_1$. Hence it further follows that

$$\sup_{\nu_1 \in \mathcal{P}_1} \mathsf{E}^\nu_\lambda(\tau^*_{\text{SRP}} - \lambda | \tau^*_{\text{SRP}} \ge \lambda) = \mathsf{E}^*_\lambda(\tau^*_{\text{SRP}} - \lambda | \tau^*_{\text{SRP}} \ge \lambda). \tag{C.11}$$

We also observe that for any stopping rule τ that satisfies the false alarm constraint $\text{FAR}(\tau) \le \alpha$, we have

$$\begin{aligned}\sup_{\nu_1 \in \mathcal{P}_1} \sup_{\lambda \ge 1} \mathsf{E}^\nu_\lambda(\tau - \lambda | \tau \ge \lambda) \\ &\ge \sup_{\lambda \ge 1} \mathsf{E}^*_\lambda(\tau - \lambda | \tau \ge \lambda) \\ &\ge \sup_{\lambda \ge 1} \mathsf{E}^*_\lambda(\tau^*_{\text{SRP}} - \lambda | \tau^*_{\text{SRP}} \ge \lambda) + o(1) \\ &= \sup_{\nu_1 \in \mathcal{P}_1} \sup_{\lambda \ge 1} \mathsf{E}^\nu_\lambda(\tau^*_{\text{SRP}} - \lambda | \tau^*_{\text{SRP}} \ge \lambda) + o(1)\end{aligned}$$

where the second relation follows from the fact that τ^*_{SRP} satisfies the asymptotic optimality of (4.13) when the true post-change distribution is $\underline{\nu}_1$, and the last equality follows from (C.11). This completes the proof of the theorem. □

We note that if the robust SRP stopping rule τ^*_{SRP} is used when $\mathcal{P}_0$ is not a singleton, the crucial step of (C.10) does not hold for $\nu_0 \ne \overline{\nu}_0$ and $\nu_1 = \underline{\nu}_1$. Thus our proof of optimality of the robust SRP stopping rule does not hold when the pre-change distribution is unknown.

C.4 Proof of Theorem 4.3.4

Proof. The proof is very similar to that of *Case 1* in Theorem 4.3.2. Since the Shiryaev test is optimal for known distributions, it is clear that the test given in (4.16) is optimal under the Bayesian criterion when the post-change distribution is $\underline{\nu}_1$. Also from the definition of $\text{PFA}(\tau^*_{\text{S}})$ it is clear that the probability of false alarm depends only on the pre-change distribution and

hence the constraint in (4.6) is met by the stopping time τ_{S}^*. Hence, it suffices to show that the value of $\mathrm{ADD}(\tau_{\mathrm{S}}^*)$ obtained under any $\nu_1 \in \mathcal{P}_1$, is no higher than the value when the true post-change distribution is $\underline{\nu}_1$.

Let us first fix $\Lambda = \lambda$. We know by the stochastic ordering condition that conditioned on $\Lambda = \lambda$, for all $i \geq \lambda$, we have $Y_i^{\nu} \succ Y_i^*$ where Y_i^* and Y_i^{ν} are as defined in the proof of Theorem 4.3.2. As before, the function,

$$f'(x_1, \ldots, x_N) \triangleq \max_{1 \leq n \leq N} \log \left(\sum_{k=1}^{n} \pi_k \exp(\sum_{i=k}^{n} x_i) \right)$$

is continuous and non-decreasing in each of its components as required by Lemma 4.3.1. Using these facts, we can show the following by proceeding exactly as in the proof of Theorem 4.3.2: Conditioned on $\Lambda = \lambda$,

$$\mathsf{E}_{\lambda}^*((\tau_{\mathrm{S}}^* - \lambda)^+ | \mathcal{F}_{\lambda-1}) \succ \mathsf{E}_{\lambda}^{\nu}((\tau_{\mathrm{S}}^* - \lambda)^+ | \mathcal{F}_{\lambda-1}).$$

Thus, we have $\mathsf{E}_{\lambda}^*((\tau_{\mathrm{S}}^* - \lambda)^+) \geq \mathsf{E}_{\lambda}^{\nu}((\tau_{\mathrm{S}}^* - \lambda)^+)$ and by averaging over λ, we get

$$\mathsf{E}^*((\tau_{\mathrm{S}}^* - \Lambda)^+) \geq \mathsf{E}^{\nu}((\tau_{\mathrm{S}}^* - \Lambda)^+).$$

□

REFERENCES

[1] 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. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[2] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, no. 6, pp. 1753–1758, 1965.

[3] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–408, 1965.

[4] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?" *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1597–, 1992.

[5] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed., ser. Springer Texts in Electrical Engineering. New York: Springer-Verlag, 1994, a Dowden & Culver Book.

[6] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. New York: Springer, 2008.

[7] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[8] P. J. Huber and V. Strassen, "Minimax tests and the Neyman-Pearson lemma for capacities," *Ann. Statist.*, vol. 1, pp. 251–263, 1973.

[9] P. J. Huber and V. Strassen, "Correction: "Minimax tests and the Neyman-Pearson lemma for capacities" (Ann. Statist. **1** (1973), 251–263)," *Ann. Statist.*, vol. 2, pp. 223–224, 1974.

[10] S. Kassam and H. Poor, "Robust techniques for signal processing: A survey," *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433 – 481, March 1985.

[11] S. Morgenthaler, "A survey of robust statistics," *Statistical Methods and Applications*, vol. 15, no. 3, pp. 271–293, February 2007.

[12] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1986.

[13] K. Zhou and J. C. Doyle, *Essentials of Robust Control*. Upper Saddle River, NJ 07458, USA: Prentice-Hall, 1998.

[14] J. Unnikrishnan and V. Veeravalli, "Algorithms for dynamic spectrum access with learning for cognitive radio," *IEEE Trans. Signal Proc.*, vol. 58, no. 2, pp. 750 –760, Feb. 2010.

[15] E. Levitan and N. Merhav, "A competitive Neyman-Pearson approach to universal hypothesis testing with applications," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2215–2229, 2002.

[16] D. Huang, J. Unnikrishnan, S. Meyn, V. Veeravalli, and A. Surana, "Statistical SVMs for robust detection, supervised learning, and universal classification," in *Proc. of IEEE Information Theory Workshop on Networking and Information Theory*, Volos, Greece, 2009, pp. 62–66.

[17] J. Unnikrishnan, D. Huang, S. Meyn, A. Surana, and V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," 2010, submitted to IEEE Trans. Inform. Theory. [Online]. Available: http://arxiv.org/abs/0909.2234

[18] A. Lapidoth, "Mismatched decoding and the multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 5, pp. 1439–1452, Sep 1996.

[19] E. Abbe, M. Medard, S. Meyn, and L. Zheng, "Finding the best mismatched detector for channel coding and hypothesis testing," *Information Theory and Applications Workshop, 2007*, pp. 284–288, 29 2007-Feb. 2 2007.

[20] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 2004.

[21] C. Pandit, S. Meyn, and V. Veeravalli, "Asymptotic robust Neyman-Pearson hypothesis testing based on moment classes," in *Proc. of IEEE International Symposium on Information Theory*, Chicago, 2004, p. 220.

[22] C. Pandit and S. P. Meyn, "Worst-case large-deviations with application to queueing and information theory," *Stoch. Proc. Applns.*, vol. 116, no. 5, pp. 724–756, May 2006.

[23] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time. I. II," *Comm. Pure Appl. Math.*, vol. 28, pp. 1–47; ibid. **28** (1975), 279–301, 1975.

[24] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *CoRR*, vol. abs/0809.0853, 2008.

[25] B. Clarke and A. R. Barron, "Information theoretic asymptotics of Bayes' methods," Univ. of Illinois, Department of Statistics, Tech. Rep. 26, July 1989.

[26] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statistics*, vol. 9, pp. 60–62, 1938.

[27] P. Harremoës, "Testing goodness-of-fit via rate distortion," in *Proc. of IEEE Information Theory Workshop on Networking and Information Theory*, Volos, Greece, June 2009, pp. 17–21.

[28] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 285–290, 1991.

[29] O. Zeitouni and M. Gutman, "Correction to: "On universal hypotheses testing via large deviations"," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, part 1, p. 698, 1991.

[30] E. Zhou, M. Fu, and S. Marcus, "Solving continuous-state POMDPs via density projection," 2010, to appear in *IEEE Trans. Autom. Control*.

[31] D. Huang, "Mismatched divergence and universal hypothesis testing," M.S. thesis, University of Illinois at Urbana-Champaign, 2009.

[32] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, no. 3, pp. 453–471, 1990.

[33] R. M. Dudley, *Uniform Central Limit Theorems*. New York: Cambridge University Press, 1999.

[34] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[35] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, Dec. 1986.

[36] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, 1971.

[37] M. Pollak, "Optimal detection of a change in distribution," *Ann. Statist.*, vol. 13, no. 1, pp. 206–227, 1985.

[38] A. N. Shiryaev, *Optimal Stopping Rules*. New York: Springer-Verlag, 1978.

[39] A. G. Tartakovsky, *Sequential Methods in the Theory of Information Systems*. Moscow: Radio i Svyaz', 1991.

[40] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes: Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[41] B. E. Brodsky and B. Darkhovsky, *Nonparametric Methods in Change-Point Problems*. Dordrecht: Kluwer, 1993.

[42] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Prob. and App.*, vol. 8, pp. 22–46, 1963.

[43] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, pp. 1753–1758, 1965.

[44] P. X. Quang, "Robust sequential testing," *Ann. Statist.*, vol. 13, no. 2, pp. 638–649, June 1985.

[45] J. Unnikrishnan, V. V. Venugopal, and S. Meyn, "Least favorable distributions for robust quickest change detection," in *Proc. of 2009 IEEE International Symposium on Information Theory, Seoul, Korea*, 2009.

[46] J. Unnikrishnan, V. V. Veeravalli, and S. Meyn, "Minimax robust quickest change detection," 2010, submitted to IEEE Trans. Inform. Theory. [Online]. Available: http://arxiv.org/abs/0911.2551

[47] R. Crow and S. Schwartz, "On robust quickest change detection procedures," in *Proc., IEEE International Symposium on Information Theory*, Trondheim, Norway, 1994, p. 258.

[48] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, UK: Cambridge University Press, 2009.

[49] A. G. Tartakovsky and A. Polunchenko, "Quickest changepoint detection in distributed multisensor systems under unknown parameters," in *Proc., 11th International Conference on Information Fusion*, Cologne, Germany, July 2008.

[50] L. Gordon and M. Pollak, "A robust surveillance scheme for stochastically ordered alternatives," *Ann. Statist.*, vol. 23, no. 4, pp. 1350–1375, 1995.

[51] T. Lai, "Sequential analysis: some classical problems and new challenges," *Stat. Sin.*, vol. 11, pp. 303–408, 2001.

[52] D. Siegmund and E. S. Venkatraman, "Using the generalized likelihood ratio statistic for sequential detection of a change-point," *Ann. Statist.*, vol. 23, no. 1, pp. 255–271, 1995.

[53] B. Brodsky and B. Darkhovsky, "Asymptotically optimal sequential change-point detection under composite hypotheses," in *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05. 44th IEEE Conference on*, Dec. 2005, pp. 7347–7351.

[54] P. J. Huber and V. Strassen, "Minimax tests and the neyman-pearson lemma for capacities," *Ann. Statist.*, vol. 1, no. 2, pp. 251–263, 1973.

[55] V. V. Veeravalli, T. Başar, and H. V. Poor, "Minimax robust decentralized detection," *IEEE Trans. Inform. Theory*, vol. 40, no. 1, pp. 35–40, Jan. 1994.

[56] H. Poor, "Robust decision design using a distance criterion," *IEEE Trans. Inform. Theory*, vol. 26, no. 5, pp. 575–587, Sept. 1994.

[57] A. Wald, *Sequential Analysis*. New York: Wiley, 1947.

[58] A. S. Polunchenko and A. G. Tartakovsky, "On optimality of the shiryaev-roberts procedure for detecting a change in distribution," December 2009, to appear in the Annals of Statistics.

[59] D. Huang and S. Meyn, "Feature extraction for universal hypothesis testing via rank-constrained optimization," presented at International Symposium on Information Theory, Austin, 2010. [Online]. Available: http://arxiv.org/abs/1001.3090

[60] W. Feller, *An Introduction to Probability Theory and its Applications Vol II*, 2nd ed. New York: John Wiley & Sons, 1971.

[61] A. Tartakovsky and V. Veeravalli, "General asymptotic bayesian theory of quickest change detection," *SIAM Theory of Probability and its Applications*, vol. 49, no. 3, pp. 458–497, 2005.

[62] C. Li, H. Dai, and H. Li, "Adaptive quickest change detection with unknown parameter," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3241 –3244.

[63] S. Meyn, A. Surana, Y. Lin, and S. Narayanan, "Anomaly detection using projective Markov models in a distributed sensor network," in *Proc. of 48th IEEE Conference on Decision and Control*, Shanghai, December 2009, pp. 4662–4669.

[64] P. Billingsley, *Convergence of Probability Measures*. New York: John Wiley & Sons, 1968.

AUTHOR'S BIOGRAPHY

Jayakrishnan Unnikrishnan was born on September 18, 1983. He received his B. Tech in electrical engineering from the Indian Institute of Technology, Madras, in 2005, and M.S. in electrical and computer engineering from the University of Illinois in 2007. Since August 2005 he has been a graduate research assistant at the Coordinate Science Laboratory, University of Illinois. He received the Vodafone Graduate Fellowship Award from the University of Illinois at Urbana-Champaign for 2007-2008 and the E. A. Reid Fellowship Award from the ECE department at University of Illinois at Urbana-Champaign for 2010-2011.