MISMATCHED DIVERGENCE AND UNIVERSAL HYPOTHESIS TESTING

BY

DAYU HUANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Adviser:

Professor Sean P. Meyn

Abstract

An important challenge in detection theory is that the size of the state space may be very large. In the context of universal hypothesis testing, two important problems pertaining to the large state space that have not been addressed before are: (1) What is the impact of a large state space on the performance of tests? (2) How does one design an effective test when the state space is large?

This thesis addresses these two problems by developing a generalization of Kullback-Leibler (KL) mismatched divergence, called *mismatched divergence*.

1. We describe a drawback of the Hoeffding test: The *asymptotic* bias and variance of the Hoeffding test are approximately proportional to the size of the state space; thus, it performs poorly when the number of test samples is comparable to the size of state space.

2. We develop a generalization of the Hoeffding test based on the mismatched divergence, called the *mismatched universal test*. We show that this test has asymptotic bias and variance proportional to the dimension of the *function class* used to define the mismatched divergence. The dimension of the function class can be chosen to be much smaller than the size of the state space, and thus our proposed test has a better finite-sample performance in terms of bias and variance.

3. We demonstrate that the mismatched universal test also has an advantage when the distribution of the null hypothesis is learned from data.

4. We develop some algebraic properties and geometric interpretations of the mismatched divergence. We also show its connection to a robust test.

5. We develop a generalization of Pinsker's inequality, which gives a lower bound of the mismatched divergence.

To my parents

Acknowledgments

This thesis would not have been possible without the support of my adviser, Sean Meyn. He has helped and guided me to appreciate the beauty of research. He has always been patient, willing to teach me even basic techniques. He has always made himself available. He encouraged me to ask any questions during our discussion and to pursue answers to them. Many results in this thesis stemmed from those small or maybe naive questions, which would not have been possible without his inspiration. He has helped me in many other aspects that cannot be listed in this short acknowledgment.

I would also like to thank my former adviser, Professor Ada Poon. She has helped me to get started in my research and to develop many basic skills such as explaining myself with a blackboard. She has encouraged me to be brave and try my best in many aspects of my graduate study. These are great assets to me.

I also want to express my gratitude to the professors with whom I have taken courses and discussed research. I especially want to thank Professor Venugopal Veeravalli, Professor Olgica Milenkovic and Professor Todd Coleman for offering continuous support on my research and other aspects of my graduate life. I would also like to thank Professor Rayadurgam Srikant for his encouragement and advice on my study.

I would like to thank my collaborator Jayakrishnan Unnikrishnan for the insightful discussions we had and his patience and help with my writing skills. I would also like to thank many of my colleagues with whom I have had many helpful research discussions. I especially want to thank Wei Dai, Farzad Hassanzadeh, Wei Chen, Kun Deng, Ankur Kulkarni and Hoa Pham. I want to thank my former and current office mates for making life in CSL an

enjoyable time; I learned a lot from them.

I would also like to thank my friends at UIUC, and especially Maokun Li, who has always been a big brother since I first came to the United States. I would like to thank my girlfriend, Fangxue Zheng, for her understanding and support during my difficult times in Urbana.

These acknowledgments would be incomplete without mention of my parents. Their support and love made me the person I am. This thesis is dedicated to them.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Recent decades have seen significant advances in data acquisition and communication technologies. In various areas from finance to science, from personal entertainment to large engineering projects, a lot of data are being collected. Many previously isolated data are now aggregated together and have become more accessible. The result is the availability of huge amounts of data that are still growing.

Consequently, we are now entering into a data-rich era, and we are still at its early stage. While our ability to collect and share data has advanced significantly, our ability to understand and use these data has not kept pace.

For example, high resolution digital cameras are almost everywhere, which has helped create huge image databases. But basic tasks such as automatically recognizing an object based on the image rather than its title are largely unsolved problems [1]. Developing techniques to make sense of image data is likely to lead to novel and promising applications, as suggested by various projects such as MOBVIS [2].

Another example is DNA microarray technology. It has made possible simultaneous profiling of large numbers of genes. Using these and other related data to understand biological system is a difficult challenge, as suggested by the acronym of the Dialogue for Reverse Engineering Assessments and Methods (DREAM) project [3].

One problem that plays an important role in the task of understanding data is the detection problem. In classical detection problems, one is given two or more candidate

hypotheses. The problem is then to decide which hypothesis is true based on data.

The challenges mentioned above suggest a new emphasis: The data is usually of *high dimension*. Or in a probabilistic context, the size of the state space is large. For example, in the face recognition problem, the number of possible values a picture can take is relatively large compared to the number of pictures taken of a particular person. Or in the problem of detecting system abnormality, we use data collected from a large number of sensors.

It is then natural to ask two questions:

1. Does the high dimensionality/large state space matter?

2. If it matters, how can this complexity be addressed?

This thesis provides rigorous answers to these two questions in the particular context of *universal hypothesis testing*. In universal hypothesis testing, we are only given information regarding one of the candidate hypotheses (which we refer as *null hypothesis*). Our task is to design a detector to decide whether this null hypothesis is true or not.

1.2 Previous Work

The study of high dimensionality in classification problems is not new. This topic is a part of the *probably almost correct* (PAC) bound of probability of classification error (see [4] and other related work), and high dimensionality is motivation for the regularization term in classification algorithms. The wonderful survey [5] provides an overview of several important techniques to handle high dimensionality. Another source of references is the textbook [6].

In the context of universal hypothesis testing, the size of the state space could be large. To our best knowledge, the impact of a large state space has not been investigated before. A closely related problem is the asymptotic statistics of the log-likelihood ratio test, studied in [7], [8] and references therein. The main result there is that the asymptotic distribution of log-likelihood ratio is a χ^2 distribution whose degree of freedom is proportional to the size

of the state space. Another related problem is the estimation of Kullback-Leibler divergence when the state space is large or is of infinite dimension (see [9], [10] and references therein).

1.3 Contributions of this Thesis

In this thesis, we propose a generalization of the Hoeffding test [11] called the mismatched universal test. We study the bias and variance of the Hoeffding test and mismatched universal test in the sequential hypothesis testing framework. We also study the bias and variance when the underlying distribution of the known hypothesis has to be *learned*. The mismatched universal test is based on the notion of *mismatched divergence*, a generalization of the Kullback-Leibler divergence. The concept of mismatched divergence was first introduced in [12] and is developed in the research project that leads to this thesis.

The results of this thesis can be summarized as follows:

1. We describe a drawback of the Hoeffding test: Its *asymptotic* bias and variance are approximately proportional to the size of the state space.

2. We develop the mismatched universal test and show that the mismatched universal test has asymptotic bias and variance proportional to the dimension of the *function class* used to define the mismatched divergence. This dimension of the function class can be chosen to be much smaller than the size of the state space, and thus our proposed test has a better performance in terms of bias and variance.

3. We demonstrate that when the distribution of the null hypothesis is learned from data, the estimator of mismatched divergence has smaller bias and variance.

4. We develop some algebraic properties and geometric interpretations of the mismatched divergence. We also show its connection to a robust hypothesis test studied in [13].

5. We develop a generalization of Pinsker's inequality, which gives a lower bound of the mismatched divergence.

Some of the results of this thesis are published in [14] or included in a submitted manuscript [15]. The mismatched divergence is connected to the I-Projection studied in [16]. Some other generalizations of the KL divergence can be found in [16] and [10]. Part of the bias and variance results can also be derived using results from [8], [16] and an unpublished technical report [17].

1.4 Credits

It should be emphasized that many results described in this thesis are the result of joint work with Jayakrishnan Unnikrishnan, Sean Meyn, Venugopal Veeravalli and Amit Surana, and therefore a large percent of credit should go to them.

Chapter 2

KL Divergence and Hoeffding Test

In this chapter, we introduce formally the universal hypothesis testing framework, KL divergence and Hoeffding test. We then describe the bias and variance issue of the Hoeffding test.

2.1 Preliminaries

2.1.1 Sequential hypothesis testing

The sequential hypothesis testing framework is given as follows: Let Z denote the state space. When Z is a finite state space, we assume without loss of generality that $Z = [N]$ where $N = |\mathsf{Z}|$ is the cardinality of Z. Let $\mathcal{P}(\mathsf{Z})$ denote the space of probability distributions on Z. In the simple i.i.d setting of binary hypothesis testing, there are two hypotheses $H0$ and $H1$. Under hypothesis Hi, $i \in \{0, 1\}$, $(Z_1, \ldots, Z_n)$ are assumed to be i.i.d. with distribution $\pi^i \in \mathcal{P}(\mathsf{Z})$. Given the observations $(Z_1, \ldots, Z_n)$, we would like to decide which of these two hypotheses is true.

There is usually a chance that we make a wrong decision and there are two types of errors: false alarm and miss. False alarm refers to the case where we decide in favor of $H1$ when the true underlying hypothesis is $H0$; miss refers to the case where we decide in favor of $H0$ when the true underlying hypothesis is $H1$. It is well known that when n is finite, usually we cannot make both errors arbitrarily small and there is a trade-off between these two types of errors. In the classical Neyman-Pearson setting, we derive a test so that it

minimizes one type of error subject to the constraint that the other type of error is no larger than some threshold. This is a well known subject and its treatment can be found in many textbooks such as [18].

When we allow n to grow into infinity, it is well known that, except in some pathological cases, we can make both errors arbitrarily small [18]. In this context, the rate (error exponent) at which the error decays becomes the object of interest. Analogous to the classical Neyman-Pearson setting, there exists a trade-off between the error exponent of the two types of errors. The goal is to derive a test so that it maximizes one error exponent subject to the constraint that the other error exponent is no smaller than some threshold.

2.1.2 Universal hypothesis testing

In many problems of practical importance, one of the distributions is not known or hard to model. For example, for systems such as the human body, or a secured computer network, only the normal state (healthy person / no intrusion in the network) π^0 is known. Therefore, it is important to derive a test that only requires the knowledge of π^0. This is the universal hypothesis testing problem. In the asymptotic Neyman-Pearson setting, the problem was first studied by Hoeffding [11] for the finite state space case. The main result is that there is a test that does not depend on π^1 and is still universally optimal in the sense that it achieves the optimal error exponent in the asymptotic Neyman-Pearson setting for any π^1. The case when Z is not finite was studied in [19].

We now explain the asymptotic universal hypothesis testing formally. Denote the sequence $(Z_1, \ldots, Z_n)$ by Z_1^n. A decision rule based on Z_1^n is a (probably randomized) binary-valued function $\phi(Z_1^n)$. We decide in favor of H_1 if $\phi(Z_1^n) = 1$, and H_0 otherwise. The two

error exponents are then defined for a test sequence $\phi := \{\phi_1, \phi_2, \dots\}$ as

$$J_\phi^0 := \liminf_{n\to\infty} -\frac{1}{n}\log(\pi^0\{\phi_n(Z_1^n) = 1\}), \tag{2.1}$$

$$J_\phi^1 := \liminf_{n\to\infty} -\frac{1}{n}\log(\pi^1\{\phi_n(Z_1^n) = 0\}). \tag{2.2}$$

The asymptotic Neyman-Pearson criterion of Hoeffding [11] is described as follows: For a given constant bound $\eta \geq 0$ on the false-alarm exponent, an optimal test is the solution to

$$\beta^*(\eta) = \sup\{J_\phi^1 : \ \textit{subject to } J_\phi^0 \geq \eta\}\,, \tag{2.3}$$

where the supremum is over all test sequences $\boldsymbol{\phi}$.

2.1.3 KL divergence and Hoeffding test

The Kullback-Leibler divergence for two probability distributions $\mu^1, \mu^0 \in \mathcal{P}(\mathsf{Z})$ is defined as

$$D(\mu^1\|\mu^0) = \langle \mu^1, \log(d\mu^1/d\mu^0)\rangle, \tag{2.4}$$

where we use the notation $\langle \mu, g\rangle := \mathsf{E}_\mu[g]$. Sometime we also use the notation $\mu(g) := \mathsf{E}_\mu[g]$.

Let $\mathcal{Q}_\alpha(\pi)$ denote the KL divergence neighborhood: $\mathcal{Q}_\alpha(\pi) = \{\mu \in \mathcal{P}(\mathsf{Z}) : D(\mu\|\pi) < \alpha\}$, for $\pi \in \mathcal{P}(\mathsf{Z})$ and $\alpha > 0$.

Define the empirical distributions $\{\Gamma^n : n \geq 1\}$ as elements of $\mathcal{P}(\mathsf{Z})$:

$$\Gamma^n(A) = \frac{1}{n}\sum_{k=1}^n \mathbb{I}\{Z_k \in A\}, \qquad A \in \mathsf{Z}.$$

At time n the Hoeffding test is a test based on the empirical distribution Γ^n. It compares the KL divergence to a threshold δ_n,

$$\phi^{\mathsf{H}}(Z_1^n) = \mathbb{I}\{D(\Gamma^n\|\pi^0) \geq \delta_n\}. \tag{2.5}$$

We remark again that the test (2.5) does not require the knowledge of π^1.

When Z is finite, the test (2.5) with a fix threshold $\delta_n = \eta$ is optimal in the following sense: for any π^1 satisfying $D(\pi^1\|\pi^0) > \eta$,

$$J^0_{\phi^{\mathsf{H}}} \geq \eta, \quad J^1_{\phi^{\mathsf{H}}} \geq \beta^*(\eta).$$

2.2 Bias and Variance Issue of the Hoeffding Test

Note that the the Hoeffding test is given by comparing the test statistic $D(\Gamma^n\|\pi^0)$ with a threshold. It can be easily shown using the strong law of large numbers that $D(\Gamma^n\|\pi^0)$ converges to $D(\pi^i\|\pi^0)$ with probability one. On the other hand, it will be shown shortly that the bias and variance of the test statistics are proportional to the size of the state space divided by the number of samples. Therefore, using the Hoeffding test requires the number of samples to be at least of the same order as $|\mathsf{Z}|$, which makes it impractical when $|\mathsf{Z}|$ is very large. This is summarized in the following result: We use the notation $\mathsf{Var}\,(X)$ to denote the variance of X: $\mathsf{Var}\,(X) = \mathsf{E}[X^2] - \mathsf{E}[X]^2$.

Theorem 2.2.1. *For the model with i.i.d. observations whose marginal distribution is $\pi = \pi^0$, the test statistic sequence $D(\Gamma^n\|\pi^0)$ has the following asymptotic bias and variance when π^0 has full support over Z:*

$$\lim_{n\to\infty} \mathsf{E}[nD(\Gamma^n\|\pi^0)] = \tfrac{1}{2}(N-1), \tag{2.6}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD(\Gamma^n\|\pi^0)] = \tfrac{1}{2}(N-1), \tag{2.7}$$

where $N = |\mathsf{Z}|$ denotes the cardinality of Z. □

We remark that while the bias could be compensated by intentionally introducing a time-varying offset to the threshold η, the variance issue cannot be easily amended.

Chapter 3

Mismatched Divergence and Generalized Pinsker's Inequality

In this chapter, we first introduce the mismatched divergence. We then show that it includes the KL divergence as a special case. We also show that—analogously to Pinsker's inequality, which connects the KL divergence to the total variation distance—the mismatched divergence based on a linear function class admits a generalized Pinsker's inequality.

3.1 Definition of Mismatched Divergence

The KL divergence $D(\mu\|\pi)$ has the following variational representation:

$$D(\mu\|\pi) = \sup_f \big(\mu(f) - \Lambda_\pi(f)\big), \tag{3.1}$$

where $\Lambda_\pi(f) = \log(\pi(e^f))$. The supremum is taken over all f such that $\Lambda_\pi(f) < \infty$ and $\mu(f)$ is well defined. We remark that the supremum is achieved by the log-likelihood ratio: $f = \log(d\mu/d\pi)$.

The mismatched divergence is defined by restricting the supremum in (3.1) to a function class $\mathcal{F}$:

$$D_{\mathcal{F}}^{\mathrm{MM}}(\mu\|\pi) := \sup_{f\in\mathcal{F}} \big(\mu(f) - \Lambda_\pi(f)\big). \tag{3.2}$$

To make sure that this is well defined, we assume that for any $f \in \mathcal{F}$, $\Lambda_\pi(f) < \infty$ and $\mu(f)$ is well defined. Usually we drop the subscript $\mathcal{F}$ when the function class used is clear from the context. The name of the mismatched divergence comes from literature on mismatched decoding [20].

When the supremum on the right-hand side of (3.2) is uniquely achieved, we define the twisted distribution as follows: let f^* denote the function that achieves the supremum, then the twisted distribution $\check{\pi}_\mu$ is defined as the distribution satisfying

$$\check{\pi}_\mu(g) := \frac{\mu(e^{f^*} g)}{\mu(e^{f^*})} \quad \text{for all } g. \tag{3.3}$$

Sometime we omit the subscript μ.

We define the mismatched divergence neighborhood $\mathcal{Q}_\alpha^{\text{MM}}(\pi)$ as

$$\mathcal{Q}_\alpha^{\text{MM}}(\pi) := \{\mu \in \mathcal{P}(\mathsf{Z}) : D^{\text{MM}}(\mu \| \pi) < \alpha\}. \tag{3.4}$$

3.1.1 Linear function class

A special function class is the finite-dimensional linear function class

$$\mathcal{F} = \{\sum_i^d r_i \psi_i : r \in \mathbb{R}^d\}, \tag{3.5}$$

where $\{\psi_i\}$ is the set of *basis* functions. Define the vector valued function $\psi = [\psi_1, \ldots, \psi_d]^T$. We usually write $f_r := \sum_i^d r_i \psi_i$ when the basis functions are clear from context. In this case, the mismatched divergence is defined by a finite-dimensional unconstrained concave problem, which can be efficiently solved using standard optimization solvers to find the global maximum:

$$D^{\text{MM}}(\mu \| \pi) = \sup_{r \in \mathbb{R}^d} \big(\mu(f_r) - \log(\pi(e^{f_r}))\big). \tag{3.6}$$

The twisted distribution has the following representation:

$$\check{\pi}_\mu(g) = \frac{\mu(e^{f_{r^*}} g)}{\mu(e^{f_{r^*}})} \quad \text{for all } g, \tag{3.7}$$

where r^* achieves the supremum in the right-hand side of (3.6).

3.2 Relationship to KL Divergence

Mismatched divergence is a lower bound of the KL divergence:

Lemma 3.2.1. *The following inequality holds for a general function class $\mathcal{F}$:*

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

The equality holds whenever the log-likelihood ratio $\log(d\mu/d\pi) \in \mathcal{F}$. *If* $f \equiv 0 \in \mathcal{F}$, *then* $D^{\text{MM}}(\mu\|\pi) \geq 0$.

Proof. The inequality follows from the variational representation of the mismatched divergence and KL divergence by using the fact that the feasible set of the function in the representation of the KL divergence is no smaller than that of the mismatched divergence. The equality follows from the fact that $\log(d\mu/d\pi)$ achieves the supremum in (3.1). □

We now consider quantizations of the KL divergence. A quantization $\{A_i\}$ is a finite partition of Z:

$$\mathsf{Z} = \cup_{i=1}^{d} A_i$$

where $\{A_i\}$ are disjoint. The quantized probability measures μ^Q and π^Q are defined over the finite state space $\{1, \ldots, d^Q\}$:

$$\mu^Q(i) := \mu(A_i), \pi^Q(i) := \pi(A_i),$$

where d^Q is the level of quantizations and Q stands for quantizations. The KL divergence with quantizations is then defined as

$$D^Q(\mu\|\pi) := D(\mu^Q\|\pi^Q).$$

The KL divergence with quantizations is very useful when one wants to estimate the KL

divergence from empirical distributions, especially when μ and π are continuous probability distributions [9]. We now show that the KL divergence with quantizations is a special case of the mismatched divergence defined using a linear function class in which the functions are indicator functions.

Lemma 3.2.2. *If the function class is taken to be* $\mathcal{F} = \{\sum_i^{d^Q} r_i\psi_i : r \in \mathbb{R}^{d^Q}\}$ *where* $\psi_i = \mathbb{I}_{x\in A_i}$, *then*

$$D^{\text{MM}}(\mu\|\pi) = D^Q(\mu\|\pi). \tag{3.8}$$

Proof. Let $\psi_j^*(i) = \mathbb{I}_{i=j}$. Denote $f_r = \sum_i^{d^Q} r_i\psi_i$ and $\bar{f}_r = \sum_i^{d^Q} r_i\bar{\psi}_i$. It is easy to see that

$$\mu(f_r) = \mu^Q(\bar{f}_r).$$

Since $\{\psi_i\}$ are indicator functions,

$$\pi(e^{f_r}\mathbb{I}_{x\in A_j}) \;\; = \;\; \pi(e^{r_j}\mathbb{I}_{x\in A_j}) = \pi^Q(e^{r_j}\mathbb{I}_{i=j}) = \pi^Q(e^{\bar{f}_r}\mathbb{I}_{i=j}).$$

Consequently,

$$\log(\pi(e^{f_r})) = \log(\pi^Q(e^{\bar{f}_r})).$$

Since the linear function class $\bar{\mathcal{F}} = \{\sum_i r_i\bar{\psi}_i : r \in \mathbb{R}^{d^Q}\}$ contains all the integrable functions, we have

$$D(\mu^Q\|\pi^Q) = \sup_{r\in\mathbb{R}^d}\big(\mu^Q(\bar{f}_r) - \log(\pi^Q(e^{\bar{f}_r}))\big).$$

Combining the above results, we obtain (3.8). □

3.3 Generalized Pinsker's Inequality

Pinsker's inequality [21] provides a lower bound on the KL divergence in terms of the total variation distance,

$$\|\mu - \pi\|_{\text{TV}} := \sup_A |\mu(A) - \pi(A)|.$$

Proposition 3.3.1. *For any two probability measures*

$$D(\mu\|\pi) \geq 2(\|\mu - \pi\|_{TV})^2. \tag{3.9}$$

Our goal in this section is to obtain an equally simple lower bound on $D^{\text{MM}}(\mu\|\pi)$ when the function class is linear. For any function $f\colon \mathsf{Z} \to \mathbb{R}$, the *span norm* is defined by $\|f\|_{\infty,\text{SP}} = (\sup f(x)) - (\inf f(x))$.

Theorem 3.3.2 (Generalized Pinsker's Inequality)**.** *For any two probability measures, the mismatched divergence based on* linear *function class $\mathcal{F}$ admits the following lower-bound:*

$$D^{\text{MM}}_{\mathcal{F}}(\mu\|\pi) \geq 2\sup\left(\frac{\mu(f_r) - \pi(f_r)}{\|f_r\|_{\infty,SP}}\right)^2, \tag{3.10}$$

where the supremum is over all non-zero $r \in \mathbb{R}^d$.

Before proceeding with the proof we remark that Theorem 3.3.2 generalizes Pinsker's inequality. To see this, take $d = 1$ and $r = 1$ and let $\psi_1(x) = \mathbb{I}_A(x)$ for arbitrary $A \in \mathcal{B}(\mathsf{Z})$. In this case we have $\|f_r\|_{\infty,\text{SP}} = r = 1$. Applying Theorem 3.3.2,

$$D(\mu\|\pi) \geq D^{\text{MM}}(\mu\|\pi) \geq 2\sup_A |\mu(A) - \pi(A)|^2.$$

This gives (3.9) since A is arbitrary.

The proof of the theorem is based on reducing the bound on $D^{\text{MM}}(\mu\|\pi)$ to a convex optimization problem that is solved using Hoeffding's inequality [22]. We recall the following consequence of Hoeffding's inequality in Proposition 3.3.3:

Proposition 3.3.3. *For any bounded function $f\colon \mathsf{Z} \to \mathbb{R}$, $\varepsilon > 0$, and any probability distributions $\nu^0, \nu^1 \in \mathcal{P}(\mathsf{Z})$ satisfying $\nu^1(f) - \nu^0(f) \geq \varepsilon$, we have*

$$D(\nu^1 \| \nu^0) \geq 2\frac{\varepsilon^2}{\|f\|^2_{\infty,SP}}.$$

The next key result used in the proof of Theorem 3.3.2 expresses solidarity between the two rate functions.

Lemma 3.3.4. *For all $\epsilon > 0, r \in \mathbb{R}^d$, and $\pi \in \mathcal{P}(\mathsf{Z})$, we have*

$$\inf\big\{D^{\text{MM}}(\mu\|\pi) : \mu(f_r) - \pi(f_r) \geq \epsilon\big\} = \inf\big\{D(\mu\|\pi) : \mu(f_r) - \pi(f_r) \geq \epsilon\big\}.$$

Proof of Theorem 3.3.2. We first prove the lower bound,

$$\inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon} D^{\text{MM}}(\mu\|\pi) \geq \inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon} D(\mu\|\pi).$$

This follows from the expression for $D^{\text{MM}}(\mu\|\pi)$ given in Theorem 7.1.1:

$$\begin{aligned}
\inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon} D^{\text{MM}}(\mu\|\pi) &= \inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon}[\sup_{\alpha} \inf_{\nu:\nu(f_\alpha)\geq\mu(f_\alpha)} D(\nu\|\pi)] \\
&\geq \sup_{\alpha} \inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon}[\inf_{\nu:\nu(f_\alpha)\geq\mu(f_\alpha)} D(\nu\|\pi)] \\
&\geq \inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon}[\inf_{\nu:\nu(f_r)\geq\mu(f_r)} D(\nu\|\pi)] \\
&= \inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon} D(\mu\|\pi).
\end{aligned}$$

Applying the simple bound,

$$D^{\text{MM}}(\mu\|\pi) = \sup_{\alpha} \inf_{\nu:\nu(f_\alpha)\geq\mu(f_\alpha)} D(\nu\|\pi) \leq D(\mu\|\pi),$$

we obtain the reverse inequality

$$\inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon} D^{\text{MM}}(\mu\|\pi) \leq \inf_{\mu:\mu(f_r)\geq\pi(f_r)+\epsilon} D(\mu\|\pi).$$

□

Proof of the Generalized Pinsker's Inequality. For any $\varepsilon > 0$ and any r, by Lemma 3.3.4 and Proposition 3.3.3, we have

$$\inf_{\mu,\pi:\mu(f_r)-\pi(f_r)\geq\epsilon} D^{\text{MM}}(\mu\|\pi) = \inf_{\mu,\pi:\mu(f_r)-\pi(f_r)\geq\epsilon} D(\mu\|\pi) \geq 2\frac{\varepsilon^2}{\|f_r\|^2_{\infty,\text{SP}}}.$$

Therefore, for any μ we can set $\varepsilon = |\mu(f_r) - \pi(f_r)|$ to obtain

$$D^{\text{MM}}(\mu\|\pi) \geq 2\frac{(\mu(f_r) - \pi(f_r))^2}{\|f_r\|^2_{\infty,\text{SP}}}.$$

□

Note that $D(\mu\|\pi) \geq D^{\text{MM}}(\mu\|\pi)$. It is natural to ask whether the generalized Pinsker's inequality provides a better lower bound for $D(\mu\|\pi)$ than (3.9). Unfortunately the answer is no. For the finite state space case we have the following lemma. The proof can be easily generalized to the general state space case.

Lemma 3.3.5. *When the state space is finite, we have*

$$\left(\frac{\mu(f) - \pi(f)}{\|f\|_{\infty,SP}}\right)^2 \leq \sup_A |\mu(A) - \pi(A)|^2.$$

Proof. Note that the left-hand side is invariant when we add a constant function to f or we multiply f by a constant, i.e.,

$$\left(\frac{\mu(f) - \pi(f)}{\|f\|_{\infty,\text{SP}}}\right)^2 = \left(\frac{\mu(\alpha(f+c)) - \pi(\alpha(f+c))}{\|\alpha(f+c)\|_{\infty,\text{SP}}}\right)^2.$$

Consequently, we have the following inequality:

$$\left(\frac{\mu(f)-\pi(f)}{\|f\|_{\infty,\mathrm{SP}}}\right)^2 \leq \sup\{(\mu(f)-\pi(f))^2 : f(z)\in[0,1] \text{ for all } z\}.$$

Observe the maximization problem on the right-hand side. The objective function $(\mu(f)-\pi(f))^2$ is a convex function in f and the constraint set $\{f : f(z)\in[0,1] \text{ for all } z\}$ is a convex set. Thus, there is an optimal solution that is also an extreme point of the constraint set, and any extreme point of the constraint set is an indicator function. Thus, there exists a set A such that

$$(\mu(\mathbb{I}_A)-\pi(\mathbb{I}_A))^2 = \sup\{(\mu(f)-\pi(f))^2 : f(z)\in[0,1] \text{ for all } z\}.$$

□

Chapter 4

Mismatched Universal Test Using a Linear Function Class

In this chapter, we introduce tests using mismatched divergence based on a linear function class. We show that it is asymptotically optimal in a relaxed Neyman-Pearson setting. We then study its bias and variance. Finally, we explain the connection between the mismatched divergence based on a linear function class and graphical models. In this and the following chapters, we restrict ourselves to the finite state space case.

4.1 Mismatched Test Based on a Linear Function Class

Our proposed universal test using mismatched divergence is given as follows:

$$\phi^{\text{MM}}(\boldsymbol{Z}) = \mathbb{I}\{D^{\text{MM}}(\Gamma^n \| \pi^0) \geq \delta_n\}, \tag{4.1}$$

where Γ^n is the empirical distribution. We call this test the *mismatched universal test*. In this chapter, we restrict ourselves to the case when the function class is linear.

The result in this chapter was developed jointly with Jayakrishnan Unnikrishnan, Sean Meyn and Venugopal Veeravalli.

4.1.1 Optimality

From the relationship between the mismatched divergence and KL divergence, it is clear that when the set $\{\psi_i\}$ spans all the functions, the test is optimal in the Neyman-Pearson sense defined in the text around Equation (2.3). This is not an isolated case: The mismatched universal test is optimal in a relaxed asymptotic Neyman-Pearson setting in which we restrict the set of possible tests, as described in the following proposition:

Proposition 4.1.1. *Suppose π^1 and π^0 satisfy $D^{\text{MM}}(\pi^1\|\pi^0) + D^{\text{MM}}(\pi^0\|\pi^1) < \infty$. When the observations $\boldsymbol{Z} = \{Z_t : t = 1, \ldots\}$ are i.i.d., then the universal test defined in* (4.1) *achieves the optimal error rate in the relaxed Neyman-Pearson setting*

$$\beta_\eta^{\text{MM}*} := \sup\{J_\phi^1 : \ \textit{subject to } J_\phi^0 \geq \eta, \phi \in \Phi\},$$

where Φ is the set of tests of the following form:

$$\Phi = \{\mathbb{I}\{\Gamma^n(f) \geq \tau\} : f \in \mathcal{F}\}.$$

This proposition is essentially [14, Proposition 3.1]. The main idea is to look at the geometric picture for the relaxed Neyman-Pearson setting. We will not give the proof here since it is not the major theme of our thesis and the proof is lengthy.

4.2 Bias and Variance

In this section, we will study the bias and variance $D^{\text{MM}}(\Gamma^n\|\pi^0)$ when the sequence of observations $\boldsymbol{Z} = \{Z_t : t = 1, \ldots\}$ are i.i.d. with marginal π. We will first consider the case when the null hypothesis is true, and then extend it to the case when the alternate hypothesis is true.

4.2.1 Bias and variance when the null hypothesis is true

When $\pi = \pi^0$, it is easy to see that $D^{\text{MM}}(\pi\|\pi^0) = 0$. Note that $D^{\text{MM}}(\mu\|\pi^0)$ is approximately quadratic when $\mu \approx \pi^0$, and the difference between Γ^n and π^0 is on the order of $1/\sqrt{n}$ by the central limit theorem. Thus it is not surprising that the bias is on the order of $1/n$. What is interesting is that the asymptotic bias and variance have a very simple expression that depends on the dimension of the function class. Defining Σ_{π^0} as

$$\Sigma_{\pi^0} = \pi^0(\psi\psi^T) - \pi^0(\psi)\pi^0(\psi^T),$$

we have the following theorem:

Theorem 4.2.1. *Suppose $\boldsymbol{Z}$ is drawn i.i.d. from a finite set Z with marginal π^0 and assume Σ_{π^0} is positive definite. Then the universal statistic has bias of order n^{-1} and variance of order n^{-2}, and the normalized asymptotic values have simple, explicit forms:*

$$\lim_{n\to\infty} n\mathsf{E}[D^{\text{MM}}(\Gamma^n\pi^0)] = \tfrac{1}{2}d, \tag{4.2}$$

$$\lim_{n\to\infty} n^2\mathsf{Var}\,[D^{\text{MM}}(\Gamma^n\pi^0)] = \tfrac{1}{2}d. \tag{4.3}$$

The assumption on Σ_{π^0} basically says that $\{\psi_i\}$ is *minimal*. That is, no nontrivial linear combination of $\{\psi_i\}$ is a constant function almost surely with respect to π^0. This assumption is not restrictive since any set of basis functions can be reduced to a set of *minimal* basis functions though d will change.

Proposition 4.2.1 suggests the following:

1. The asymptotic bias and variance of the mismatched universal test can be much smaller than the Hoeffding test. Moreover, they can be controlled by properly selecting the function class and thus provide a solution when $|\mathsf{Z}|$ is large and the number of samples n is limited.

2. When the log-likelihood ratio $\log(d\pi/d\pi^0)$ is in the function class, we have

$$D^{\text{MM}}(\pi\|\pi^0) = D(\pi\|\pi^0).$$

Thus the test based on the mismatched divergence is asymptotically optimal in the usual asymptotic Neyman-Pearson setting.

3. The bias term suggests that instead of fixing δ_n to be η, a time-varying $\delta_n = \eta + \frac{d}{2n}$ may perform better in practice when the number of samples n is finite.

The proof of Theorem 4.2.1 basically follows from analyzing the Taylor series expansion of $D^{\text{MM}}(\Gamma^n\|\pi^0)$. There is an issue of proving the convergence of mean and variance from convergence in distribution. This issue is addressed by the following lemma proved in [15]. Let $\mathsf{Cov}\,(X)$ denote the covariance matrix of vector X as: $\mathsf{Cov}\,(X) = \mathsf{E}[(X - \mathsf{E}[X])(X - \mathsf{E}[X])^{\intercal}]$.

Lemma 4.2.2. *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. $\lim_{n\to\infty} -\frac{1}{n}\log \mathsf{P}\{S^n \notin K\} > 0.$

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*

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

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

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

□

Proof of Theorem 4.2.1. To apply Lemma 4.2.2, h is specialized to be $h(\mu) := D^{\text{MM}}(\mu\|\pi^0)$. We take $X^i = (\mathbb{I}_{z_1}(Z_i), \mathbb{I}_{z_2}(Z_i), \ldots, \mathbb{I}_{z_N}(Z_i))^T$, and $\mathsf{Z} = [0,1]^N$. Take $\Xi = \mathsf{Cov}\left(X\right)$. Define the matrix the Ψ as $\Psi_{i,j} = \psi_i(j)$. It is easy to see that $\Sigma_{\pi^0} = \Psi\Xi\Psi^T$.

We demonstrate that

$$M = \nabla^2 h(\pi^0) = \Psi^T(\Sigma_{\pi^0})^{-1}\Psi, \tag{4.4}$$

and prove that the other technical conditions of Lemma 4.2.2 are satisfied. The rest will follow from Lemma 4.2.2, as

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

and similarly

$$\mathsf{tr}\left(M\Xi M\Xi\right) = \mathsf{tr}\left(I_d\right) = d.$$

The condition Σ_{π^0} being positive definite indicates that the objective function of the right-hand side of (3.6) is strictly concave and thus has a unique maximum for each μ. Let $r(\mu)$ be the maximizer for a given μ. Then

$$h(\mu) = \mu(f_{r(\mu)}) - \Lambda_{\pi^0}(f_{r(\mu)}).$$

Recall that $\check{\pi}_\mu$ is the twisted distribution defined in (3.3). Define $\check{\Sigma}_\mu$ as

$$\check{\Sigma}_\mu = \check{\pi}_\mu(\psi\psi^T) - \check{\pi}_\mu(\psi)\check{\pi}_\mu(\psi^T).$$

The first order optimality condition in the right-hand side of (3.6) gives

$$\mu(\psi) - \check{\pi}_\mu(\psi) = 0.$$

On taking the derivative with respect to μ_z with $z \in \mathsf{Z}$, we have

$$\psi(z) - \check{\Sigma}_\mu \frac{\partial r(\mu)}{\partial \mu(z)} = 0.$$

Then it is straightforward to show that

$$\frac{\partial}{\partial \mu(z)} h(\mu) = f_{r(\mu)}(z),$$

$$\frac{\partial^2}{\partial \mu(z) \partial \mu(\bar{z})} h(\mu) = \psi^{\scriptscriptstyle T}(z) \frac{\partial r(\mu)}{\partial \mu(\bar{z})} = \psi^{\scriptscriptstyle T}(z) \check{\Sigma}_\mu^{-1} \psi(\bar{z}).$$

When $\mu = \pi^0$, we have $r(\pi^0) = 0$ and $\check{\Sigma}_\mu = \Sigma_{\pi^0}$. Thus,

$$\frac{\partial^2}{\partial \mu(z) \partial \mu(\bar{z})} h(\pi^0) = \psi^{\scriptscriptstyle T}(z) \Sigma_{\pi^0}^{-1} \psi(\bar{z}).$$

We now verify the remaining conditions required in Lemma 4.2.2:

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

2. 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.

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

4. Pick a compact set K that contains π^0 as an interior point, and

$$K \subset \{\mu \in \mathcal{P}(\mathsf{Z}) : \max_u |\mu(u) - \pi^0(u)| < \tfrac{1}{2} \min_u |\pi^0(u)|\}.$$

This choice of K ensures that $\lim_{n\to\infty} -\frac{1}{n}\log \mathsf{P}\{S^n \notin K\} > 0$. Since $r(\mu)$ is continuously differentiable on K, we conclude that h is C^2 on K.

□

4.2.2 Bias and variance when the alternate hypothesis is true

There is more than one way to derive the result for the asymptotic bias and variance of $D^{\text{MM}}(\Gamma^n\|\pi^0)$ when $\pi \neq \pi^0$. Here we show that the case when $\pi \neq \pi^0$ can be derived from the case when $\pi = \pi^0$ using the following lemma:

Lemma 4.2.3. *Suppose the supremum in $D^{\text{MM}}(\mu\|\pi^0)$ and $D^{\text{MM}}(\pi^1\|\pi^0)$ are both achieved. Denote $\check{\pi} = \check{\pi}_{\pi^1}$. We have*

$$D^{\text{MM}}(\mu\|\pi^0) = D^{\text{MM}}(\mu\|\check{\pi}) + D^{\text{MM}}(\pi^1\|\pi^0) + \langle \mu - \pi^1, \log(\check{\pi}/\pi^0)\rangle.$$

Here we use the theory of I-projection [16] to derive Lemma 4.2.3. Let $\mathcal{L}$ denote the linear family of probability distributions:

$$\mathcal{L}(\mu) = \{\nu : \nu(\psi_i) = \mu(\psi_i), \text{ for all } i\}$$

Let $\mathcal{E}$ denote the exponential family of probability distributions:

$$\mathcal{E} = \{\nu : \nu(z) = \frac{\pi^0(z)e^{f_r(z)}}{\pi^0(e^{f_r})}, r \in \mathbb{R}^d\}.$$

Then by the theory of I-projection, $\check{\pi}_\mu$ is the unique intersection of the linear family and exponential family [16]:

$$\mathcal{L}(\mu) \cap \mathcal{E} = \{\check{\pi}_\mu\},$$

and

$$D^{\text{MM}}(\mu\|\pi^0) = D(\check{\pi}_\mu\|\pi^0).$$

Proof of Lemma 4.2.3. From the definition of the exponential family, $\check{\pi}$, $\check{\pi}_\mu$ and π^0 belong to the same exponential family. Therefore, $D^{\text{MM}}(\mu\|\pi^0) = D(\check{\pi}_\mu\|\pi^0)$, $D^{\text{MM}}(\mu\|\check{\pi}) = D(\check{\pi}_\mu\|\check{\pi})$ and $D^{\text{MM}}(\pi^1\|\pi^0) = D(\check{\pi}\|\pi^0)$. Consequently,

$$\begin{aligned}
D^{\text{MM}}(\mu\|\pi^0) &= D(\check{\pi}_\mu\|\pi^0) = D(\check{\pi}_\mu\|\check{\pi}) + \langle\check{\pi}_\mu, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&= D(\check{\pi}_\mu\|\check{\pi}) + \langle\mu, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&= D^{\text{MM}}(\mu\|\check{\pi}) + \langle\mu, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&= D^{\text{MM}}(\mu\|\check{\pi}) + \langle\check{\pi}, \log(\frac{\check{\pi}}{\pi^0})\rangle + \langle\mu - \check{\pi}, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&= D^{\text{MM}}(\mu\|\check{\pi}) + D(\check{\pi}\|\pi^0) + \langle\mu - \check{\pi}, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&= D^{\text{MM}}(\mu\|\check{\pi}) + D^{\text{MM}}(\pi^1\|\pi^0) + \langle\mu - \check{\pi}, \log(\frac{\check{\pi}}{\pi^0})\rangle \\
&= D^{\text{MM}}(\mu\|\check{\pi}) + D^{\text{MM}}(\pi^1\|\pi^0) + \langle\mu - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle,
\end{aligned}$$

where the third equality $\langle\mu, \log(\frac{\check{\pi}}{\pi^0})\rangle = \langle\check{\pi}_\mu, \log(\frac{\check{\pi}}{\pi^0})\rangle$ follows from the fact that $\log(\frac{\check{\pi}}{\pi^0}) \in \mathcal{F}$ and $\check{\pi}_\mu \in \mathcal{L}(\mu)$; and the last equality $\langle\check{\pi}, \log(\frac{\check{\pi}}{\pi^0})\rangle = \langle\pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle$ follows from a similar reasoning. □

Applying Lemma 4.2.3, we obtain:

$$D^{\text{MM}}(\Gamma^n\|\pi^0) = D^{\text{MM}}(\Gamma^n\|\check{\pi}) + D^{\text{MM}}(\pi^1\|\pi^0) + \langle\Gamma^n - \check{\pi}, \log(\check{\pi}/\pi^0)\rangle. \tag{4.5}$$

The decomposition suggests that the bias and the variance of $D^{\text{MM}}(\Gamma^n\|\pi^0)$ come from the first and third term. The first term $D^{\text{MM}}(\Gamma^n\|\check{\pi})$ can be studied using an argument similar to that of the case $\pi = \pi^0$ and shown to have asymptotic bias of order n^{-1} and variance of order n^{-2}. The third term has a mean 0 and the central limit theorem applying to Γ^n suggests that the variance of the second term is of order n^{-1}. This observation suggests the following statement and an approach to prove it: The bias of the overall term is mainly contributed by the first term and is of order n^{-1}; when n is large, the variance of the overall

term is mainly contributed by the third term and is of order n^{-1}. In the rest of this section we will make this precise. We remark that when n is small, the variance of the first term could be very significant.

Theorem 4.2.4. *Suppose $\boldsymbol{Z}$ is drawn i.i.d. from a finite set Z with marginal $\pi = \pi^1 \neq \pi^0$. Assume Σ_{π^0} is positive definite and $\pi \preceq \pi^0$. Then the universal statistic has bias of order n^{-1} and variance of order n^{-1}, and given by the explicit forms:*

$$\lim_{n\to\infty} n\mathsf{E}[D^{\text{MM}}(\Gamma^n\|\pi^0) - D^{\text{MM}}(\pi\|\pi^0)] = \tfrac{1}{2}\mathsf{tr}\,(\check{\Sigma}_\pi^{-1}\Sigma_{\pi^0}), \tag{4.6}$$

$$\lim_{n\to\infty} n\mathsf{Var}\,[D^{\text{MM}}(\Gamma^n\|\pi^0)] = \mathsf{Cov}\,(\log(\check{\pi}/\pi^0)), \tag{4.7}$$

where $\check{\pi} = \check{\pi}_{\pi^1}$ and

$$\check{\Sigma}_\pi = \check{\pi}(\psi\psi^T) - \check{\pi}(\psi)\check{\pi}(\psi^T).$$

When $\log(\pi^1/\pi^0) \in \mathcal{F}$, the bias has a simple form:

$$lim_{n\to\infty} n\mathsf{E}[D^{\text{MM}}(\Gamma^n\|\pi^0) - D^{\text{MM}}(\pi\|\pi^0)] = \tfrac{1}{2}d. \tag{4.8}$$

To prove Theorem 4.2.4 we first investigate the bias and variance of the two terms in (4.5), as summarized in the following two lemmas:

Lemma 4.2.5. *Under the assumptions of Proposition 4.2.4, we have*

$$\lim_{n\to\infty} n\mathsf{E}[D^{\text{MM}}(\Gamma^n\|\check{\pi})] = \tfrac{1}{2}\mathsf{tr}\,(\check{\Sigma}_\pi^{-1}\Sigma_{\pi^0}), \tag{4.9}$$

$$\lim_{n\to\infty} n^2\mathsf{Var}\,[D^{\text{MM}}(\Gamma^n\|\check{\pi})] = \tfrac{1}{2}\mathsf{tr}\,(\check{\Sigma}_\pi^{-1}\Sigma_{\pi^0}\check{\Sigma}_\pi^{-1}\Sigma_{\pi^0}). \tag{4.10}$$

Proof. The proof is similar to that of Proposition 4.2.1 and is left to the Appendix. □

Lemma 4.2.6. *Under the assumptions of Proposition 4.2.4, we have*

$$E[\langle \Gamma^n - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle] = 0, \tag{4.11}$$

$$n\mathsf{Var}\left[\langle \Gamma^n - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle\right] = \mathsf{Cov}\left(\log(\check{\pi}/\pi^0)\right). \tag{4.12}$$

Proof. The proof is trivial. □

Proof of Theorem 4.2.4. Using Lemma 4.2.7 we obtain that supremum in the definition of the mismatched divergence is achieved and the conditions of Lemma 4.2.3 are satisfied. Combining (4.5), (4.9) and (4.11), we obtain (4.6). When $\log(\pi^1/\pi^0) \in \mathcal{F}$, we have $\check{\pi} = \pi^1$ and obtain (4.8). We now compute the variance. To prove (4.7) we first use the short-hand notations

$$X^n = D^{\text{MM}}(\Gamma^n \| \check{\pi}), \quad Y^n = \langle \Gamma^n - \pi^1, \log(\frac{\check{\pi}}{\pi^0})\rangle.$$

The variance is then expressed as

$$n\mathsf{Var}\left(X^n + Y^n\right) = n\mathsf{Var}\left(X^n\right) + n\mathsf{Var}\left(Y^n\right) + 2n\mathsf{E}\big[(X^n - \mathsf{E}[X^n])(Y^n - \mathsf{E}[Y^n])\big],$$

where the last term can be bounded using the Cauchy-Schwarz inequality:

$$n|\mathsf{E}\big[(X^n - \mathsf{E}[X^n])(Y^n - \mathsf{E}[Y^n])| \le \sqrt{n\mathsf{Var}\left(X^n\right) n\mathsf{Var}\left(Y^n\right)}.$$

Since (4.10) and (4.12) imply,

$$\lim_{n\to\infty} n\mathsf{Var}\left(X^n\right) = 0, \quad , \lim_{n\to\infty} n\mathsf{Var}\left(Y^n\right) \le \infty,$$

we have

$$\lim_{n\to\infty} n\mathsf{E}\big[(X^n - \mathsf{E}[X^n])(Y^n - \mathsf{E}[Y^n]) = 0.$$

Therefore,

$$\lim_{n\to\infty} n\mathsf{Var}\,(X^n + Y^n) = \lim_{n\to\infty} n\mathsf{Var}\,(Y^n).$$

Substituting the right-hand side using (4.12), we obtain (4.7). □

Lemma 4.2.7. *If $\mu \preceq \pi$ for all i and Σ_π is positive definite, then the right-hand side of the definition of mismatched divergence based on linear function class* (3.6) *has a unique maximizer.*

The proof is given in the appendix.

Specializing Theorem 4.2.4 to the KL divergence with quantizations, we have

Corollary 4.2.8. *Suppose $\boldsymbol{Z}$ is drawn i.i.d. from a finite set Z with marginal $\pi = \pi^1 \neq \pi^0$ and $\pi^1 \preceq \pi^0$. Then the quantized divergence $D^Q(\Gamma^n\|\pi^0)$ has bias of order n^{-1} and variance of order n^{-1}, and the normalized asymptotic values have explicit forms:*

$$\lim_{n\to\infty} n\mathsf{E}[D^Q(\Gamma^n\|\pi^0) - D^Q(\mu\|\pi^0)] = \tfrac{1}{2}(d^Q - 1),$$
$$\lim_{n\to\infty} n\mathsf{Var}\,[D^Q(\Gamma^n\|\pi^0)] = \mathsf{Var}_{\,\mu^Q}\big(\log(\frac{d\mu^Q}{d{\pi^0}^Q})\big).$$

4.3 Application to Graphical Models

Graphical models can be used to model interactions between random variables and are useful in many applications. The reference [23] is an excellent tutorial. Here we only consider a special case. Let $\{X_i, i = 1, \ldots, K\}$ be a set of random variables taking values in a finite set which we assume without loss of generality to be $[M] = \{1, 2, \ldots, M\}$. We assume that their joint distribution is modeled using an exponential family:

$$\Pr\{X_1 = x_1, \ldots X_M = x_N\} = C\exp\Big\{\sum_{i,a} \lambda_{i,a}\mathbb{I}\{x_i = a\} + \sum_{i,j,a,b} \theta_{i,j,a,b}\mathbb{I}\{x_i = a\}\mathbb{I}\{x_j = b\}\Big\}, \tag{4.13}$$

where C is a normalizing constant. Thus, a distribution π is specified by the set of weights $\{\theta_{i,j,a,b}\}$ and $\{\lambda_{i,a}\}$. The distribution is associated with an undirected graph $G = (V, E)$. For each i, X_i is associated with a vertex v_i in V. Each edge $e_{i,j}$ is associated with the set of weights $\{\theta_{i,j,a,b}, a \in [M], b \in [M]\}$. There is no edge between v_i, v_j if and only if all the weights in $\{\theta_{i,j,a,b}, a \in [M], b \in [M]\}$ are zero.

Consider a universal hypothesis testing problem where in the null hypothesis $\{X_i\}$ has distribution π^0 and weights $\Theta^0 = \{\theta^0_{i,j,a,b}\}$, $\Lambda^0 = \{\lambda^0_{i,a}\}$; and in the alternate hypothesis $\{X_i\}$ has distribution π^1 and weights $\Theta^1 = \{\theta^1_{i,j,a,b}\}$, $\Lambda^1 = \{\lambda^1_{i,a}\}$. Only the weights Θ^0 and Λ^0 are known, and the graphical structure (namely the edges) are known. Our task is to decide whether the set of weights is Θ^0, or not.

From the theory above, we know that the Hoeffding test has asymptotic bias and variance given by the following corollary.

Corollary 4.3.1. *Suppose the null hypothesis is true. The universal hypothesis testing statistics $D(\Gamma^n \| \pi^0)$ have asymptotic bias and variance given by*

$$\begin{aligned} \lim_{n\to\infty} \mathsf{E}[nD(\Gamma^n \| \pi^0)] &= \tfrac{1}{2}(M^{|V|} - 1), \\ \lim_{n\to\infty} \mathsf{Var}\,[nD(\Gamma^n \| \pi^0)] &= \tfrac{1}{2}(M^{|V|} - 1). \end{aligned}$$

Note here that the definition of KL divergence is extended to the multi-dimensional distribution using the variational representation in (3.1).

We may also use the prior knowledge of the graph structure, and using the mismatched universal test with the following function class:

$$\mathcal{F} = \Big\{ \sum_{i,a} \lambda_{i,a} \mathbb{I}\{x_i = a\} + \sum_{i,j,a,b} \theta_{i,j,a,b} \mathbb{I}\{x_i = a\} \mathbb{I}\{x_j = b\}, \text{for all } \Lambda,\, \Theta \text{ consistent with } G \Big\}. \tag{4.14}$$

From the connection between exponential families and mismatched divergence, for any two distributions π^1, π^0 consistent with the graphical model, the log-likelihood ratio $\log(\pi^1/\pi^0)$

is in the function class $\mathcal{F}$. Therefore we have the following corollary.

Corollary 4.3.2. *Suppose the null hypothesis is true. The mismatched universal test using $\mathcal{F}$ given by* (4.14) *is optimal in the asymptotic Neyman-Pearson setting. The test statistics $D^{\text{MM}}(\Gamma^n\|\pi^0)$ have asymptotic bias and variance given by*

$$\begin{aligned}\lim_{n\to\infty} \mathsf{E}[nD(\Gamma^n\|\pi^0)] &= \tfrac{1}{2}\big((M-1)^2|E| + (M-1)|V|\big),\\ \lim_{n\to\infty} \mathsf{Var}\,[nD(\Gamma^n\|\pi^0)] &= \tfrac{1}{2}\big((M-1)^2|E| + (M-1)|V|\big).\end{aligned}$$

The reduction in bias and variance is due to the restriction of pairwise interactions and the edge structure. The reduction is more significant when the graph is sparse: For a complete graph, the bias and variance are proportional to $((M-1)|V|(|V|-1)/2+|V|)$. For a graph that is a tree structure, the bias and variance are proportional to $((M-1)(|V|-1)+|V|)$.

For example, consider a model with binary outputs, $x_i \in \{0,1\}$ for each $1 \le i \le 10$. That is, $M = 2$ and $|V| = 10$. For the Hoeffding test, the variance is given by $\frac{1}{2}(2^{10}-1)$. If the graph is complete and we use the mismatched universal test, the variance is given by $\frac{55}{2}$. If the graph is a tree and we use mismatched universal test, the variance is given by $\frac{19}{2}$.

Chapter 5

Mismatched Universal Test Using a General Function Class

In this chapter, we extend the theory in Chapter 4 to more general cases, in which we allow the function class to be nonlinear. We also show that the general mismatched universal test includes a robust hypothesis test studied in [13].

5.1 Bias and Variance of a Robust Test

In this section, we study the bias and variance of the robust test studied in [13] by exploring its connection to the mismatched universal test. In the robust hypothesis testing framework, the null hypothesis π^0 is only known to belong to a moment class $\mathbb{P}$ defined by a set of functions $\{\psi_i, i = 1, \ldots, d\}$:

$$\mathbb{P} = \{\varpi : \varpi(\psi_i) = c_i, i = 1, \ldots, d\}$$

The robust test is given by

$$\Phi = \mathbb{I}\{\inf_{\varpi \in \mathbb{P}} D(\mu \| \varpi) \geq \tau\}.$$

Loosely speaking, $\inf_{\varpi \in \mathbb{P}} D(\Gamma^n \| \varpi)$ measures the worst-case error exponent when the true distribution belongs to $\mathbb{P}$.

We assume the following regularity condition in this section, which also appears in [13].

The result in this chapter was developed jointly with Jayakrishnan Unnikrishnan, Sean Meyn and Venugopal Veeravalli.

Assumption 5.1.1. *Assume that $\psi_1, \ldots, \psi_d$ are continuous over Z, and c lies in the interior points of the set of feasible moment vectors, defined as*

$$\Delta := \{\pi(\psi), \text{ for some } \pi \in \mathsf{P}(\mathsf{Z})\}.$$

Define

$$R(\psi, r_0) := \{r \in \mathbb{R}^d : r_0 + r^{\scriptscriptstyle T}\psi(x) > 0 \text{ for all } x \in \mathsf{Z}\}.$$

The following lemma follows from Theorem 1.4 and the statement in Section 3.3 in [13]:

Lemma 5.1.2. *[13] Suppose Assumption 5.1.1 holds, then*

$$\inf_{\varpi \in \mathbb{P}} D(\mu \| \varpi) = \sup\{\mu(\log(r_0 + r^{\scriptscriptstyle T}\psi)) : r_0 + r^{\scriptscriptstyle T}c = 1, r \in R(\psi, r_0)\}, \tag{5.1}$$

and there exists an optimizer satisfying $r_0 \neq 0$.

The conclusion that the optimizer satisfies $r_0 \neq 0$ is indicated by the proof of Theorem 1.4 in [13] since r_0 is a Lagrangian multiplier for the constraint $\pi(1) = 1$ in $\inf_{\varpi \in \mathbb{P}} D(\mu \| \varpi)$, and it always has nonzero sensitivity.

Using Lemma 5.1.2, we can show that the robust test is a special mismatched universal test:

Theorem 5.1.3. *Suppose that Z is compact, the functions $\{\psi_i\}$ are continuous. The function class $\mathcal{F}$ is defined as*

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

The robust test statistics have the following alternative representation:

$$\inf_{\varpi \in \mathbb{P}} D(\mu \| \varpi) = D^{\text{MM}}(\mu \| \pi^0),$$

where π^0 is any distribution satisfying $\pi^0 \in \mathbb{P}$.

Proof. For each $\mu \in \mathcal{M}_1$, by applying Lemma 5.1.2, we obtain

$$\begin{aligned}
\inf_{\varpi \in \mathbb{P}} D(\mu \| \varpi) &= \sup_{(r_0, r): r \in R(\psi, r_0), r_0 + r^T c = 1} \{\mu(\log(r_0 + r^T \psi))\} \\
&= \sup_{(r_0, r): r \in R(\psi, r_0), r_0 + r^T c = 1} \{\mu(\log(r_0 + r^T \psi)) - \log(\pi^0(r_0 + r^T \psi))\} \\
&= \sup_{(r_0, r): r \in R(\psi, r_0)} \{\mu(\log(r_0 + r^T \psi)) - \log(\pi^0(r_0 + r^T \psi))\} \\
&= \sup_{(r_0, r'): r' r_0 \in R(\psi, r_0)} \{\mu(\log(r_0(1 + r'^T \psi))) - \log(\pi^0(r_0(1 + r'^T \psi))\} \\
&= \sup_{r' \in R(\psi, 1)} \{\mu(\log(1 + r'^T \psi)) - \log(\pi^0((1 + r'^T \psi))\} \\
&= \sup_{r' \in R(\psi, 1)} \{\mu(\log(1 + r'^T \psi)) - \Lambda_{\pi^0}\big(\log(\pi^0(1 + r'^T \psi))\big)\} \\
&= \sup_{f \in \mathcal{F}} \{\mu(f) - \Lambda_{\pi^0}(f)\} \\
&= D^{\mathrm{MM}}(\mu \| \pi).
\end{aligned}$$

The second equality follows because when $\pi^0(r_0 + r^T \psi) = 1$, we have $\log(\pi^0(r_0 + r^T \psi)) = 0$. The fifth equality follows from the fact that $\mu(\log(r^T \psi)) - \Lambda_\pi(\log(r^T \psi))$ is invariant when r is multiplied by a positive real number. □

Based on this connection, we have the following result on the bias and variance of the robust test, which is a special case of a more general result that we will prove later.

Proposition 5.1.4. *Suppose $\boldsymbol{Z}$ is drawn i.i.d. from a finite set Z with marginal $\pi^0 \in \mathbb{P}$ and Assumption 5.1.1 holds. Assume $\Sigma_{\pi^0} := \pi^0(\psi\psi^T) - \pi^0(\psi)\pi^0(\psi^T)$ is positive definite. Then the universal statistic has bias of order n^{-1} and variance of order n^{-2}, and the normalized asymptotic values have simple, explicit forms:*

$$\begin{aligned}
\lim_{n \to \infty} n \mathsf{E}[\inf_{\varpi \in \mathbb{P}} D(\Gamma^n \| \varpi)] &= \tfrac{1}{2} d, \\
\lim_{n \to \infty} n^2 \mathsf{Var}[\inf_{\varpi \in \mathbb{P}} D(\Gamma^n \| \varpi)] &= \tfrac{1}{2} d.
\end{aligned}$$

The proof will be given after we derive the more general case.

5.2 Bias and Variance for Mismatched Universal Tests using a General Function Class

In this section, we consider a general case: The function class is only assumed to have a d-dimensional parameterization r:

$$\mathcal{F} = \{f_r, r \in \mathbb{R}^d\}.$$

While in the linear case the test is optimal in the relaxed asymptotic Neyman-Pearson setting as claimed in Proposition 4.1.1, the optimality is not necessarily true in the general case. The problem is that $\mathcal{F}$ is not necessarily a pointed cone and the condition of Theorem 7.1.1 fails. On the other hand, the bias and variance result can be generalized to the general case, as given by the following theorem:

Theorem 5.2.1. *Suppose that the observation sequence $\boldsymbol{Z}$ is i.i.d. with marginal π. Suppose that there exists r^* satisfying $f_{r^*} = \log(\pi/\pi^0)$ and $f_r(z)$ is C^2 in r in an open neighborhood B_1 of r^* for every $z \in \mathsf{Z}$. Further, suppose that*

1. *There is an open neighborhood B of π, such that for any $\mu \in B$, the supremum in the definition of $D^{\text{MM}}(\mu\|\pi^0)$ is uniquely achieved.*
2. *The matrix $\Sigma_\pi := \pi(\nabla f_r(\nabla f_r)^T) - \pi(\nabla f_r)\pi^0((\nabla f_r)^T)\big|_{r=r^*}$ is positive definite.*

Then,

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

$$\lim_{n\to\infty} \mathsf{E}[nD^{\text{MM}}(\Gamma^n\|\pi^0)] = \tfrac{1}{2}d, \tag{5.2}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[nD^{\text{MM}}(\Gamma^n\|\pi^0)] = \tfrac{1}{2}d. \tag{5.3}$$

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

$$\lim_{n\to\infty} \mathsf{E}[n(D^{\text{MM}}(\Gamma^n\|\pi^0) - D(\pi^1\|\pi^0))] = \tfrac{1}{2}d, \tag{5.4}$$

$$\lim_{n\to\infty} \mathsf{Var}\,[n^{\frac{1}{2}} D^{\text{MM}}(\Gamma^n\|\pi^0)] = \sigma_1^2. \tag{5.5}$$

Similarly to the case of the linear function class, the second result where $\pi = \pi^1 \neq \pi^0$ can be derived from the first result using the following lemma which generalizes Lemma 4.2.3:

Lemma 5.2.2. *Suppose the supremum in $D^{\text{MM}}(\mu\|\pi^0)$ and $D^{\text{MM}}(\pi^1\|\pi^0)$ are both achieved. Denote $\check{\pi} = \check{\pi}_{\pi^1}$ which is defined in* (3.3)*. Define $\mathcal{G} = \mathcal{F} - f_{r^*} := \{f_r - f_{r^*} : r \in \mathbb{R}^d\}$. Then we have*

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

To prove this lemma, we need the following equality from [15, Proposition II.3] which holds when the supremum in the definition of the mismatched divergence is achieved:

$$D^{\text{MM}}(\mu\|\pi^0) = D(\mu\|\pi^0) - \inf_{\nu\in\mathcal{E}_\pi} D(\mu\|\nu) = D(\mu\|\pi^0) - D(\mu\|\check{\pi}_\mu). \tag{5.6}$$

Proof. In the following identities, the first, third and fifth equalities follow from (5.6).

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

□

Proof of Theorem 5.2.1. (1) We first prove the result for the case $\pi = \pi^0$, and the argument

is similar to that of Theorem 4.2.1. To apply Lemma 4.2.2, h is specialized to be $h(\mu) := D^{\text{MM}}(\mu\|\pi^0)$. Take $X^i = (\mathbb{I}_{z_1}(Z_i), \mathbb{I}_{z_2}(Z_i), \ldots, \mathbb{I}_{z_N}(Z_i))^T$. Let $\mathsf{X} = [0,1]^N$ and $\Xi = \mathsf{Cov}\,(X)$. Redefine the matrix Ψ as $\Psi_{i,j} = (\nabla f_r)_i(j)\big|_{r=r^*}$. It is easy to see that $\Sigma_{\pi^0} = \Psi\Xi\Psi^T$.

We demonstrate that

$$M = \nabla^2 h(\pi^0) = \Psi^T(\Sigma_{\pi^0})^{-1}\Psi, \tag{5.7}$$

and prove that the other technical conditions of Lemma 4.2.2 are satisfied. The rest will follow from Lemma 4.2.2, since

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

and similarly

$$\mathsf{tr}\,(M\Xi M\Xi) = \mathsf{tr}\,(I_d) = d.$$

By the assumption, when $\mu \in B$ we can define a function $r(\mu)$ such that $r(\mu)$ is the maximizer of the definition of the mismatched divergence. We will first prove that around an open neighborhood of π^0, $r(\mu)$ is a continuously differentiable function of μ. The first order optimality condition in the right-hand side of (3.6) gives

$$\mu(\nabla f_r) - \frac{\pi^0(e^{f_r}\nabla f_r)}{\pi^0(e^{f_r})} = 0. \tag{5.8}$$

The derivative of the left-hand side of (5.8) with respect to r is given by

$$\begin{aligned} & \nabla\big(\mu(\nabla_r f_r) - \frac{\pi^0(e^{f_r}\nabla f_r)}{\pi^0(e^{f_r})}\big) \\ = \;& \mu(\nabla^2 f_r) - [\frac{\pi^0\big(e^{f_r}\nabla f_r \nabla f_r^T\big) + \pi^0\big(e^{f_r}\nabla^2 f_r\big)}{\pi^0(e^{f_r})} - \frac{\pi^0\big(e^{f_r}\nabla f_r\big)\pi^0\big(e^{f_r}\nabla f_r^T\big)}{(\pi^0(e^{f_r}))^2}]. \end{aligned} \tag{5.9}$$

When $\mu = \pi^0$, (5.8) is satisfied with $f_r = 0$ by the hypothesis. Then the derivative (5.9) is given by the negative of $\Sigma_{\pi^0} = \pi^0(\nabla f_r(\nabla f_r)^T) - \pi^0(\nabla f_r)\pi^0((\nabla f_r)^T)\big|_{r=r^*}$ which is positive definite by the hypothesis. Therefore, by the implicit function theorem, around an open

neighborhood $U \subseteq B \cap r^{-1}(B_1)$ around $\mu = \pi^0$, $r(\mu)$ is continuously differentiable.

On taking the derivative of (5.8) with respect to $\mu(z)$ with $z \in \mathsf{Z}$, we have

$$\nabla f_r(z) + \nabla\big[\mu(\nabla_r f_r) - \frac{\pi^0(e^{f_r}\nabla f_r)}{\pi^0(e^{f_r})}\big]\frac{\partial r(\mu)}{\partial \mu(z)}\big|_{r=r(\mu)} = 0. \tag{5.10}$$

When $\mu = \pi^0$, we have

$$\nabla f_r(z)\big|_{r=r*} = \Sigma_{\pi^0}\frac{\partial r(\mu)}{\partial \mu(z)}\big|_{\mu=\pi^0}. \tag{5.11}$$

It is straightforward to see that

$$\frac{\partial}{\partial \mu(z)}h(\mu) = f_{r(\mu)}(z),$$

$$\frac{\partial^2}{\partial \mu(z)\partial \mu(\bar{z})}h(\mu) = \nabla f_r(z)^T\big|_{r=r^*}\frac{\partial r(\mu)}{\partial \mu(\bar{z})}. \tag{5.12}$$

When $\mu = \pi^0$, note that $r(\mu) = r_0$ and applying (5.11), we have

$$\frac{\partial^2}{\partial \mu(z)\partial \mu(\bar{z})}h(\mu)\big|_{\mu=\pi^0} = (\nabla f_r(z))^T(\Sigma_{\pi^0})^{-1}\nabla f_r(\bar{z})\big|_{r=r^*}. \tag{5.13}$$

Now since $r(\mu)$ is continuously differentiable on U, and $f_r(z)$ is smooth in r for each z, we have that $\frac{\partial^2}{\partial \mu(z)\partial \mu(\bar{z})}f(\mu) = \nabla f_{r(\mu)}(z)^T\frac{\partial r(\mu)}{\partial \mu(\bar{z})}$ is continuous on U. Then we can pick a compact set K such that

$$K \subset U \cap \{\mu \in \mathcal{P}(\mathsf{Z}) : \max_u |\mu(u) - \pi^0(u)| < \tfrac{1}{2}\min_u |\pi^0(u)|\},$$

and K contains π^0 as an interior point. It follows that $-\frac{1}{n}\log \mathsf{P}\{S^n \notin K\} > 0$. In sum, we can pick K so that all the technical conditions on K outlined in Lemma 4.2.2 are satisfied.

(2) To prove the result for the case $\pi = \pi^1$, we can use an argument similar to that of Theorem 4.2.4, and use Lemma 5.2.2 in place of Lemma 4.2.3. □

Proof of Proposition 5.1.4. It suffices to prove that for the robust test, the conditions of Theorem 5.2.1 are satisfied.

The maximum at π^0 is clearly achieved by $r = 0$, and it is easy to see that the function f_r is of C^2 in the open neighborhood near $r = 0$.

When $r_0 = 1$, the Hessian of $\log(\mu(r_0 + r^T\psi))$ with respect to r is given by $H(\mu) = \mu\big(\frac{\psi\psi^T}{(r_0+r^T\psi)^2}\big)$. At $\mu = \pi^0$, it is equal to Σ_{π^0} which is positive definite. Therefore, when μ is in a neighborhood B of π^0, the Hessian $H(\mu)$ is also positive definite. Thus the maximizer in (5.1) is unique for $r_0 = 1$. Thus the maximizer of mismatched divergence using function class $\mathcal{F}$ is unique. □

Chapter 6

Bias and Variance in Learning

In most applications, the underlying distribution π^0 is not given and has to be learned from data. In this chapter we study a simple case in which m i.i.d. samples with marginal π^0 are given as the training data. We then use the resulting empirical distribution $\bar{\Gamma}_m$ in place of π^0 in the test statistics, i.e. $D^{\text{MM}}(\Gamma^n \| \bar{\Gamma}_m)$ instead of $D^{\text{MM}}(\Gamma^n \| \pi^0)$, where Γ^n is the empirical distribution from π. However, $D^{\text{MM}}(\Gamma^n \| \bar{\Gamma}_m)$ can be unbounded when the support of $\bar{\Gamma}_m$ is a strict subset of the support of Γ^n. Thus, we use the following test instead:

$$\phi^{\text{MM}}(\boldsymbol{Z}) = \mathbb{I}\{D^{\text{MM}}_{\mathcal{F}}(\Gamma^n \| \bar{\Gamma}_m) \wedge \bar{M} \geq \delta_n\}, \tag{6.1}$$

where $\bar{M}$ is a constant chosen large enough so that $\bar{M}$ is much larger than $\max_z \log(\frac{1}{\pi^0(z)})$.

It is clear that the test statistic $D^{\text{MM}}_{\mathcal{F}}(\Gamma^n \| \bar{\Gamma}_m) \wedge \bar{M}$ will converge to $D^{\text{MM}}(\pi \| \pi^0)$ asymptotically when both n and m go to infinity. Motivated by results in the previous chapters, we would like to investigate its finite length properties. In this chapter we derive the asymptotic bias and variance of $D^{\text{MM}}(\pi^0 \| \bar{\Gamma}_m) \wedge \bar{M}$. Extending this result to the case where $\pi \neq \pi^0$ is an ongoing study.

Our main result in this chapter is given in the following proposition:

Proposition 6.0.3. *Suppose $\boldsymbol{Z}$ is drawn i.i.d. from a finite set Z with marginal π^0 and assume Σ_{π^0} is positive definite. Let $\bar{\Gamma}_m(z) = \frac{1}{m}\sum_{i=1}^m \mathbb{I}\{Z_i = z\}$. Then $D^{\text{MM}}(\pi^0 \| \bar{\Gamma}_m) \wedge \bar{M}$ has bias of order n^{-1} and variance of order n^{-2}, and the normalized asymptotic values have*

the following simple, explicit forms:

$$\lim_{m\to\infty} m\mathsf{E}[D^{\text{MM}}(\pi^0\|\bar{\Gamma}_m)\wedge\bar{M}] = \tfrac{1}{2}d,$$
$$\lim_{m\to\infty} m^2\mathsf{Var}\,[D^{\text{MM}}(\pi^0\|\bar{\Gamma}_m)\wedge\bar{M}] = \tfrac{1}{2}d.$$

Proposition 6.0.3 suggests that

1. When using the empirical distribution in place of the true underlying distribution, the bias and variance of test statistics using mismatched divergence can be much smaller than that using KL divergence. This suggests the possibility that using mismatched divergence could require less training data, though this requires confirmation through further experimental study and analysis.

2. The bias term suggests that we should use a threshold $\delta_{n,m}$ that also depends on the number of training samples m.

Proof of Proposition 6.0.3. The proof is similar to that of Theorem 4.2.1. To apply Lemma 4.2.2, h is specialized to be $h(\mu) := D^{\text{MM}}(\pi^0\|\mu)\wedge\bar{M}$ and take $X^i = (\mathbb{I}_{z_1}(Z_i),\mathbb{I}_{z_2}(Z_i),\ldots,\mathbb{I}_{z_N}(Z_i))^T$ and $\mathsf{X} = [0,1]^d$. Let $\Xi = \mathsf{Cov}\,(X)$. Redefine the matrix Ψ as $\Psi_{i,j} = \psi_i(j)$. Also denote the vector valued function $\psi = [\psi_1,\ldots,\psi_d]^T$. It is easy to see that $\Sigma_{\pi^0} = \Psi\Xi\Psi^T$.

We will demonstrate the gradient and Hessian of $h(\pi^0)$ are given by

$$\nabla h(\pi^0) = -\mathbf{1}, \tag{6.2}$$

$$M = \nabla^2 h(\pi^0) = \mathbf{1}\mathbf{1}^T + (\mu(\psi)\mathbf{1}^T + \Psi)^T(\Sigma_{\pi^0})^{-1}(\mu(\psi)\mathbf{1}^T + \Psi), \tag{6.3}$$

and prove that the other technical conditions of Lemma 4.2.2 are satisfied. Note that $\Xi\mathbf{1} = 0$. The rest will follow from Lemma 4.2.2, since

$$\nabla h(\pi^0)^T(\mu - \pi^0) = -\mathbf{1}^T(\mu - \pi^0) = 0, \tag{6.4}$$

as required in the lemma. The limiting values in the lemma are

$$\begin{aligned}
\mathsf{tr}\,(M\Xi) &= \mathsf{tr}\,(\mathbf{1}\mathbf{1}^T\Xi + (\mu(\psi)\mathbf{1}^T + \Psi)^T(\Sigma_{\pi^0})^{-1}(\mu(\psi)\mathbf{1}^T + \Psi)\Xi) \\
&= \mathsf{tr}\,\big(\Psi^T(\Sigma_{\pi^0})^{-1}\Psi\Xi + (\mu(\psi)\mathbf{1}^T + \Psi)^T(\Sigma_{\pi^0})^{-1}\mu(\psi)\mathbf{1}^T\Xi + (\mu(\psi)\mathbf{1}^T)^T(\Sigma_{\pi^0})^{-1}\Psi\Xi\big) \\
&= \mathsf{tr}\,(\Psi^T(\Sigma_{\pi^0})^{-1}\Psi\Xi) + \mathsf{tr}\,((\mu(\psi)\mathbf{1}^T)^T(\Sigma_{\pi^0})^{-1}\Psi\Xi) \\
&= \mathsf{tr}\,((\Sigma_{\pi^0})^{-1}\Psi\Xi\Psi^T) + \mathsf{tr}\,(\Xi\mathbf{1}\mu(\psi)^T(\Sigma_{\pi^0})^{-1}\Psi) \\
&= \mathsf{tr}\,(I_d) + 0 = d, \qquad (6.5)
\end{aligned}$$

and similarly

$$\mathsf{tr}\,(M\Xi M\Xi) = \mathsf{tr}\,(I_d) = d.$$

The Hessian of $\pi^0(f_r) - \Lambda_\mu(f_r)$ at $\mu = \pi^0$ is given by the positive defnite matrix Σ_{π^0}. Thus, the objective function of the right-hand side of (3.6) is strictly concave and thus has a unique maximum for each μ in an open neighborhood B of π^0. Let $r(\mu)$ be the maximizer for a given μ.

The first order optimality condition in the right-hand side of (3.6) gives

$$\pi^0(\psi) - \frac{\mu(e^{r(\mu)^T\psi}\psi)}{\mu(e^{r(\mu)^T\psi})} = 0 \quad \text{for all } i.$$

On taking the derivative with respect to μ_z with $z \in \mathsf{Z}$, we have

$$\begin{aligned}
0 &= -\frac{\partial}{\partial\mu(z)}\Big(\frac{\mu(e^{r^T\psi}\psi)}{\mu(e^{r^T\psi})}\Big)\Big|_{r=r(\mu)} - \nabla_r\Big(\frac{\mu(e^{r^T\psi}\psi)}{\mu(e^{r^T\psi})}\Big)\Big|_{r=r(\mu)}\frac{\partial r(\mu)}{\partial\mu(z)} \\
&= -\Big(\frac{e^{r^T\psi(z)}\psi(z)}{\mu(e^{r^T\psi})} - \frac{\mu(e^{r^T\psi}\psi)e^{r^T\psi(z)}}{\mu(e^{r^T\psi})^2}\Big)\Big|_{r=r(\mu)} \\
&\quad -\Big[\frac{\mu(e^{r^T\psi}\psi\psi^T)}{\mu(e^{r^T\psi})} - \frac{\mu(e^{r^T\psi}\psi)\mu(e^{r^T\psi}\psi^T)}{\mu(e^{r^T\psi})^2}\Big]\Big|_{r=r(\mu)}\frac{\partial r(\mu)}{\partial\mu(z)}.
\end{aligned}$$

When $\mu = \pi^0$, we have $r(\pi^0) = 0$. Consequently,

$$-(\psi(z) - \mu(\psi)) = \Sigma_{\pi^0} \frac{\partial r(\mu)}{\partial \mu(z)}\Big|_{\mu=\pi^0}. \tag{6.6}$$

Since $h(\pi^0) = 0$ and $h(\mu)$ is continuous in B, there exists an open set $B_1 \subseteq B$ such that $h(\mu) \leq \bar{M}$ for $\mu \in B_1$. Thus, for $\mu \in B_1$, $h(\mu) = D^{\text{MM}}(\pi^0 \| \mu)$. The following is straightforward:

$$\frac{\partial}{\partial \mu(z)} h(\mu) = -\frac{e^{r^T \psi(z)}}{\mu(e^{r^T \psi})},$$

$$\frac{\partial^2}{\partial \mu(z) \partial \mu(\bar{z})} h(\mu) = \frac{e^{r^T \psi(z)} e^{r^T \psi(\bar{z})}}{\mu(e^{r^T \psi})^2}\Big|_{r=r(\mu)} - \Big(\frac{e^{r^T \psi(z)} \psi(z)}{\mu(e^{r^T \psi})} - \frac{\mu(e^{r^T \psi} \psi) e^{r^T \psi(z)}}{\mu(e^{r^T \psi})^2}\Big)^T\Big|_{r=r(\mu)} \frac{\partial r(\mu)}{\partial \mu(\bar{z})}.$$

When $\mu = \pi^0$, we obtain

$$\frac{\partial}{\partial \mu(z)} h(\mu) = -1,$$

$$\frac{\partial^2}{\partial \mu(z) \partial \mu(\bar{z})} h(\mu)\Big|_{\mu=\pi^0} = 1 + (\psi(z) - \mu(\psi))^T \Sigma_{\pi^0}^{-1} (\psi(z) - \mu(\psi)).$$

We now verify the remaining conditions required in applying Lemma 4.2.2:

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

2. The function h is uniformly bounded.

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

4. Pick a compact set K that contains π^0 as an interior point and

$$K \subset B_1 \cap \{\mu \in \mathcal{P}(\mathsf{Z}) : \max_u |\mu(u) - \pi^0(u)| < \tfrac{1}{2} \min_u |\pi^0(u)|\}.$$

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 B_1, it follows that h is C^2 on $K \subset B_1$.

□

Chapter 7

Other Properties of Mismatched Divergence

In this chapter we derive other properties of the mismatched divergence: mainly its difference and similarity to the KL divergence, and its geometric interpretation.

An important property of the KL divergence $D(\mu\|\pi)$ is that it is convex with respect to (μ, π). The mismatched divergence inherits this property:

Lemma 7.0.4. *$D^{\text{MM}}(\mu\|\pi)$ is convex in (μ, π).*

Note that this is stronger than being convex in both μ and π.

Proof. For a given f, $\mu(f)$ is linear in μ. Since $\pi(e^f)$ is linear in π, $\Lambda_\pi(f)$ is concave in π. Therefore, $\mu(f) - \Lambda_\pi(f)$ is convex in (μ, π). The result follows as it is well known that the supremum of a set of convex functions is convex [24]. □

The other properties in this chapter are specialized to the case where the function class is linear.

7.1 Linear Function Class

Recall that the mismatched divergence using linear function class can be written as:

$$D(\mu\|\pi) = \sup_{r\in\mathbb{R}^d} \big(\mu(f_r) - \Lambda_\pi(f_r)\big).$$

7.1.1 Geometric interpretations

Here we give a geometric interpretation of the mismatched divergence. The result also holds when the function class $\mathcal{F}$ is a pointed cone.[1]

For a given function $f \in \mathcal{F}$ and $c \in \mathbb{R}$, we define a subset of $\mathcal{P}(\mathsf{Z})$ by $\mathcal{H} = \{\mu \in \mathcal{P}(\mathsf{Z}) : \langle \mu, f \rangle = c\}$. This set is interpreted as a hyper-plane, even though it is restricted to the simplex $\mathcal{P}$. The associated "half spaces" are defined by

$$\mathcal{H}^-_{f,c} = \{\mu \in \mathcal{P}(\mathsf{Z}) : \mu(f) \le c\}, \qquad \mathcal{H}^+_{f,c} = \{\mu \in \mathcal{P}(\mathsf{Z}) : \mu(f) \ge c\}. \tag{7.1}$$

The set $\mathcal{Q}^{\text{GM}}_\alpha(\pi) \subset \mathcal{M}_1$ is defined as the intersection of all half spaces defined using functions from $\mathcal{F}$ that contain $\mathcal{Q}_\alpha(\pi)$. Formally, for each $\alpha \ge 0$,

$$\mathcal{Q}^{\text{GM}}_\alpha(\pi) := \bigcap \big\{ H^-_{f,c} : \mathcal{Q}_\alpha(\pi) \subset H^-_{f,c}, f \in \mathcal{F}, c \in \mathbb{R} \big\}. \tag{7.2}$$

Theorem 7.1.1. *When $\mathcal{F} \subset \mathcal{F}^*_\pi$ is a pointed cone, the mismatched divergence has the following geometrical interpretation:*

$$D^{\text{MM}}(\mu\|\pi) = \inf\{\alpha : \mu \in \mathcal{Q}^{\text{GM}}_\alpha(\pi)\} = \sup_{f \in \mathcal{F}} \inf_{\nu} \{D(\nu\|\pi) : \nu(f) \ge \mu(f)\}. \tag{7.3}$$

To prove this theorem, we need the following lemma.

Lemma 7.1.2. *For any $G : \mathsf{Z} \to \mathbb{R}$ and $c \in R$,*

$$\inf\{D(\mu\|\pi) : \mu(G) \ge c\} = \sup_{\theta \ge 0} \big(\theta c - \Lambda_\pi(\theta G) \big). \tag{7.4}$$

Proof. This is a direct consequence of Sanov's and Cramer's theorems. □

[1] A cone is pointed if it includes the origin.

Proof of Theorem 7.1.1.

$$\begin{aligned}\mu \notin \mathcal{Q}_\alpha^{\mathrm{GM}}(\pi) &\Rightarrow \text{ There exists } f \in \mathcal{F} \text{ such that } \inf_\nu\{D(\nu\|\pi) : \nu(f) \geq \mu(f)\} \geq \alpha \\ &\Rightarrow \sup_{f\in\mathcal{F}} \inf_\nu\{D(\nu\|\pi) : \nu(f) \geq \mu(f)\} \geq \alpha.\end{aligned}$$

Consequently, for any $\epsilon > 0$, we have

$$\sup_{f\in\mathcal{F}} \inf_\nu\{D(\nu\|\pi) : \nu(f) \geq \mu(f)\} + \epsilon \in \{\alpha : \mu \in \mathcal{Q}_\alpha^{\mathrm{GM}}(\pi)\}.$$

Since ϵ is arbitrary, this means that we have

$$\inf\{\alpha : \mu \in \mathcal{Q}_\alpha^{\mathrm{GM}}(\pi)\} \leq \sup_{f\in\mathcal{F}} \inf_\nu\{D(\nu\|\pi) : \nu(f) \geq \mu(f)\}.$$

Similarly, we also have

$$\begin{aligned}\mu \in \mathcal{Q}_\alpha^{\mathrm{GM}}(\pi) &\Rightarrow \text{ For any } f \in \mathcal{F}, \inf_\nu\{D(\nu\|\pi) : \nu(f) \geq \mu(f)\} \leq \alpha \\ &\Rightarrow \sup_{f\in\mathcal{F}} \inf_\nu\{D(\nu\|\pi) : \nu(f) \geq \mu(f)\} \leq \alpha,\end{aligned}$$

which implies that

$$\inf\{\alpha : \mu \in \mathcal{Q}_\alpha^{\mathrm{GM}}(\pi)\} \geq \sup_{f\in\mathcal{F}} \inf_\nu\{D(\nu\|\pi) : \nu(f) \geq \mu(f)\}.$$

Therefore,

$$\begin{aligned}\inf\{\alpha : \mu \in \mathcal{Q}_\alpha^{\mathrm{GM}}(\pi)\} &= \sup_{f\in\mathcal{F}} \inf_\nu\{D(\nu\|\pi) : \nu(f) \geq \mu(f)\} \\ &= \sup_{f\in\mathcal{F}} \sup_{\theta\geq 0}\{\theta\mu(f) - \Lambda_\pi(\theta f)\} \\ &= \sup_{f\in\mathcal{F}}\{\mu(f) - \Lambda_\pi(f)\},\end{aligned}$$

where the last equality follows from the assumption that $\mathcal{F}$ is a pointed cone. □

7.1.2 Necessary and sufficient conditions for $D^{\text{MM}}(\mu\|\pi) = 0$ and $D^{\text{MM}}(\mu\|\pi) = \infty$

It is natural to ask how to interpret $D^{\text{MM}}(\mu\|\pi) = 0$ and $D^{\text{MM}}(\mu\|\pi) = \infty$. When $D^{\text{MM}}(\mu\|\pi) = 0$, loosely speaking, μ and π cannot be strictly separated by any hyperplane defined using any function $f \in \mathcal{F}$. Formally we have the following:

Lemma 7.1.3. *The following three statements are equivalent:*[2]

1. $D^{\text{MM}}(\mu\|\pi) = 0$,
2. $D^{\text{MM}}(\pi\|\mu) = 0$,
3. $\mu(f_r) = \pi(f_r)$ *for all* r.

Here we give a proof based on the generalized Pinsker's inequality.

Proof. We first prove (1) indicates (3): By the generalized Pinsker's inequality, we obtain from $D^{\text{MM}}(\mu\|\pi) = 0$ that

$$\sup \left(\frac{\mu(f_r) - \pi(f_r)}{\|f_r\|_{\infty,\text{SP}}} \right)^2 = 0.$$

Consequently, $\mu(f_r) = \pi(f_r)$ for all r. We now prove that (3) indicates (2): Since $\mu(f_r) = \pi(f_r)$ for all r, we have

$$D^{\text{MM}}(\pi\|\mu) = \sup_r \{\mu(f_r) - \Lambda_\pi(f_r)\} = \sup_r \{\pi(f_r) - \Lambda_\pi(f_r)\} = D^{\text{MM}}(\pi\|\pi) = 0,$$

where the last equality is obtained using the following chain of inequalities from Lemma 3.2.1:

$$0 \leq D^{\text{MM}}(\pi\|\pi) \leq D(\pi\|\pi) = 0.$$

[2]To the author's best knowledge, the fact the first and second are equivalent was first proved using a different approach by Jayakrishnan Unnikrishnan.

Since the second statement is symmetric in μ and π, the above argument also indicates that (3) implies (1) and (2) implies (3). $\square$

The case when $D^{\text{MM}}(\mu\|\pi) = \infty$ is a little bit more complicated. We use ess $\sup_\pi(f)$ to denote the essential supremum $\inf\{\alpha : \pi(\{x : f(x) > \alpha\}) = 0\}$. The following lemma illustrates a sufficient condition and a necessary condition:

Lemma 7.1.4. *If there exists $f \in \mathcal{F}$ such that $\mu(f) >$ ess $\sup_\pi(f)$, then $D^{\text{MM}}(\mu\|\pi) = \infty$; If the supremum in the definition of $D^{\text{MM}}(\mu\|\pi)$ is not achieved, in particular if $D^{\text{MM}}(\mu\|\pi) = \infty$, then there exists $f \in \mathcal{F}$ such that $\mu(f) \geq$ ess $\sup_\pi(f)$.*

Note that the necessary and sufficient conditions differ in whether the equality holds. The following two examples in which $\mu(f) =$ ess $\sup_\pi(f)$ for all $f \in \mathcal{F}$ illustrate that the lemma is almost the best possible.

Example 7.2. *Consider the one-dimensional function class $\psi(x) = x$. Let $\mu(\{0\}) = \pi(\{0\}) = 1$. Then $\mu(f) =$ ess $\sup_\pi(f)$ for any $f \in \mathcal{F}$ and $D^{\text{MM}}(\mu\|\pi) = 0$.*

Example 7.3. *Consider again the one-dimensional function class $\psi(x) = x$. Let μ and π be the discrete probability measure: $\mu(\{0\}) = 1$; for all positive integer k, $\pi(\{-\frac{1}{k}\}) = 2^{-k}$. It is easy to see that for any $f \in \mathcal{F}$,* ess $\sup_\pi(f) = 0 = \mu(f)$. *We now show that $D^{\text{MM}}(\mu\|\pi) = \infty$:*

$$\begin{aligned} -D^{\text{MM}}(\mu\|\pi) &= \inf_\theta \log\Big(\sum_{k=1}^{\infty} e^{-(\frac{\theta}{k}+k\log 2)}\Big) \\ &\leq \inf_{\theta\geq 0} \log\Big(\sum_{k=1}^{M} e^{(\frac{\theta}{k}+k\log 2)} + 2^{-M}\Big) \\ &\leq \inf_{\theta\geq 0} \log\big(Me^{-2\sqrt{\theta\log 2}} + 2^{-M}\big) \\ &= -M\log(2). \end{aligned}$$

Since this holds for any $M > 0$, we have $D^{\text{MM}}(\mu\|\pi) = \infty$.

Proof of Lemma 7.1.4. We first prove that $\mu(f) > \text{ess sup}_\pi(f)$ implies $D^{\text{MM}}(\mu\|\pi) = \infty$. The following is straightforward: for any $\alpha > 0$,

$$\frac{\Lambda_\pi(\alpha f)}{\alpha} \le \operatorname*{ess\,sup}_\pi(f),$$

Consequently,

$$\liminf_{\alpha\to\infty} \frac{\mu(\alpha f) - \Lambda_\pi(\alpha f)}{\alpha} \ge \mu(f) - \operatorname*{ess\,sup}_\pi(f) > 0.$$

Thus

$$D^{\text{MM}}(\mu\|\pi) \ge \liminf_{\alpha\to\infty} \mu(\alpha f) - \Lambda_\pi(\alpha f) = \infty.$$

We now prove that if the supremum is not achieved, then $\mu(f) \ge \text{ess sup}_\pi(f)$ by giving a construction of one such function: Since the supremum in the definition of mismatched divergence is not achieved, there exists $\{f_n = r_n^T\psi\} \subset \mathcal{F}$ such that

$$\liminf_{n\to\infty} \mu(f_n) - \Lambda_\pi(f_n) > 0. \tag{7.5}$$

Thus, taking a subsequence if necessary, we can assume that the sequence of vectors $\{r_n\}$ associated with $\{f_n\}$ satisfies $\|r_n\| \to \infty$ as $n \to \infty$ and $\mu(f_n) - \Lambda_\pi(f_n) > 0$ for every n. Define $g_n = \frac{f_n}{\|r_n\|}$. By considering a subsequence if necessary, we can assume without loss of generality that the sequence g_n is convergent point-wise. Define $g_\infty = \lim_{n\to\infty} g_n$. Clearly, $g_\infty \in \mathcal{F}$. We will prove that g_∞ satisfies $\mu(g_\infty) \ge \text{ess sup}_\pi(g_\infty)$.

Let $b_0 = \text{ess sup}_\pi(g_\infty)$. We have for any $\epsilon > 0$,

$$\pi\{x : g_\infty \ge b_0 - \frac{1}{2}\epsilon\} > 0.$$

Since $\mathbb{I}\{g_n \ge b_0 - \epsilon\}\mathbb{I}\{g_\infty \ge b_0 - \frac{1}{2}\epsilon\}$ converges to $\mathbb{I}\{g_\infty \ge b_0 - \frac{1}{2}\epsilon\}$ point-wise, by the

dominated convergence theorem, there is an $n(\epsilon)$ such that for $n > n(\epsilon)$

$$\pi\{x : g_n \geq b_0 - \epsilon\} > 0.$$

Therefore,

$$\begin{aligned}
\frac{1}{\|r_n\|}\Lambda_\pi(f_n) &= \frac{1}{\|r_n\|}\Lambda_\pi(\|r_n\|g_n) \\
&= \frac{1}{\|r_n\|}\log\big(\pi(e^{\|r_n\|g_n})\big) \\
&\geq \frac{1}{\|r_n\|}\log\big(\pi\{x : g_n(x) \geq b_0 - \epsilon\}e^{\|r_n\|(b_0-\epsilon)}\big)\big).
\end{aligned}$$

Therefore,

$$\liminf_{n\to\infty} \frac{1}{\|r_n\|}\Lambda_\pi(f_n) \geq b_0 - \epsilon.$$

Since this holds for any $\epsilon > 0$, we have

$$\liminf_{n\to\infty} \frac{1}{\|r_n\|}\Lambda_\pi(f_n) \geq b_0. \tag{7.6}$$

Consequently,

$$\limsup_{n\to\infty} \frac{1}{\|r_n\|}\big(\mu(f_n) - \Lambda_{\pi^0}(f_n)\big) \leq \mu(g_\infty) - b_0.$$

On the other hand, we obtain from the fact that $\mu(f_n) - \Lambda_\pi(f_n) > 0$ for every n:

$$0 \leq \limsup_{n\to\infty} \frac{1}{\|r_n\|}\big(\mu(f_n) - \Lambda_{\pi^0}(f_n)\big).$$

Therefore,

$$0 \leq \mu(g_\infty) - b_0 = \mu(g_\infty) - \operatorname*{ess\,sup}_\pi(g_\infty).$$

Thus, $g_\infty \in \mathcal{F}$ is a function that satisfies $\mu(g_\infty) \geq \text{ess sup}_\pi(g_\infty)$. □

Chapter 8

Conclusions

We investigated the asymptotic bias and variance of the Hoeffding test and mismatched universal test. We have shown that the asymptotic bias and variance of mismatched universal test can be much smaller than the Hoeffding test. In addition, we showed that the mismatched universal test includes a robust test as a special case. Consequently, the bias and variance of the robust test increase proportionally to the co-dimension of the uncertainty set.

We also investigated the performance of the test when the distribution of the null hypothesis is learned from data. As a preliminary result, we showed that the bias and variance depend on the number of training samples as well as the dimensionality of the function class.

We developed other properties of the mismatched divergence. In particular, we showed that the mismatched divergence admits a generalized Pinsker's inequality.

For future work, there are many important problems:

1. The mismatched universal test is optimal when the log-likelihood ratio is in the function class. It is not clear what the performance is when the log-likelihood ratio is not in the function class. One question in this direction is how many distributions can be distinguished using a mismatched divergence test based on function classes of a given dimension.

2. The performance of the mismatched universal test depends on the function class used. Therefore, it is important to study how to choose the function class.

3. We made some preliminary study on how the function class impacts the test when

the underlying distribution is learned. An interesting question is to derive PAC type bounds, and study how the dimensionality affects the probability of error.

4. We have shown that a robust test is a special case of mismatched divergence. One question in this direction is to find connections between mismatched divergence and other distance/divergence, such as the f-divergence in [16] and other generalizations defined in [10].

Appendix A

Proofs of Lemmas 4.2.5 and 4.2.7

A.1 Proof of Lemma 4.2.5

Proof of Lemma 4.2.5. In our case, to apply Lemma 4.2.2, h is specialized to be $h(\mu) := D^{\text{MM}}(\mu\|\pi^0)$, and take $X^i = (\mathbb{I}_{z_1}(Z_i), \mathbb{I}_{z_2}(Z_i), \ldots, \mathbb{I}_{z_N}(Z_i))^T$, and $\mathsf{Z} = [0,1]^N$. Take $\Xi = \mathsf{Cov}\,(X)$. Define the matrix Ψ as $\Psi_{i,j} = \psi_i(j)$. Also denote the vector valued function $\psi = [\psi_1, \ldots, \psi_d]^T$. It is easy to see that $\Sigma_{\pi^0} = \Psi\Xi\Psi^T$.

We demonstrate that

$$M = \nabla^2 h(\pi^0) = \Psi^T(\check{\Sigma}_\pi)^{-1}\Psi,$$

and prove that the other technical conditions of Lemma 4.2.2 are satisfied. The rest follows from Lemma 4.2.2, since

$$\mathsf{tr}\,(M\Xi) = \mathsf{tr}\,((\check{\Sigma}_\pi)^{-1}\Psi\Xi\Psi^T) = \mathsf{tr}\,(\check{\Sigma}_\pi^{-1}\Sigma_{\pi^0}),$$

and similarly

$$\mathsf{tr}\,(M\Xi M\Xi) = \mathsf{tr}\,(\check{\Sigma}_\pi^{-1}\Sigma_{\pi^0}\check{\Sigma}_\pi^{-1}\Sigma_{\pi^0}).$$

The condition Σ_{π^0} being positive definite indicates that the objective function of the right-hand side of (3.6) is strictly concave and thus has a unique maximum for each μ. Let $r(\mu)$ be the maximizer for a given μ. Then

$$h(\mu) = \mu(f_{r(\mu)}) - \Lambda_{\pi^0}(f_{r(\mu)}).$$

Recall that $\check{\pi}_\mu$ is the twisted distribution defined in (3.3). Define $\check{\Sigma}_\mu$ as

$$\check{\Sigma}_{\mu,i,j} = \check{\pi}_\mu(\psi_i\psi_j) - \check{\pi}_\mu(\psi_i)\check{\pi}_\mu(\psi_j),$$

The first order optimality condition in the right-hand side of (3.6) gives

$$\mu(\psi) - \check{\pi}_\mu(\psi) = 0.$$

On taking the derivative with respect to μ_z with $z \in \mathsf{Z}$, we have

$$\psi(z) - \check{\Sigma}_\mu \frac{\partial r(\mu)}{\partial \mu(z)} = 0.$$

Then it is straingforward to show that

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

$$\frac{\partial^2}{\partial \mu(z)\partial \mu(\bar{z})} h(\mu) = \psi^T(z) \frac{\partial r(\mu)}{\partial \mu(\bar{z})} = \psi^T(z)\check{\Sigma}_\mu^{-1}\psi(\bar{z}).$$

When $\mu = \pi^1$, we have $r(\pi) = 0$ and $\check{\Sigma}_\mu = \check{\Sigma}_\pi$. Thus,

$$\frac{\partial^2}{\partial \mu(z)\partial \mu(\bar{z})} f(\pi) = \sum_i \psi_i(z)\check{\Sigma}_\pi^{-1}\psi(\bar{z}).$$

We now verify the remaining conditions required in Lemma 4.2.2:

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

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

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

4. Pick a compact K that contains π^0 as an interior point and

$$K \subset \{\mu \in \mathcal{P}(\mathsf{Z}) : \max_u |\mu(u) - \pi^0(u)| < \tfrac{1}{2} \min_u |\pi^0(u)|\}.$$

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 K, it follows that h is C^2 on K.

□

A.2 Proof of Lemma 4.2.7

Proof of Lemma 4.2.7. The supremum is of course achieved when $\mu = \pi$. Thus we only need to prove the case $\mu \neq \pi$. Using Lemma 7.1.4, since $\mu \preceq \pi$ and $\mu \neq \pi$, for any f $\mu(f) < \text{ess sup}_\pi(f)$; therefore the supremum is achieved. Since the Hessian of $\mu(f_r) - \Lambda_\pi(f_r)$ is given by Σ_π which is positive definite, we have that the function $\mu(f_r) - \Lambda_\pi(f_r)$ is strictly concave and the maximizer is unique. □

References

[1] Y. Jing and S. Baluja, "Pagerank for product image search," in *WWW '08: Proceedings of the 17th International Conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 307 – 316.

[2] "Mobvis project." [Online]. Available: http://www.mobvis.org/

[3] G. Stolovitzky, R. J. Prill, and A. Califano, "Lessons from the DREAM2 challenges," *Annals of the New York Academy of Sciences*, vol. 1158, pp. 159 – 195, 2009.

[4] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264 – 280, 1971.

[5] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: A survey of some recent advances," *ESAIM: Probability and Statistics*, vol. 9, pp. 323 – 375, Nov. 2005.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[7] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, pp. 60 – 62, 1938. [Online]. Available: http://www.jstor.org/stable/2957648

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

[9] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3064 – 3074, Sep. 2005.

[10] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," Department of Statistics, UC Berkeley, Tech. Rep. 764, Jan. 2007.

[11] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, vol. 36, pp. 369 – 401, 1965. [Online]. Available: http://www.jstor.org/stable/2238145

[12] E. Abbe, M. Medard, S. Meyn, and Z. Lizhong, "Finding the best mismatched detector for channel coding and hypothesis testing," in *Information Theory and Applications Workshop, 2007*, 29 Feb. 2007, pp. 284 – 288.

[13] C. Pandit and S. Meyn, "Worst-case large-deviation asymptotics with application to queueing and information theory," *Stochastic Processes and their Applications*, vol. 116, no. 5, pp. 724 – 756, 2006.

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

[15] J. Unnikrishnan, D. Huang, S. Meyn, A. Surana, and V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," *IEEE Transactions on Information Theory*, submitted for publication. [Online]. Available: http://arxiv.org/abs/0909.2234

[16] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417 – 528, 2004.

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

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

[19] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 285 – 290, Mar. 1991.

[20] N. Merhav, G. Kaplan, A. Lapidoth, and S. S. Shitz, "On information rates for mismatched decoders," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1953 – 1967, Nov. 1994.

[21] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes.* Moscow, U.S.S.R.: Izv. Akad. Nauk, 1960.

[22] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13 – 30, 1963. [Online]. Available: http://www.jstor.org/stable/2282952

[23] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1 - 2, pp. 1 – 305, 2008.

[24] D. P. Bertsekas, A. Nedic, and A. Ozdaglar, *Convex Analysis and Optimization.* Nashua, NH, USA: Athena Scientific, 2003.