December 2016

ROBUST KULLBACK-LEIBLER DIVERGENCE AND ITS APPLICATIONS IN UNIVERSAL HYPOTHESIS TESTING AND DEVIATION DETECTION

Pengfei Yang
Syracuse University

Abstract

The Kullback-Leibler (KL) divergence is one of the most fundamental metrics in information theory and statistics and provides various operational interpretations in the context of mathematical communication theory and statistical hypothesis testing. The KL divergence for discrete distributions has the desired continuity property which leads to some fundamental results in universal hypothesis testing. With continuous observations, however, the KL divergence is only lower semi-continuous; difficulties arise when tackling universal hypothesis testing with continuous observations due to the lack of continuity in KL divergence.

This dissertation proposes a robust version of the KL divergence for continuous alphabets. Specifically, the KL divergence defined from a distribution to the Lévy ball centered at the other distribution is found to be continuous. This robust version of the KL divergence allows one to generalize the result in universal hypothesis testing for discrete alphabets to that for continuous observations. The optimal decision rule is developed whose robust property is provably established for universal hypothesis testing.

Another application of the robust KL divergence is in deviation detection: the problem of detecting deviation from a nominal distribution using a sequence of independent and identically distributed observations. An asymptotically δ-optimal detector is then developed for deviation detection where the Lévy metric becomes a very natural distance measure for deviation from the nominal distribution.

Lastly, the dissertation considers the following variation of a distributed detection

problem: a sensor may overhear other sensors' transmissions and thus may choose to refine its output in the hope of achieving a better detection performance. While this is shown to be possible for the fixed sample size test, asymptotically (in the number of samples) there is no performance gain, as measured by the KL divergence achievable at the fusion center, provided that the observations are conditionally independent. For conditionally dependent observations, however, asymptotic detection performance may indeed be improved when overhearing is utilized.

ROBUST KULLBACK-LEIBLER DIVERGENCE AND ITS APPLICATIONS IN UNIVERSAL HYPOTHESIS TESTING AND DEVIATION DETECTION

by

Pengfei Yang
B.S.(Statistics), University of Science and Technology of China, 2010

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical and Computer Engineering

Syracuse University
December 2016

ACKNOWLEDGEMENTS

My deepest gratitude goes first and foremost to my advisor Prof. Biao Chen for all the guidance and inspiration during my PhD study. It is only due to his patience and support that I am able to overcome the difficulties and finish this dissertation. I have learned a lot from his sharp insight on the problem, incredible carefulness and integrity, which will benefit me in my whole life. It was a great pleasure to be his student.

I would also like to thank Prof. Pramod K. Varshney, Prof. Pinyuan Chen, Prof. Yingbin Liang, Prof. M. Cenk Gursoy and Prof. Lixin Shen for taking the time to be my committee members and giving me helpful suggestions.

I am also grateful for my fellow labmates: Ge Xu, Wei Liu, Kapil Borle, Fangfang Zhu, Yu Zhao, Fangrong Peng, Shengyu Zhu, Yang Liu and Tiexing Wang for their helpful discussions in my PhD study. I also thank my friends for the good time we had together.

Last but not least, I am deeply indebted to my parents, my sister and my fiancée Mingxuan Tan for their unconditional support and love, who always stand by me through good and bad times, and encourage me to pursue my own dreams, which has led to this great journey of my PhD study.

The author would like to acknowledge the generous support from the National Science Foundation under Grant CCF1218289 and the Air Force Office of Scientific Research under Grants FA9550-10-1-0458 and FA9550-16-1-0077.

TABLE OF CONTENTS

Abstract **i**

Acknowledgements **vi**

List of Figures **ix**

1 Introduction **1**

1.1 Motivation and Problem Statement 1

1.2 Summary of Contributions 7

1.3 Organization 10

2 Robust Kullback-Leibler Divergence **11**

2.1 Preliminaries 11

2.2 Main Theorem 15

2.3 Proof of Theorem 2.2.1 18

2.4 Summary 32

3 Robust Universal Hypothesis Testing and Deviation Detection **33**

3.1 Framework and Criteria 33

3.2 An Overlooked Fact 35

3.3 Robust Universal Hypothesis Testing 39

3.3.1 Related Work 39

3.3.2 Some observations 41

3.3.3 Robust Universal Hypothesis Testing 43
3.4 Deviation Detection 49
3.4.1 Introduction 49
3.4.2 Problem Formulation and Solution 50
3.5 Summary 56

4 Computation and Estimation of the Robust KL Divergence 58
4.1 Computation of the Robust KL Divergence 58
4.2 Estimation of the Robust KL divergence 60
4.2.1 P_0 is known 60
4.2.2 P_0 is unknown 63
4.3 Summary 65

5 To Listen or Not: Distributed Detection with Asynchronous Transmissions 66
5.1 Introduction 66
5.2 Problem Statement 68
5.3 Distributed Detection with Asynchronous Transmissions 72
5.3.1 Conditionally independent observations 72
5.3.2 Conditionally dependent observations 74
5.4 Examples for fixed sample size test 75
5.4.1 A discrete example 75
5.4.2 A continuous example 77
5.5 Summary 78

6 Conclusion and Future Research 79
6.1 Conclusion 79
6.2 Future Work 82

References 85

LIST OF FIGURES

2.1 The Lévy ball centered at standard normal distribution with radius 0.045. . 14

2.2 The upper bound of $D(\mu||B_L(P_0, \delta_0))$ for different δ_0. 18

2.3 Illustration of $u_{-\delta}, u_\delta, u^\delta_{-\delta}$ and u^δ_δ 27

3.1 The shaded region is $B_L(\mu_n, 2\delta)$ and the solid line is P_0. 42

3.2 Lévy ball of the standard normal distribution and the step function 48

4.1 Examples of the estimate of the robust KL divergence. 62

5.1 Distributed network 68

5.2 Comparision betweent Q_{xy} and $\tilde{Q}_{xy}$ of conditionally dependent example. . 75

5.3 The ROC curves for the discrete example. 76

5.4 The ROC curves for the continuous example. 78

CHAPTER 1
INTRODUCTION

1.1 Motivation and Problem Statement

For distributions defined on discrete alphabets, the Kullback-Leibler (KL) divergence from distribution μ to another distribution P_0 is

$$D(\mu||P_0) = \sum_i \mu_i \log \frac{\mu_i}{P_i}. \tag{1.1}$$

For continuous distributions defined on the real line, the KL divergence between μ and P_0 is

$$D(\mu||P_0) = \int_{\mathcal{R}} d\mu \log \frac{d\mu}{dP_0}. \tag{1.2}$$

For either discrete or continuous alphabets, let $\mathcal{P}$ denote the corresponding probability space. In both discrete and continuous cases, the sublevel set and the superlevel set of the KL divergence to a fixed distribution P_0, are disjoint, i.e., for $0 < \eta_0 < \eta_1$,

$$\{\mu \in \mathcal{P} : D(\mu||P_0) \leq \eta_0\} \quad \text{and} \quad \{\mu \in \mathcal{P} : D(\mu||P_0) \geq \eta_1\} \quad \text{are disjoint.} \tag{1.3}$$

Two sets are separated if each is disjoint from the other's closure. The topology that specifies the "closure" is defined differently for the discrete and continuous cases. For the discrete case with finite, say, m elements, $\mathcal{P}$ is a compact subset of the m-dimensional Euclidean space, so $\mathcal{P}$ equipped with the Euclidean metric is a compact metric space. Assume P_0 is non zero at all m elements, then the KL divergence is continuous in the pair (μ, P_0). It is easy to see that the two sets in (1.3) are closed thus separated in the discrete case.

However, this is not true in the continuous case. For the continuous case, $\mathcal{P}$ equipped with the L$\acute{e}vy$ metric is a metric space, which is compatible with respect to the weak topology induced by weak convergence. Weak convergence can be defined in multiple equivalent ways, one is using the L$\acute{e}vy$ metric: a sequence of distributions weakly converges to a distribution if the corresponding L$\acute{e}vy$ distance converges to zero (c.f. Lemma 2.1.1).

In the continuous case, the KL divergence is lower semicontinuous, which is equivalent to stating that the sublevel set in (1.3) is closed. However, it is not upper semicontinuous thus the superlevel set is not closed in $\mathcal{P}$. P_0 belongs to the sublevel set, the following shows that P_0 is also a limit point of the superlevel set. Choose any distribution that is not absolutely continuous with respect to P_0, the linear combination of this distribution and P_0 is not absolutely continuous with respect to P_0 either, thus the KL divergence becomes infinity. But the L$\acute{e}vy$ metric between this combination and P_0 converges to 0 when the weight on P_0 goes to 1. This example takes advantage of distributions that are not absolutely continuous with respect to P_0. Actually, even when constrained to distributions that are absolutely continuous with respect to the fixed P_0, it can be shown that P_0 is still a limit point of the superlevel set. Therefore the two sets in (1.3) are not separated in the continuous case.

The above difference between discrete and continuous cases, in essence, stems from the difference in the continuity property: the KL divergence is continuous in the discrete case but not continuous with respect to weak convergence in the continuous case. As it turns out, such a fundamental difference in the KL divergence between the continuous and discrete

observations plays an important role in deviation detection, robust universal hypothesis testing, and KL divergence estimation to be defined in subsequent sections.

We start with the deviation detection problem. The normal state, i.e., the null hypothesis, is characterized by a proximity set, which consists of distributions that are close to a nominal distribution P_0; any *significant* departure to this nominal distribution constitutes the alternative hypothesis. Deviation detection has numerous engineering and societal applications. The classical quality control problem can often be formulated as a deviation detection problem. Normal operation leads to observation sequences (e.g., product measurements or other quantifiable metrics) that are expected to follow a nominal distribution P_0, which is known as it can be learned from past operations. Abnormality in the distribution of the output sequence is indication of the operation irregularity; yet the precise state of the anomaly, if it occurs, is often not a known *a priori* thus a sensible approach is to model the quality control problem as a deviation detection problem where the abnormal state can be any distribution significantly different from the nominal distribution.

Clearly, the first question in formulating the deviation detection problem is the choice of distance metrics.

Problem 1. *What is an appropriate metric that characterizes the distance between distributions in deviation detection?*

There are numerous metrics available to quantify the distance between distributions; some of them are not strict distance metrics in the sense that they do not satisfy the usual requirement for metrics, namely non-negativity, symmetry, and triangle inequality. Some of the widely used metrics include: Hellinger distance, Kullback-Leibler (KL) divergence, Kolmogorov metric, L$\acute{e}vy$ metric, Prokhorov metric, Separation distance, total variation distance, Wasserstein metric and χ^2 distance.

It is tempting to try the KL divergence given that this is probably the most widely used "distance metric" in the literature. With the KL divergence, the uncertainty sets of the deviation detection are just the above mentioned sublevel and superlevel sets in (1.3). In

the discrete case, the two sets are separated. The same statement holds for some other metrics, such as the total variation. Indeed, with discrete observations, almost all the above distance metrics can be used to define uncertainty set and it often comes down to which one is easy to work with or may lead to a detector that is simple to implement. However, with probability measures defined on the real line, the KL divergence, is not a suitable choice for defining the proximity set. This is true for some other metrics as well.

The deviation detection problem falls into the general framework of robust detection, which has been the subject of extensive studies since the seminal work of Huber and Strassen [1, 30]. In robust detection, the conditional probability distributions given *each* hypothesis are specified to belong to some uncertainty sets. The goal of robust detection is often to optimize the worst case performance over the uncertainty sets.

Naturally, the uncertainty sets of the general robust detection problems are disjoint. Otherwise the problem becomes degenerate, since the worst case would correspond to any distribution that belongs to both hypotheses. Furthermore, we need these two uncertainty sets to be separated, that is, the two sets can not be arbitrarily close to each other.

With finite alphabets, the uncertainty sets are usually characterized by continuous functions such as moments or above-mentioned metrics. For example, the two sets in (1.3) can be used to define the uncertainty sets, in this case "disjoint" usually implies "separated". In general, for the finite alphabets case, as long as the two uncertainty sets are disjoint, requiring two sets to be separated is redundant.

Such is not the case with continuous alphabets, and the following requirement is essential.

- The two uncertainty sets should be separated, in the sense that the closure of the respective sets should be non-overlapping. Here the closure is defined with respect to weak convergence of probability measures.

This separation requirement gives rise to difficulties in some well defined detection problems when dealing with distributions defined on the real line. One such example is the

deviation detection problem when the deviation is defined using either the KL divergence or the total variation; another example is the moment constrained detection problem.

We can draw more insight from an unexpected property of the KL divergence: for a fixed P_0, the closure of the KL divergence surface ($D(\mu||P_0) = \eta$) is the KL divergence ball ($D(\mu||P_0) \leq \eta$). Consequently, the closure of the superlevel set is the entire probability space. As such, defining the uncertainty sets using the KL divergence would encounter significant issues that render the deviation detection problem meaningless.

Recall that the separation requirement defined above is with respect to weak convergence. On the other hand, it is well known that weak convergence is equivalent to convergence in the Levy metric (again, c.f. Lemma 2.1.1). This leads to the next question.

Problem 2. *What is the relation between the Lévy metric and the KL divergence?*

Let us assume one can find the appropriate metric for the deviation detection problem. The next problem is to design the optimal detector. As mentioned earlier, deviation detection falls into the framework of the robust detection problem. In robust detection, to minimize the worst case performance over the uncertainty classes, the solution typically involves identifying a pair of least favorable distributions (LFDs), and subsequently designing a simple hypothesis test between the LFDs. In the Huber and Strassen framework, the proofs of existence of LFDs rely on the so-called joint stochastic boundedness property of the uncertainty classes. Such a property, however, does not hold for the deviation detection problem.

To facilitate the analysis, instead of solving the fixed sample size problem, we follow Hoeffding's approach in solving the universal hypothesis testing problem, which was first formulated in [2]. Hoeffding used the generalized Neyman-Pearson (NP) criterion, which evaluates the asymptotic efficiency by considering the error exponents instead of the error probabilities. The universal hypothesis testing problem is to decide whether an independent and identically distributed (i.i.d.) sequence of random variables has originated from a known distribution P_0 or another unknown distribution. This problem was treated in [30]

where the underlying alphabet is assumed to be discrete. The statistic of Hoeffding's detector is

$$D(\hat{\mu}_n||P_0), \tag{1.4}$$

where $\hat{\mu}_n$ is the empirical distribution. However, in the continuous case, one can not directly compare the KL divergence from an empirical distribution to a continuous distribution. Attempts to reconstruct a similar decision rule for continuous observations have been fruitless. Zeitouni and Gutman extended Hoeffding's work to continuous distributions [14] at the cost of a strictly weaker optimality. Zeitouni and Gutman's detector, as described in Chapter 3, is rather complicated, which leads to the following question.

Problem 3. *Why does not the detector (1.4) not work in the continuous case, and how to generalize and improve the detector proposed by Zeitouni and Gutman?*

One reason that detector (1.4) is not valid for continuous observations is that the complementary set of the KL divergence open ball is not closed with respect to weak convergence. However, Zeitouni and Gutman's treatment of continuous observations sheds light on a potential approach to circumvent the difficulty. Zeitouni and Gutman's detector first expands the empirical distribution to a L$\acute{e}$vy ball centered at the empirical distribution with radius δ, then compares the KL divergence from that L$\acute{e}$vy ball to P_0 and performs a $\delta-$smooth operation on the detector. This leads to the following conjecture.

Problem 4. *If there is uncertainty under null hypothesis and the uncertainty is defined using a L$\acute{e}$vy ball centered at P_0, will the generalized empirical likelihood ratio test be the optimal test?*

It turns out the above conjecture is equivalent to the following,

Problem 5. *Is the KL divergence from a distribution to a known L$\acute{e}$vy ball continuous with respect to weak convergence?*

This thesis will start with the answer of Problem 5. The result is subsequently applied to a slew of inference problems including robust universal detection, deviation detection and estimation of the KL divergence.

1.2 Summary of Contributions

For distributions with continuous alphabets, the KL divergence between two distributions is only lower semicontinuous and not upper semicontinuous with respect to weak convergence. These properties have consequences when applying KL divergence to a number of inference problems involving continuous observations. Examples include extending universal hypothesis testing currently developed for discrete alphabets to continuous alphabets and estimating the KL divergence between distributions with continuous alphabets. In this thesis, the classical KL divergence is generalized to that involving a distribution set defined using the L$\acute{e}$vy metric, which is compatible with the weak convergence of probability measures.

The KL divergence defined between two sets is the infimum of the KL divergence over the sets. In Chapter 2, we establish that the KL divergence from a distribution to the L$\acute{e}$vy ball of another known distribution is continuous with respect to weak convergence. We refer to this KL divergence as the robust KL divergence in this dissertation. Besides, the robust KL divergence is shown to be bounded and the supremum is represented as a function of the radius of the L$\acute{e}$vy ball due to the infimum operation in defining the robust KL divergence.

The intuition of the continuous property of the robust KL divergence is the following. The classical KL divergence is a function of two distributions, and its value may vary arbitrarily with small perturbation in one of the distributions with respect to the Levy metric. The reason is because L$\acute{e}$vy metric is strictly weaker than the KL divergence, i.e., convergence in the KL divergence necessarily implies convergence in the Lévy metric but not

the other way around. For the robust KL divergence, which compares one distribution to a L$\acute{e}$vy ball of another distribution, small perturbations in the first distribution can be tolerated by the L$\acute{e}$vy ball.

The following lists the important intermediate steps towards proving the continuity of the robust KL divergence. Some of the statements are themselves significant results in that they differ from that of the classical KL divergence.

- The robust KL divergence of discretized distributions will converge to the robust KL divergence of the original distributions as the number of quantization levels increases.
- The robust KL divergence is defined as the infimum over the L$\acute{e}$vy ball and the infimum can be attained by a distribution within the L$\acute{e}$vy ball.
- The robust KL divergence is continuous in the radius of the L$\acute{e}$vy ball.
- The robust KL divergence and the quantized robust KL divergence are convex functions.
- The supremum of the robust KL divergence over a L$\acute{e}$vy ball can be achieved by a distribution which is the combination of two distributions: the lower bound distribution and the upper bound distribution of the L$\acute{e}$vy ball. The freedom of such distribution is the combination point on the real line. Therefore, the problem of finding the supremum is reduced from an infinite dimension problem to a one dimension problem.
- The supremum of the robust KL divergence over a L$\acute{e}$vy ball converges to the robust KL divergence as the radius of the L$\acute{e}$vy ball diminishes. Therefore, the robust KL divergence is upper semicontinuous.
- The robust KL divergence is lower semicontinuous.

Chapter 2 proves the above properties and the continuous property of the robust KL divergence. The continuous property plays an important role in finding the optimal decision

rule for robust universal hypothesis testing and the deviation detection problem. Furthermore, it facilitates the estimation of the robust KL divergence from a sample sequence.

Chapter 3 examines the robust universal hypothesis testing problem, which is the generalization of the universal hypothesis testing problem to the robust setting. Zeitouni and Gutman successfully gave a $\delta-$optimal detector. However, the detector is hard to realize given the sample sequence.

In robust universal hypothesis testing, the nominal distribution P_0 is replaced by a L$\acute{e}$vy ball surrounding P_0, which is denoted as $\mathcal{P}_0$. The robust universal hypothesis testing problem is not sensitive to P_0: under the null hypothesis, the samples might come from a distribution that is very close to P_0, the optimal detector should be robust to such uncertainty. The reason to use the L$\acute{e}$vy metric among various distance metrics is that the L$\acute{e}$vy metric is the weakest [18], in other words, this L$\acute{e}$vy ball contains all distributions which are close enough to the nominal distribution using any other metrics. Although we generalize the universal hypothesis testing problem to the robust setting, we show that the generalized empirical likelihood ratio test is optimal. The test is intuitive and the test statistic is easy to compute.

In the deviation detection problem we first investigate which distance metrics are appropriate for characterizing the two uncertainty sets. By choosing the KL divergence or the total variation, the two uncertainty sets will be arbitrarily close to each other with respect to weak convergence. Thus we turn to the L$\acute{e}$vy metric, which can separate the probability space into two separated sets. Finally, we show that the generalized likelihood ratio test also applies to the deviation detection problem.

Chapter 4 investigates the computation and estimation of the robust KL divergence. Computing the optimal detector of the robust universal detection problem is shown to be equivalent to solving a convex optimization problem whose solution can be readily obtained via a standard convex program. Besides, the estimate of the robust KL divergence is then shown to converge almost surely.

In Chapter 5, we examine a variation of the canonical distributed detection system: a sensor may overhear other sensors' transmissions and thus may choose to refine its output in the hope of achieving a better detection performance. We show that while this is indeed possible for the fixed sample size test, asymptotically (in the number of samples) there is no performance gain, as measured by the KL divergence achievable at the fusion center, provided that the observations are conditionally independent. For conditionally dependent observations, however, we demonstrate that asymptotic detection performance may indeed be improved when overhearing is utilized.

1.3 Organization

The rest of the thesis is organized as follows. In Chapter 2, we prove that the robust KL divergence is continuous. In Chapter 3, we investigate the robust universal hypothesis testing and deviation detection problems. In Chapter 4, we compute the robust KL divergence of the empirical distribution via a convex optimization approach, and provided a procedure for estimating the KL divergence. Chapter 5 deals with the problem of distributed detection with asynchronous transmissions, using the KL divergence as an optimizing metric.

CHAPTER 2
ROBUST KULLBACK-LEIBLER DIVERGENCE

2.1 Preliminaries

The KL divergence was first introduced in [3]. It is often used as a metric to quantify the distance between two probability distributions. The KL divergence is also known as information divergence or relative entropy. For finite alphabets, the KL divergence between a probability distribution $\mu = (\mu_1, \mu_2, \cdots, \mu_n)$ and another distribution $P_0 = (p_1, p_2, \cdots, p_n)$ is

$$D(\mu||P_0) = \sum_{i=1}^{n} \mu_i \log \frac{\mu_i}{p_i}. \tag{2.1}$$

Notice that the order of the two distributions matters in the definition hence it is not a symmetric function of the two distributions thus is not a true distance metric. For distributions defined on the real line $\mathcal{R}$, the KL divergence between μ and P is defined as

$$D(\mu||P) = \int_{\mathcal{R}} d\mu \log \frac{d\mu}{dP}. \tag{2.2}$$

The KL divergence is one of the most fundamental metrics in information theory, statistics [4], machine learning and signal processing. For example, a special case of the KL divergence is the mutual information which has various operational interpretations in channel coding and data compression [5]. In hypothesis testing, the KL divergence controls the decay rates of error probabilities (e.g., see Stein's lemma [5] and Sanov's theorem [6]). In statistical machine learning, the KL divergence is used to extend machine learning algorithms to distributional features. The KL divergence can be employed as a similarity measure in image registration or multimedia classification [7–9]. In distributed signal processing, the KL divergences are frequently exploited as the metric to be optimized as a proxy to detection error probability which is otherwise intractable [10–12].

$D(\mu||P_0)$ is jointly convex and lower semi-continuous in the pair (μ, P_0) [13]. Furthermore, for the discrete case, $D(\mu||P_0)$ is continuous in P_0 and continuous in μ if $\min p_i > 0$. For finite alphabets, one can directly compute the KL divergence between an empirical distribution with another distribution, or between two empirical distributions. However, in the continuous case, given the samples and the empirical distribution $\hat{\mu}_n$, computing $D(\hat{\mu}_n||P_0)$ directly using the definition (2.2) is meaningless as one can get infinity since $\hat{\mu}_n$ and P_0 may have different support sets. In addition, the non-continuity of the KL divergence for continuous distributions often leads to difficulties in various inference problems. One example is universal hypotheses testing - while simple and intuitive results have been obtained for the discrete case, the attempt to generalize that to the continuous case has been laborious and the resulting detector is inexplicably complex to analyze or even implement. [14].

In this chapter, a novel property of the KL divergence is identified for distributions defined on the real line. Although for a fixed distribution P_0, $D(\mu||P_0)$ is not continuous in μ, the infimum of the KL divergence from μ to a L$\acute{e}$vy ball centered at P_0 is continuous in μ. This will be elaborated in Theorem 2.2.1. This continuity property plays an important role in solving the robust universal hypothesis testing problem and the deviation detection problem in Chapter 3.

Before we proceed to the main theorem, let us introduce some necessary definitions and notations. Denote the space of probability distributions on $(\mathcal{R}, \mathcal{F})$ as $\mathcal{P}$, where $\mathcal{R}$ is the real line and $\mathcal{F}$ is the sigma-algebra that contains all the Borel sets of $\mathcal{R}$. For $P \in \mathcal{P}$, $P(S)$ is defined for the set $S \in \mathcal{F}$. A clear and simple notation commonly used, is $P(t) := P((-\infty, t])$, since P and its corresponding cumulative distribution function (CDF) are equivalent, i.e., one is uniquely determined by the other [16].

The L$\acute{e}$vy metric d_L between distributions $F \in \mathcal{P}$ and $G \in \mathcal{P}$ is defined as follows,

$$d_L(F, G) := \inf\{\epsilon : F(x - \epsilon) - \epsilon \leq G(x) \leq F(x + \epsilon) + \epsilon, \forall x \in \mathcal{R}\}. \tag{2.3}$$

The L$\acute{e}$vy metric makes $(\mathcal{P}, d_L)$ a metric space [6], i.e., we have,

$$d_L(\mu, P) = 0 \quad \Leftrightarrow \quad \mu = P, \tag{2.4}$$

$$d_L(\mu, P) \quad = \quad d_L(P, \mu), \tag{2.5}$$

$$d_L(\mu, P) \quad \leq \quad d_L(\mu, Q) + d_L(Q, P). \tag{2.6}$$

The L$\acute{e}$vy ball centered at $P_0 \in \mathcal{P}$ with radius δ, is defined as

$$B_L(P_0, \delta) = \{P \in \mathcal{P} : d_L(P, P_0) \leq \delta\}. \tag{2.7}$$

Fig.2.1 plots the CDF of the standard normal distribution and its L$\acute{e}$vy ball with radius 0.045. A distribution falls inside the shaded area if and only if its distance to the standard normal distribution, as measured by the Levy metric d_L, is less than or equal to 0.045.

There are many equivalent ways to define the weak convergence, one is given in the following.

Definition 1. *(Weak convergence [15, 16]) For P_n, $P \in \mathcal{P}$, we say P_n weakly converges to P and write $P_n \xrightarrow{w} P$, if $P_n(x) \to P(x)$ for all x such that P is continuous at x.*

The L$\acute{e}$vy metric is strongly related to the concept of the weak convergence of proba-

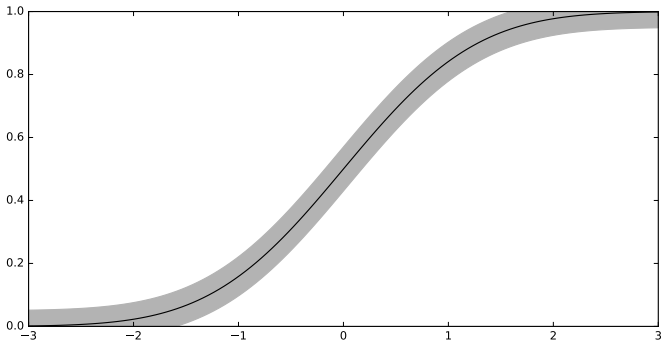


Fig. 2.1: The L$\acute{e}$vy ball centered at standard normal distribution with radius 0.045.

bility measures.

Lemma 2.1.1. *[15,16] For sequences in $\mathcal{P}$ whose limit is also in $\mathcal{P}$, the weak convergence and convergence in the d_L are equivalent, i.e., if $(P_n \in \mathcal{P})$ is a sequence in $\mathcal{P}$ and $P \in \mathcal{P}$, then $P_n \xrightarrow{w} P$ iff $d_L(P_n, P) \to 0$.*

The set of all partitions $\mathcal{A} = (A_1, \cdots, A_{|\mathcal{A}|})$ of $\mathcal{R}$ into a finite number of sets A_i is denoted by Π. Partition of $\mathcal{A}$ over $P \in \mathcal{P}$ is denoted as $P^{\mathcal{A}}$, which can be represented as a $|\mathcal{A}|$ dimensional vector $(P(A_1), P(A_2), \cdots, P(A_{|\mathcal{A}|})) \in \mathcal{R}^{|\mathcal{A}|}$. For convenience, for set $\Gamma \subseteq \mathcal{P}$, we define

$$\Gamma^{\mathcal{A}} := \{P^{\mathcal{A}} : P \in \Gamma\}.$$

Definition 2 (The KL divergence [17])**.** *The KL divergence between $P \in \mathcal{P}$ and $Q \in \mathcal{P}$ is defined as,*

$$D(P||Q) = \sup_{\mathcal{A} \in \Pi} D(P^{\mathcal{A}}||Q^{\mathcal{A}}), \tag{2.8}$$

where

$$D(P^{\mathcal{A}}||Q^{\mathcal{A}}) = \sum_{i=1}^{|\mathcal{A}|} P(A_i) \log \frac{P(A_i)}{Q(A_i)}. \tag{2.9}$$

The above definition is consistent with the classical definition using the Radon-Nikodym

derivative as in (2.2). For convenience, for sets $\Gamma_1, \Gamma_2 \subseteq \mathcal{P}$ and $S_1, S_2 \subseteq \mathcal{R}^n$, we will write

$$D(\Gamma_1||\Gamma_2) := \inf_{\gamma_1\in\Gamma_1,\gamma_2\in\Gamma_2} D(\gamma_1||\gamma_2), \quad D(S_1||S_2) = \inf_{\mathbf{x}\in S_1,\mathbf{y}\in S_2} D(\mathbf{x}||\mathbf{y}).$$

In addition, $\underline{\lim}$ and $\overline{\lim}$ denote $\liminf$ and $\limsup$, respectively.

2.2 Main Theorem

Theorem 2.2.1. *For a distribution $P_0 \in \mathcal{P}$, if $P_0(t)$ is continuous in t, then for any $\delta_0 > 0$, $D(\mu||B_L(P_0, \delta_0))$ is continuous in μ with respect to the weak convergence.*

The non trivial part is to show $D(\mu||B_L(P_0, \delta_0))$ is upper semicontinuous in μ, which is proved in Lemma 2.3.5. Lemma 2.3.6 proves $D(\mu||B_L(P_0, \delta_0))$ is lower semicontinuous in μ. Therefore, $D(\mu||B_L(P_0, \delta_0))$ is continuous in μ. The complete proof is lengthy and is included in Section 2.3. Important intermediate steps are summarized below.

- We first quantize the real line into a set of finite intervals, then the quantized robust KL divergence will converge to the original robust KL divergence as the quantization becomes finer. The proof is in essence proving that a max-min inequality is in fact an equality (Lemma 2.3.1).

- The robust KL divergence is defined as the infimum over the L$\acute{e}$vy ball and it is established that there exists a distribution inside the L$\acute{e}$vy ball that achieves the infimum (Lemma 2.3.1).

- The robust KL divergence is continuous in the radius of the L$\acute{e}$vy ball (Lemma 2.3.2).

- The robust KL divergence and the quantized robust KL divergence are convex functions of μ and $\mu^{\mathcal{A}}$, respectively (Lemma 2.3.3).

- The supremum of the robust KL divergence over a L$\acute{e}$vy ball can be achieved by distributions consisting of only two parts, one is the distribution corresponding to the

lower bound of the L$\acute{e}$vy ball, the other is the distribution corresponding to the upper bound of the L$\acute{e}$vy ball. Therefore, the problem of finding the supremum is reduced from an infinite dimension problem to a one dimension problem (Lemma 2.3.4).

- The supremum of the robust KL divergence over a L$\acute{e}$vy ball converges to the robust KL divergence as the L$\acute{e}$vy ball diminishes, i.e., its radius goes to 0. Therefore, the robust KL divergence is upper semicontinuous (Lemma 2.3.5).

- The robust KL divergence is lower semicontinuous (Lemma 2.3.6).

The continuity property in Theorem 2.2.1 does not hold if the ball is constructed using other measures such as the total variation or the KL divergence. The total variation is defined as follows.

Definition 3. *[18] The total variation between $P \in \mathcal{P}$ and $Q \in \mathcal{P}$ is*

$$d_{TV}(P, Q) := \sup_{S \in \mathcal{F}} |P(S) - Q(S)|.$$

Assume $P_0(t)$ is continuous in t, denote by $B_{TV}(P_0, \delta_0)$ and $B_{KL}(P_0, \delta_0)$ distribution balls defined by the total variation and the KL divergence, respectively, in a manner similar to that of the Levy ball in (2.7). We will show that there exists a sequence $P_n \xrightarrow{w} P_0$, while neither $D(P_n||B_{TV}(P_0, \delta_0))$ nor $D(P_n||B_{KL}(P_0, \delta_0))$ converges to the desired limit, i.e., $D(P_0||B_{TV}(P_0, \delta_0))$ and $D(P_0||B_{KL}(P_0, \delta_0))$. For any $n > 0$, we can always choose a $P_n \in B_L(P_0, 1/n)$ such that $P_n(t)$ is a step function. Let $S_n := \{x \in \mathcal{R} : P_n(x) - P_n(x-) > 0\}$, S_n is the set of all jump points of $P_n(t)$.

- For the total variation case, from the data processing inequality,

$$\begin{aligned} D(P_n||B_{TV}(P_0,\delta_0)) &= \inf_{\{P\in B_{TV}(P_0,\delta_0)\}} D(P_n||P) \\ &\geq \inf_{\{P\in B_{TV}(P_0,\delta_0)\}} P_n(S_n)\log\frac{P_n(S_n)}{P(S_n)} + P_n(S_n^c)\log\frac{P_n(S_n^c)}{P(S_n^c)} \\ &= \inf_{\{P\in B_{TV}(P_0,\delta_0)\}} 1\log\frac{1}{P(S_n)} + 0\log\frac{0}{P(S_n^c)} \\ &\geq 1\log\frac{1}{P_0(S_n)+\delta_0} \\ &= \log\frac{1}{\delta_0}, \end{aligned}$$

then

$$\begin{aligned} \underline{\lim}_{n\to\infty} D(P_n||B_{TV}(P_0,\delta_0)) &\geq \log\frac{1}{\delta_0} \\ &> 0 \\ &= D(P_0||B_{TV}(P_0,\delta_0)). \end{aligned}$$

Thus $D(P_n||B_{TV}(P_0,\delta_0)) \nrightarrow D(P_0||B_{TV}(P_0,\delta_0))$ even though $P_n \xrightarrow{w} P_0$.

- As for the KL divergence case, for any P such that $D(P||P_0) \leq \delta_0$, $D(P_n||P) = \infty$, therefore $D(P_n||B_{KL}(P_0,\delta_0)) \nrightarrow D(P_0||B_{KL}(P_0,\delta_0))$.

The assumption that $P_0(t)$ is continuous in t is also necessary, otherwise the continuous property of the robust KL divergence does not necessarily hold. We construct the following example to illustrate this point. Let P_0 be the distribution that $P_0(t) = 0$ for $t < 0$ and $P_0(t) = 1$ for $t \geq 0$, i.e., it is a degenerate random variable that equals to 0 with probability 1. Let μ_i be the distribution that $\mu_i(t) = 0$ for $t < 0.5 + \frac{1}{i}$ and $\mu_i(t) = 1$ for $t \geq 0.5 + \frac{1}{i}$. Then $\mu_i \xrightarrow{w} \mu$ as $i \to \infty$, where $\mu(t) = 0$ for $t < 0.5$ and $\mu(t) = 1$ for $t \geq 0.5$. However, we can see that, as $\mu_i \xrightarrow{w} \mu$,

$$\lim_{i\to\infty} D(\mu_i||B_L(P_0,0.5)) = \lim_{i\to\infty} \log\frac{1}{0.5} = \log 2 \neq D(\mu||B_L(P_0,0.5)) = 0. \qquad (2.10)$$

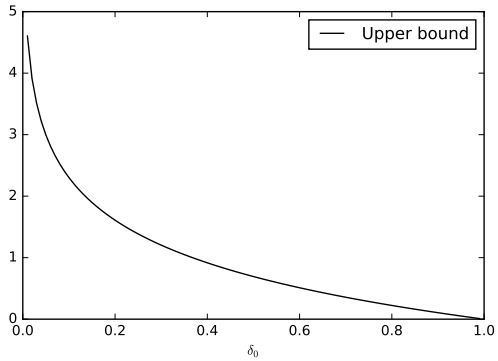


Fig. 2.2: The upper bound of $D(\mu||B_L(P_0, \delta_0))$ for different δ_0.

$D(\mu||P_0)$ is unbounded, but $D(\mu||B_L(P_0, \delta_0))$ is bounded. The upper bound is given in the following and its proof can be found in Proposition 1 in the next section.

$$\sup_{\mu, P_0 \in \mathcal{P}} D(\mu||B_L(P_0, \delta_0)) = \log \frac{1}{\delta_0}.$$

Fig. 2.2 shows how δ_0 controls the upper bound of $D(\mu||B_L(P_0, \delta_0))$.

Theorem 2.2.1 sheds some light on the dynamics of the KL divergence of continuous distributions. Also, this continuity property provides a convenient approach for solving the robust version of the universal hypothesis testing problem for the continuous case. This will be elaborated in Chapter 3. Chapter 4 explores the estimation of the robust KL divergence, again, by utilizing the continuity of the robust KL divergence.

2.3 Proof of Theorem 2.2.1

The following lemma generalizes (2.8) for the classical KL divergence to the robust KL divergence.

Lemma 2.3.1. *For $\mu, P_0 \in \mathcal{P}$ and $\delta_0 > 0$, $D(\mu||B_L(P_0, \delta_0)) = \sup_{\mathcal{A} \in \Pi} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0))$.*

Proof. $\mathcal{P}$ is not compact with respect to the weak convergence, a consequence of the countably additive property in the axiomatic definition of probability [16]. Let $\mathcal{M}$ denote the

space of *finitely* additive and non-negative set functions on $(\mathcal{R}, \mathcal{F})$ with $M(\mathcal{R}) = 1$ for $M \in \mathcal{M}$. Thus, we relax the countable additivity to finite additivity and as a consequence, $\mathcal{P} \subseteq \mathcal{M}$ and $\mathcal{M}$ is compact with respect to the weak convergence [17]. Similarly, we define $M(t) := M((-\infty, t])$ and $M(t)$ is a right continuous non-decreasing function on $\mathcal{R}$. As with $P(t)$ and $P \in \mathcal{P}$, $M(t)$ and $M \in \mathcal{M}$ are equivalent since one is uniquely determined by the other.

$\mathcal{P}$ is equivalent to the set of right continuous non-decreasing functions on $\mathcal{R}$ with $P(-\infty) = 0$ and $P(\infty) = 1$ if $P \in \mathcal{P}$ [16], while M is equivalent to the set of right continuous non-decreasing functions on $\mathcal{R}$ with $M(-\infty) \geq 0$ and $M(\infty) \leq 1$ if $M \in \mathcal{M}$. The L*é*vy metric d_L and the KL divergence extend unchanged to $F \in \mathcal{M}$ and $G \in \mathcal{M}$ [16, 17]. The following three steps constitute the proof of the lemma,

$$
\begin{aligned}
D(\mu||B_L(P_0, \delta_0)) &= D(\mu||\bar{B}_L(P_0, \delta_0)), && (2.11)\\
D(\mu||\bar{B}_L(P_0, \delta_0)) &= \sup_{\mathcal{A} \in \Pi} D(\mu^{\mathcal{A}}||\bar{B}_L^{\mathcal{A}}(P_0, \delta_0)), && (2.12)\\
\sup_{\mathcal{A} \in \Pi} D(\mu^{\mathcal{A}}||\bar{B}_L^{\mathcal{A}}(P_0, \delta_0)) &= \sup_{\mathcal{A} \in \Pi} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0)), && (2.13)
\end{aligned}
$$

where $\bar{B}_L(P_0, \delta_0) := \{P \in \mathcal{M} : d_L(P, P_0) \leq \delta_0\}$. We now prove (2.11)-(2.13).

- $D(\mu||\bar{B}_L(P_0, \delta_0)) = \inf_{\{P \in \bar{B}_L(P_0, \delta_0)\}} D(\mu||P)$, $\bar{B}_L(P_0, \delta_0)$ is closed with respect to the weak convergence, thus is compact since $\mathcal{M}$ is compact. To prove (2.11), let

$$P_\mu := \arg \inf_{\{P \in \bar{B}_L(P_0, \delta_0)\}} D(\mu||P),$$

the existence of P_μ is guaranteed since $D(\mu||P)$ is lower semicontinuous and lower semicontinuous function attains its infimum on a compact set. Assume $P_\mu \in \mathcal{M} \setminus \mathcal{P}$, then there exists a $\delta > 0$ such that $P_\mu(-\infty) \geq \delta$ or $P_\mu(+\infty) \leq 1 - \delta$. Without loss of generality, we can assume $P_\mu(-\infty) = \delta$ and $P_\mu(+\infty) = 1$, as other cases can be proved similarly. Let s denote the minimum t such that $P_0(t - \delta_0) \geq \delta_0$, we construct

P'_μ as follows.

– If $P_\mu(s) = \delta$, let

$$P'_\mu(t) = \begin{cases} 0 & \text{if } t < s, \\ P_\mu(t) & \text{if } t \geq s. \end{cases}$$

Since $\inf_t P'_\mu(t) = 0$ and $\sup_t P'_\mu(t) = 1$, $P'_\mu \in \mathcal{P}$. In addition, it can be easily verified that $d_L(P'_\mu, P_0) \leq \delta_0$. Therefore, $P'_\mu \in B_L(P_0, \delta_0)$ and $D(\mu||P'_\mu) = D(\mu||P_\mu)$.

– If $P_\mu(s) > \delta$, let

$$P'_\mu(t) = \begin{cases} \frac{(P_\mu(t)-\delta)P_\mu(s)}{P_\mu(s)-\delta} & \text{if } t < s, \\ P_\mu(t) & \text{if } t \geq s. \end{cases}$$

$\inf_t P'_\mu(t) = 0$ and $\sup_t P'_\mu(t) = 1$ thus $P'_\mu \in \mathcal{P}$. For $t < s$,

$$\frac{(P_\mu(t) - \delta)P_\mu(s)}{P_\mu(s) - \delta} \leq P_\mu(t) \Leftrightarrow P_\mu(t) \leq P_\mu(s),$$

then,

$$P_0(t - \delta_0) - \delta_0 < 0 \leq P'_\mu(t) \leq P_\mu(t) \leq P_0(t + \delta_0) + \delta_0 \Longrightarrow d_L(P'_\mu, P_0) \leq \delta_0.$$

Therefore, we have $P'_\mu \in B_L(P_0, \delta_0)$. Also P'_μ achieves the infimum since,

$$\begin{aligned} D(\mu||P'_\mu) &= \int_{-\infty}^{s-} d\mu(t) \log \frac{d\mu(t)}{dP'_\mu(t)} + \int_s^{\infty} d\mu(t) \log \frac{d\mu(t)}{dP'_\mu(t)} \\ &= \mu(s-) \log \frac{P_\mu(s) - \delta}{P_\mu(s)} + \int_{-\infty}^{s-} d\mu(t) \log \frac{d\mu(t)}{dP_\mu(t)} + \int_s^{\infty} d\mu(t) \log \frac{d\mu(t)}{dP_\mu(t)} \\ &= \mu(s-) \log \frac{P_\mu(s) - \delta}{P_\mu(s)} + D(\mu||P_\mu) \\ &\leq D(\mu||P_\mu). \end{aligned}$$

Therefore in either case,

$$\exists P'_\mu \in B_L(P_0, \delta_0) \quad \text{s.t.} \quad P'_\mu \quad \text{achieves} \quad \inf_{\{P \in \bar{B}_L(P_0, \delta_0)\}} D(\mu||P). \tag{2.14}$$

- Lemma 2.4 in [15] shows $D(B_L(P_0, \delta_0)||\mu) = \sup_{\mathcal{A} \in \Pi} D(B_L^{\mathcal{A}}(P_0, \delta_0)||\mu^{\mathcal{A}})$, using a parallel proof, one can show $D(\mu||\bar{B}_L(P_0, \delta_0)) = \sup_{\mathcal{A} \in \Pi} D(\mu^{\mathcal{A}}||\bar{B}_L^{\mathcal{A}}(P_0, \delta_0))$, i.e., (2.12) holds.
- For any $\mathcal{A} \in \Pi$, $\bar{B}_L^{\mathcal{A}}(P_0, \delta_0) = B_L^{\mathcal{A}}(P_0, \delta_0)$, therefore (2.13) holds.

□

Lemma 2.3.2. *Given $\mu, P_0 \in \mathcal{P}$ and $\delta_0 > 0$, if $P_0(t)$ is continuous in t, then $D(\mu||B_L(P_0, \delta_0))$ is continuous in δ_0.*

Proof. Let $\delta \in (0, \delta_0)$. $D(\mu||B_L(P_0, \delta_0))$ is left continuity in δ_0 if,

$$D(\mu||\{P \in \mathcal{P} : d_L(P, P_0) < \delta_0\}) = D(\mu||B_L(P_0, \delta_0)). \tag{2.15}$$

It is easy to see that $D(\mu||\{P \in \mathcal{P} : d_L(P, P_0) < \delta_0\}) \geq D(\mu||B_L(P_0, \delta_0))$. So we only need to show the other direction,

$$D(\mu||\{P \in \mathcal{P} : d_L(P, P_0) < \delta_0\}) \leq D(\mu||B_L(P_0, \delta_0)). \tag{2.16}$$

Denote

$$P_\delta = \arg \inf_{\{P \in B_L(P_0, \delta)\}} D(\mu||P), \quad P_{\delta_0} = \arg \inf_{\{P \in B_L(P_0, \delta_0)\}} D(\mu||P).$$

The existence of P_δ and P_{δ_0} is guaranteed by (2.14). For any $0 < \lambda < 1$,

$$d_L(\lambda P_\delta + (1 - \lambda) P_{\delta_0}, P_0) < \delta_0,$$

which may not hold if $P_0(t)$ is not continuous in t. Then,

$$\begin{aligned} D(\mu||\{P \in \mathcal{P} : d_L(P, P_0) < \delta_0\}) &\leq \lim_{\lambda \to 0+} D(\mu||\lambda P_\delta + (1-\lambda)P_{\delta_0}) && (2.17)\\ &\leq \lim_{\lambda \to 0+} \lambda D(\mu||P_\delta) + (1-\lambda)D(\mu||P_{\delta_0}) && (2.18)\\ &= D(\mu||P_{\delta_0}) && (2.19)\\ &= D(\mu||B_L(P_0, \delta_0)). && (2.20) \end{aligned}$$

Therefore $D(\mu||B_L(P_0, \delta_0))$ is left continuous in δ_0.

The rest is to show $D(\mu||B_L(P_0, \delta_0))$ is right continuous in δ_0. Since $D(\mu||B_L(P_0, \delta_0))$ is decreasing in δ_0, we only need to show:

$$\lim_{n\to\infty} D\left(\mu||B_L(P_0, \delta_0 + \frac{1}{n})\right) \geq D(\mu||B_L(P_0, \delta_0)). \tag{2.21}$$

From (2.14), there exists $P_n \in B_L(P_0, \delta_0 + \frac{1}{n})$ such that $D(\mu||P_n) = D\left(\mu||B_L(P_0, \delta_0 + \frac{1}{n})\right)$. $\mathcal{M}$ is compact, P_n converges to $P^* \in \mathcal{M}$. Since $P^* \in \bar{B}_L(P_0, \delta_0 + \frac{1}{n})$ for any n, $P^* \in \bar{B}_L(P_0, \delta_0)$. We have,

$$\begin{aligned} \lim_{n\to\infty} D\left(\mu||B_L(P_0, \delta_0 + \frac{1}{n})\right) &= \lim_{n\to\infty} D(\mu||P_n) && (2.22)\\ &\geq D(\mu||P^*) && (2.23)\\ &\geq D(\mu||\bar{B}_L(P_0, \delta_0)) && (2.24)\\ &= D(\mu||B_L(P_0, \delta_0)), && (2.25) \end{aligned}$$

where (2.23) comes from the fact that the KL divergence is lower semicontinuous and the last equality was proved in (2.11). Therefore $D(\mu||B_L(P_0, \delta_0))$ is right continuous in δ_0. □

Lemma 2.3.3. *For $\mu, P_0 \in \mathcal{P}$ and $\delta_0 > 0$, $D(\mu||B_L(P_0, \delta_0))$ is a convex function of μ. Also, for any partition $\mathcal{A}$, $D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0))$ is convex in $\mu^{\mathcal{A}}$.*

Proof. Let $P_i = \arg\inf_{\{P \in B_L(P_0,\delta_0)\}} D(\mu_i||P)$ for $i = 1, 2$. For any $0 < \lambda < 1$, $\lambda P_1 + (1-\lambda)P_2 \in B_L(P_0, \delta_0)$, thus,

$$\begin{aligned} D(\lambda\mu_1 + (1-\lambda)\mu_2||B_L(P_0,\delta_0)) &\leq D(\lambda\mu_1 + (1-\lambda)\mu_2||\lambda P_1 + (1-\lambda)P_2) \\ &\leq \lambda D(\mu_1||P_1) + (1-\lambda)D(\mu_2||P_2) \\ &= \lambda D(\mu_1||B_L(P_0,\delta_0)) + (1-\lambda)D(\mu_2||B_L(P_0,\delta_0)). \end{aligned}$$

Therefore, $D(\mu||B_L(P_0,\delta_0))$ is a convex function of μ. That $D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0,\delta_0))$ is convex in $\mu^{\mathcal{A}}$ follows a similar argument. □

Lemma 2.3.4. *Given $\mu_0, P_0 \in \mathcal{P}$ and $\delta, \delta_0 > 0$, we have*

$$\sup_{\mu \in B_L(\mu_0,\delta)} D(\mu||B_L(P_0,\delta_0)) = \sup_{x \in \mathcal{R}} D(\mu_x^\delta||B_L(P_0,\delta_0)),$$

where

$$\mu_x^\delta(t) = \begin{cases} \max(0, \mu_0(t-\delta) - \delta)) & \text{if } t < x, \\ \min(1, \mu_0(t+\delta) + \delta)) & \text{if } t \geq x. \end{cases} \tag{2.26}$$

Proof. We have

$$\begin{aligned} \sup_{\mu \in B_L(\mu_0,\delta)} D(\mu||B_L(P_0,\delta_0)) &= \sup_{\mu \in B_L(\mu_0,\delta)} \sup_{\mathcal{A} \in \Pi} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0,\delta_0)) && (2.27) \\ &= \sup_{\mathcal{A} \in \Pi} \sup_{\mu \in B_L(\mu_0,\delta)} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0,\delta_0)) && (2.28) \\ &= \sup_{\mathcal{A} \in \Pi} \sup_{\mu^{\mathcal{A}} \in B_L^{\mathcal{A}}(\mu_0,\delta)} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0,\delta_0)). && (2.29) \end{aligned}$$

Equality (2.27) comes from Lemma 2.3.1. Fix a partition $\mathcal{A}$, without loss of generality we can assume $|\mathcal{A}| = n$ and $\mathcal{A} = \{(-\infty, a_1], (a_1, a_2], \cdots, (a_{n-2}, a_{n-1}], (a_{n-1}, \infty)\}$. The

partition $\mathcal{A}$ over the probability space $\mathcal{P}$ can be represented as an n-dimensional polytope,

$$\mathcal{P}^{\mathcal{A}} = \{(x_1, x_2, \cdots, x_n) \in \mathcal{R}^n : \sum_i x_i = 1 \text{ and } \forall i, 0 \leq x_i \leq 1\}. \tag{2.30}$$

Similarly, the partition $\mathcal{A}$ over the set $B_L(\mu_0, \delta)$ is also an n-dimensional polytope inside $\mathcal{P}^{\mathcal{A}}$,

$$B_L^{\mathcal{A}}(\mu_0, \delta) = \{(x_1, x_2, \cdots, x_n) \in \mathcal{P}^{\mathcal{A}} : \forall 1 \leq j \leq n-1, L_j \leq \sum_{i=1}^{j} x_i \leq U_j\}, \tag{2.31}$$

where $L_j = \max(0, \mu_0(a_j - \delta) - \delta), U_j = \min(1, \mu_0(a_j + \delta) + \delta)$. We can assume for any $1 \leq j \leq n-2$, $U_j > L_{j+1}$, otherwise we can make $\mathcal{A}$ finer such that the new partition (denoted as $\mathcal{A}$ again) has the property that $a_{j+1} \leq a_j + \delta$ for $1 \leq j \leq n-2$. It can be verified that for each $1 \leq j \leq n-2, U_j > L_{j+1}$. The reason that such an $\mathcal{A}$ can be finite is that $\mu_0(t)$ is a bounded non-decreasing function.

A vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is a vertex of $B_L^{\mathcal{A}}(\mu_0, \delta)$ if and only if $\sum_{i=1}^{j} x_i$ equals L_j or U_j for any $1 \leq j \leq n-1$, $\sum_{i=1}^{n} x_i = 1$ and $0 \leq x_i \leq 1$. Since $U_j > L_{j+1}$ for any $1 \leq j \leq n-2$, for a vertex $\mathbf{x}$, once $\sum_{i=1}^{j} x_i = U_j$ for some j, then for any $k > j$ we have $\sum_{i=1}^{k} x_i = U_k$.

Therefore there are n vertices $\mathbf{x}^1, \cdots, \mathbf{x}^n$ of $B_L^{\mathcal{A}}(\mu_0, \delta)$ that satisfy the property that $\sum_{i=1}^{j} x_i^k = L_j$ for $j < k$, $\sum_{i=1}^{j} x_i^k = U_j$ for $j \geq k$. Or equivalently, if we denote $L_0 = 0$ and $U_n = 1$, for $1 \leq k \leq n$,

$$x_i^k = \begin{cases} L_i - L_{i-1} & \text{if } i < k, \\ U_i - L_{i-1} & \text{if } i = k, \\ U_i - U_{i-1} & \text{if } i > k. \end{cases}$$

From Lemma 2.3.3, $D(\cdot || B_L^{\mathcal{A}}(P_0, \delta_0))$ is a convex function, thus the supremum on the poly-

tope $B_L^{\mathcal{A}}(\mu_0, \delta)$ is achieved at its vertices. Let

$$\mu_x^\delta(t) = \begin{cases} \max(0, \mu_0(t-\delta) - \delta)) & \text{if } t < x, \\ \min(1, \mu_0(t+\delta) + \delta)) & \text{if } t \geq x. \end{cases} \tag{2.32}$$

Then any $\mathbf{x}^k$ is a quantization of μ_x^δ over the partition $\mathcal{A}$ for some x.

$$\begin{aligned} \sup_{\mu^{\mathcal{A}} \in B_L^{\mathcal{A}}(\mu_0,\delta)} D(\mu^{\mathcal{A}} || B_L^{\mathcal{A}}(P_0, \delta_0)) &= \max_k D(\mathbf{x}^k || B_L^{\mathcal{A}}(P_0, \delta_0)) && (2.33) \\ &\leq \sup_x D((\mu_x^\delta)^{\mathcal{A}} || B_L^{\mathcal{A}}(P_0, \delta_0)) && (2.34) \\ &\leq \sup_{\mathcal{A} \in \Pi} \sup_{x \in \mathcal{R}} D((\mu_x^\delta)^{\mathcal{A}} || B_L^{\mathcal{A}}(P_0, \delta_0)), && (2.35) \\ &= \sup_{x \in \mathcal{R}} \sup_{\mathcal{A} \in \Pi} D((\mu_x^\delta)^{\mathcal{A}} || B_L^{\mathcal{A}}(P_0, \delta_0)), && (2.36) \\ &= \sup_{x \in \mathcal{R}} D(\mu_x^\delta || B_L(P_0, \delta_0)), && (2.37) \end{aligned}$$

the last equality comes from Lemma 2.3.1. From (2.29) and (2.37), we have

$$\sup_{\mu \in B_L(\mu_0,\delta)} D(\mu || B_L(P_0, \delta_0)) \leq \sup_{x \in \mathcal{R}} D(\mu_x^\delta || B_L(P_0, \delta_0)).$$

For the other direction, since $\mu_x^\delta \in B_L(\mu_0, \delta)$,

$$\sup_{\mu \in B_L(\mu_0,\delta)} D(\mu || B_L(P_0, \delta_0)) \geq \sup_{x \in \mathcal{R}} D(\mu_x^\delta || B_L(P_0, \delta_0)).$$

Therefore, we have,

$$\sup_{\mu \in B_L(\mu_0,\delta)} D(\mu || B_L(P_0, \delta_0)) = \sup_{x \in \mathcal{R}} D(\mu_x^\delta || B_L(P_0, \delta_0)).$$

□

Proposition 1. $\sup_{\mu, P_0 \in \mathcal{P}} D(\mu || B_L(P_0, \delta_0) = \log \frac{1}{\delta_0}$.

Proof. We construct a distribution $S_0 \in \mathcal{P}$ such that $S_0(t) = 0$ for $t < 0$, and $S_0(t) = 1$ for

$t \geq 0$, then $\mathcal{P} = B_L(S_0, 1)$ since d_L is bounded by 1. According to Lemma 2.3.4,

$$\sup_{\mu \in B_L(S_0,1)} D(\mu||B_L(P_0, \delta_0)) = \sup_{x \in \mathcal{R}} D(\mu_x^\delta||B_L(P_0, \delta_0)), \tag{2.38}$$

where $\mu_x^1(t) = 0$ for $t < x$, and $\mu_x^1(t) = 1$ for $t \geq x$, then,

$$\begin{aligned}
& \sup_{x \in \mathcal{R}} D(\mu_x^\delta||B_L(P_0, \delta_0)) \\
= & \sup_{x \in \mathcal{R}} \log \frac{1}{\min(1, P_0(x + \delta_0) + \delta_0) - \max(0, P_0(x - \delta_0) - \delta_0)} && (2.39) \\
= & \log \frac{1}{\inf_{x \in \mathcal{R}}(\min(1, P_0(x + \delta_0) + \delta_0) - \max(0, P_0(x - \delta_0) - \delta_0))} && (2.40) \\
= & \log \frac{1}{\delta_0}, && (2.41)
\end{aligned}$$

the last equality comes from the fact that

$$\min(1, P_0(t + \delta_0) + \delta_0) - \max(0, P_0(t - \delta_0) - \delta_0) \geq \delta_0$$

and

$$\lim_{x \to \infty} \min(1, P_0(x + \delta_0) + \delta_0) - \max(0, P_0(x - \delta_0) - \delta_0) = \delta_0,$$

which means a finitely additive measure belongs to $\mathcal{M} \setminus \mathcal{P}$ can always achieves the supremum for any P_0. □

Lemma 2.3.5. *Given $P_0 \in \mathcal{P}$ and $\delta_0 > 0$, if $P_0(t)$ is continuous in t, then $D(\mu||B_L(P_0, \delta_0))$ is upper semicontinuous in μ with respect to the weak convergence.*

Proof. For any fixed $\mu_0 \in \mathcal{P}$, the statement is equivalent to show that when $\delta \to 0$,

$$\lim_{\delta \to 0} \sup_{\mu \in B_L(\mu_0, \delta)} D(\mu||B_L(P_0, \delta_0)) \leq D(\mu_0||B_L(P_0, \delta_0)). \tag{2.42}$$

From Lemma 2.3.4, it is equivalent to show

$$\limsup_{\delta\to 0}\sup_{x\in\mathcal{R}} D(\mu_x^\delta||B_L(P_0,\delta_0)) \le D(\mu_0||B_L(P_0,\delta_0)).$$

Denote $u_{-\delta}$ as the left boundary of support set $\mathcal{S}(\mu(t+\delta))$ and $u_{-\delta}^\delta := \arg\inf_x \mu(x+\delta) = 1-\delta$, denote u_δ as the left boundary of $\mathcal{S}(\mu(t-\delta))$ and $u_\delta^\delta := \arg\inf_x \mu(x-\delta) = 1$. Fig.2.3 plots these locations. We will prove

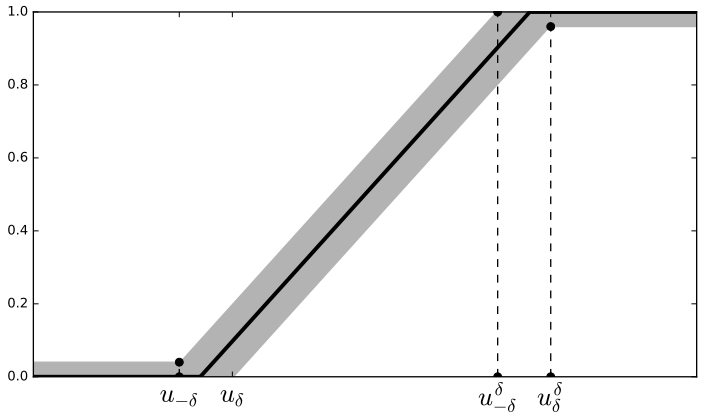


Fig. 2.3: Illustration of $u_{-\delta}, u_\delta, u_{-\delta}^\delta$ and u_δ^δ. The solid line represents μ_0 and shaded region represents $B_L(\mu_0,\delta)$.

$$\lim_{\delta\to 0}\sup_{\mu\in B_L(\mu_0,\delta)} D(\mu||B_L(P_0,\delta_0)) \le D(\mu_0||B_L(P_0,\delta_0-\delta_1))$$

for any $\delta_1 > 0$. Now fix δ_1, we will prove $D(\mu_x^\delta||B_L(P_0,\delta_0))$ can be uniformly bounded as x varies. Denote

$$P_{\delta_0-\delta_1} := \arg\inf_{\{P\in B_L(P_0,\delta_0-\delta_1)\}} D(\mu_0||P).$$

For fixed $\delta < \delta_1$, let

$$P_{\delta_0-\delta_1}^{\delta,u}(t) = (1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1, \;\; P_{\delta_0-\delta_1}^{\delta,l}(t) = (1-\delta_1)P_{\delta_0-\delta_1}(t-\delta).$$

To get $P_{\delta_0-\delta_1}^{\delta,u}(t)$, we first shift $P_{\delta_0-\delta_1}(t)$ to the left by δ, then scale it by $(1-\delta_1)$ and shift

it up by δ_1; similarly to get $P^{\delta,l}_{\delta_0-\delta_1}(t)$, we shift $P_{\delta_0-\delta_1}(t)$ to the right by δ, then scale it by $(1-\delta_1)$. Clearly

$$d_L(P^{\delta,u}_{\delta_0-\delta_1}, P_{\delta_0-\delta_1}) \leq \delta_1, \quad d_L(P^{\delta,l}_{\delta_0-\delta_1}, P_{\delta_0-\delta_1}) \leq \delta_1. \tag{2.43}$$

For any x, construct $P^x_{\delta_0-\delta_1}$ in a similar way as μ^δ_x,

$$P^x_{\delta_0-\delta_1}(t) = \begin{cases} P^{\delta,l}_{\delta_0-\delta_1}(t) & \text{if } t < x, \\ P^{\delta,u}_{\delta_0-\delta_1}(t) & \text{if } t \geq x. \end{cases} \tag{2.44}$$

$P^x_{\delta_0-\delta_1} \in B_L(P_0, \delta_0)$ since

$$\begin{aligned} d_L(P^x_{\delta_0-\delta_1}, P_0) &\leq d_L(P^x_{\delta_0-\delta_1}, P_{\delta_0-\delta_1}) + d_L(P_{\delta_0-\delta_1}, P_0) && (2.45) \\ &\leq \delta_1 + (\delta_0 - \delta_1) = \delta_0, && (2.46) \end{aligned}$$

where the first inequality holds because $(\mathcal{P}, d_L)$ is a metric space, d_L satisfies the triangle inequality, the second inequality comes from (2.43) and the definition of $P_{\delta_0-\delta_1}$.

From Proposition 1 $D(\mu_0||P_{\delta_0-\delta_1}) = D(\mu_0||B_L(P_0, \delta_0 - \delta_1)) < \infty$, therefore μ_0 is absolutely continuous with respect to $P_{\delta_0-\delta_1}$. From the construction of μ^δ_x and $P^x_{\delta_0-\delta_1}$, we can see that μ^δ_x is absolutely continuous with respect to $P^x_{\delta_0-\delta_1}$ as well. Therefore we have

$$\begin{aligned} \limsup_{\delta\to 0} \sup_{x\in\mathcal{R}} D(\mu^\delta_x||B_L(P_0, \delta_0)) &= \limsup_{\delta\to 0} \sup_{x\in\mathcal{R}} \inf_{\{P\in B_L(P_0,\delta_0)\}} D(\mu^\delta_x||P) && (2.47) \\ &\leq \limsup_{\delta\to 0} \sup_{x\in\mathcal{R}} D(\mu^\delta_x||P^x_{\delta_0-\delta_1}). && (2.48) \end{aligned}$$

We now prove $D(\mu^\delta_x||P^x_{\delta_0-\delta_1})$ can be uniformly bounded as x varies.

- For $x < u_{-\delta}$,

$$
\begin{aligned}
& D(\mu_x^\delta || P^x_{\delta_0-\delta_1}) \\
\leq\ & \delta \log \frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)+\delta) \log \frac{d(\mu_0(t+\delta)+\delta)}{d((1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1)} && (2.49)\\
=\ & \delta \log \frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)) \log \frac{d(\mu_0(t+\delta))}{(1-\delta_1)d(P_{\delta_0-\delta_1}(t+\delta))} && (2.50)\\
=\ & \delta \log \frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)) \log \frac{1}{(1-\delta_1)} \\
& + \int_{u_{-\delta}}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)) \log \frac{d(\mu_0(t+\delta))}{d(P_{\delta_0-\delta_1}(t+\delta))} && (2.51)\\
=\ & \delta \log \frac{\delta}{\delta_1} + (1-\delta) \log \frac{1}{(1-\delta_1)} + \int_{\mathcal{S}_l}^{u^\delta_{-\delta}+\delta} d(\mu_0(t)) \log \frac{d(\mu_0(t))}{d(P_{\delta_0-\delta_1}(t))}, && (2.52)
\end{aligned}
$$

  when $\delta \to 0$, the above converges to

$$\log \frac{1}{(1-\delta_1)} + D(\mu_0 || P_{\delta_0-\delta_1}).$$

- For $u_{-\delta} \leq x \leq u_\delta$,

$$
\begin{aligned}
& D(\mu_x^\delta || P^x_{\delta_0-\delta_1}) \\
=\ & (u_0(x+\delta)+\delta) \log \frac{(u_0(x+\delta)+\delta)}{(1-\delta_1)P_{\delta_0-\delta_1}(x+\delta)+\delta_1} \\
& + \int_{x+}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)+\delta) \log \frac{d(\mu_0(t+\delta)+\delta)}{d((1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1)} && (2.53)\\
\leq\ & \delta \log \frac{\delta}{\delta_1} + (u_0(x+\delta)) \log \frac{(u_0(x+\delta))}{(1-\delta_1)P_{\delta_0-\delta_1}(x+\delta)} \\
& + \int_{x+}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)) \log \frac{d(\mu_0(t+\delta))}{(1-\delta_1)d(P_{\delta_0-\delta_1}(t+\delta))} && (2.54)\\
\leq\ & \delta \log \frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)) \log \frac{d(\mu_0(t+\delta))}{(1-\delta_1)d(P_{\delta_0-\delta_1}(t+\delta))}, && (2.55)
\end{aligned}
$$

  which degenerate to the case of $x < u_{-\delta}$ since (2.55) is the same as (2.49).

- For $u_\delta < x \le u^\delta_{-\delta}$,

$$
\begin{aligned}
& D(\mu^\delta_x || P^x_{\delta_0-\delta_1}) \\
= & \int_{u_\delta}^{x-} d(\mu_0(t-\delta)-\delta) \log \frac{d(\mu_0(t-\delta)-\delta)}{d((1-\delta_1)P_{\delta_0-\delta_1}(t-\delta))} \\
& + (\mu_0(x+\delta)+\delta-(\mu_0(x-\delta)-\delta)) \log \frac{(\mu_0(x+\delta)+\delta-(\mu_0(x-\delta)-\delta))}{(1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1-(1-\delta_1)P_{\delta_0-\delta_1}(t-\delta)} \\
& + \int_{x+}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)+\delta) \log \frac{d(\mu_0(t+\delta)+\delta)}{d((1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1)} \qquad (2.56) \\
= & \int_{u_\delta}^{x-} d(\mu_0(t-\delta)) \log \frac{d(\mu_0(t-\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t-\delta)} \\
& + (2\delta+\mu_0(x+\delta)-\mu_0(x-\delta)) \log \frac{2\delta+\mu_0(x+\delta)-\mu_0(x-\delta)}{\delta_1+(1-\delta_1)(P_{\delta_0-\delta_1}(t+\delta)-P_{\delta_0-\delta_1}(t-\delta))} \\
& + \int_{x+}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)) \log \frac{d(\mu_0(t+\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t+\delta)} \qquad (2.57) \\
\le & \int_{u_\delta}^{x-} d(\mu_0(t-\delta)) \log \frac{d(\mu_0(t-\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t-\delta)} \\
& + 2\delta \log \frac{2\delta}{\delta_1} + \int_{x-\delta}^{x+\delta} d\mu_0(t) \log \frac{d\mu_0(t)}{(1-\delta_1)dP_{\delta_0-\delta_1}(t)} \\
& + \int_{x+}^{u^\delta_{-\delta}} d(\mu_0(t+\delta)) \log \frac{d(\mu_0(t+\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t+\delta)} \qquad (2.58) \\
= & 2\delta \log \frac{2\delta}{\delta_1} + \int_{u_\delta-\delta}^{u^\delta_{-\delta}+\delta} d\mu_0(t) \log \frac{d\mu_0(t)}{(1-\delta_1)dP_{\delta_0-\delta_1}(t)} \qquad (2.59) \\
= & 2\delta \log \frac{2\delta}{\delta_1} + (1-2\delta) \log \frac{1}{1-\delta_1} + \int_{u_\delta-\delta}^{u^\delta_{-\delta}+\delta} d\mu_0(t) \log \frac{d\mu_0(t)}{dP_{\delta_0-\delta_1}(t)}, \qquad (2.60)
\end{aligned}
$$

  when $\delta \to 0$, the above converges to

$$\log \frac{1}{1-\delta_1} + D(\mu_0 || P_{\delta_0-\delta_1}).$$

- Other symmetric cases can be solved similarly.

From the above arguments, we have

$$\limsup_{\delta \to 0} \sup_{x \in \mathcal{R}} D(\mu^\delta_x || B_L(P_0, \delta_0)) \le \log \frac{1}{1-\delta_1} + D(\mu_0 || B_L(P_0, \delta_0 - \delta_1)).$$

Notice this is true for any δ_1, let $\delta_1 \to 0$, we have

$$\begin{aligned}
\lim_{\delta\to 0}\sup_{x\in\mathcal{R}} D(\mu_x^{\delta}||B_L(P_0,\delta_0)) &\leq \lim_{\delta_1\to 0}\left(\log\frac{1}{1-\delta_1} + D(\mu_0||B_L(P_0,\delta_0-\delta_1))\right) && (2.61)\\
&= \lim_{\delta_1\to 0} D(\mu_0||B_L(P_0,\delta_0-\delta_1)) && (2.62)\\
&= D(\mu_0||B_L(P_0,\delta_0)), && (2.63)
\end{aligned}$$

the last equality comes from the fact that $D(\mu_0||B_L(P_0,\delta_0))$ is left continuous in δ_0 if $P_0(t)$ is continuous in t (Lemma 2.3.2). □

Lemma 2.3.6. *Given $P_0 \in \mathcal{P}$ and $\delta_0 > 0$, $D(\mu||B_L(P_0,\delta_0))$ is lower semicontinuous in μ with respect to the weak convergence.*

Proof. Assume $\mu_n \xrightarrow{w} \mu_0$. From (2.14), we know there exists $P_n \in B_L(P_0,\delta_0)$ such that $D(\mu_n||P_n) = D(\mu_n||B_L(P_0,\delta_0))$. Since $\bar{B}_L(P_0,\delta_0)$ is compact, there exists a subsequence of P_n (which we again denote by n) that converge to $P_{\mu_0} \in \bar{B}_L(P_0,\delta_0)$. $D(\mu||P_{\mu_0}) \leq \underline{\lim}_{n\to\infty} D(\mu_n||P_n)$ because $(\mu_n,P_n) \to (\mu_0,P_{\mu_0})$ and the KL divergence is lower semicontinuous. Therefore we have

$$\begin{aligned}
D(\mu_0||B_L(P_0,\delta_0)) &= D(\mu_0||\bar{B}_L(P_0,\delta_0)) && (2.64)\\
&\leq D(\mu_0||P_{\mu_0}) && (2.65)\\
&\leq \underline{\lim}_{n\to\infty} D(\mu_n||P_n) && (2.66)\\
&= \underline{\lim}_{n\to\infty} D(\mu_n||B_L(P_0,\delta_0)) && (2.67)
\end{aligned}$$

where (2.64) comes from (2.11). Therefore, according to the definition, $D(\mu||B_L(P_0,\delta_0))$ is lower semicontinuous in μ. □

2.4 Summary

It is straightforward to prove $D(\mu||B_L(P_0, \delta_0))$ is lower semicontinuous in μ; proving that it is also upper semicontinuous is tricky. The key step of the long proof in the previous section is Lemma 2.3.4, which is explained below.

For a fixed P_0, with small perturbation on μ, $D(\mu||P_0)$ may vary in an arbitrary manner, thus $D(\mu||P_0)$ is not upper semicontinuous. $B_L(P_0, \delta_0)$ provides the maximum freedom for tolerating the perturbation on μ, since the L$\acute{e}$vy metric is the weakest among other metrics. For all perturbations on μ that are within $B_L(\mu, \delta)$, the largest variation of $D(\mu||B_L(P_0, \delta_0))$ is achieved by a distribution whose CDF is constructed by shifting the $\mu(t)$ both horizontally and vertically to the edge of of $B_L(\mu, \delta)$. Such shifts can be tolerated by $B_L(P_0, \delta_0)$, so as the level of perturbation on μ decreases to 0, and the corresponding variation in $D(\mu||B_L(P_0, \delta_0))$ diminishes.

By proving $D(\mu||B_L(P_0, \delta_0))$ is both upper semicontinuou and lower semicontinuous in Lemma 2.3.5 and 2.3.6, we know that if $P_0(t)$ is continuous in t, $D(\mu||B_L(P_0, \delta_0))$ is continuous in μ with respect to the weak convergence. Therefore, for a fixed P_0, the sublevel and the superlevel sets of $D(\mu||B_L(P_0, \delta_0))$ are both closed, i.e., the probability space can be divided into two separated sets by the robust KL divergence. As elaborated in Chapter 1, such a separation is desired when constructing robust detectors for both universal hypothesis test and deviation detection. This is addressed in the next Chapter.

CHAPTER 3
ROBUST UNIVERSAL HYPOTHESIS TESTING AND DEVIATION DETECTION

3.1 Framework and Criteria

In this chapter, we focus our attention on the binary hypothesis testing problem. Each hypothesis is characterized by an uncertainty set. Denote by $\mathcal{P}$ the set of all probability measures defined on $\mathcal{R}$. Consider a sequence of observations $(X_0, \cdots, X_{n-1}) = X^n$ which are i.i.d. random variables with distribution $P \in \mathcal{P}$. Given X^n, one needs to determine whether P belongs to one of the two hypotheses, i.e.,

$$\mathcal{H}_1 : P \in \mathcal{P}_1, \quad \mathcal{H}_2 : P \in \mathcal{P}_2, \tag{3.1}$$

where $\mathcal{P}_1$ and $\mathcal{P}_1$ are two uncertainty sets that belong to $\mathcal{P}$.

The difficulty of the above problem is the composite nature of both hypotheses. A standard approach to designing decision rules in the above setting is the minimax NP criterion, first introduced by Huber [30] to optimize the worst case performance of the composite hypothesis testing problem. The decision rules thus obtained are said to be robust to the

uncertainty sets. The minimax NP criterion imposes a uniform constraint on the type-I error for all $P_1 \in \mathcal{P}_1$; subject to this constraint, we seek an acceptance region that minimizes the worst type-II error across $P_2 \in \mathcal{P}_2$. Thus we have the following constrained optimization problem:

$$\min_{S_n} \sup_{P_2 \in \mathcal{P}_2} P_2^n(S_n) \quad \text{s.t.} \quad \sup_{P_1 \in \mathcal{P}_1} P_1^n(S_n^c) \leq \alpha, \tag{3.2}$$

where $S_n \in \mathcal{R}^n$ is the acceptance region of $\mathcal{H}_1$.

For the above problem, the widely used method is identifying the LFDs within the classes, by establishing the property of jointly stochastic boundedness. Then the solution to the robust detection problem is a likelihood ratio test (LRT) between the pair of LFDs. It turns out that finding the LFDs is not quite simple, and the most effective approach is often to first guess the solution and then verify that it is indeed the LFD pair.

There are cases that the jointly stochastic boundedness does not hold. For example the deviation detection problem in Section 3.4 is one such an exception. However, one can still try to find an asymptotically optimal detector under the classical Hoeffding's approach using the generalized NP criterion, which evaluates the asymptotic efficiency by considering error exponents instead of error probabilities.

Let ϕ be the sequence of detectors $\{\phi^n(x_0, \cdots, x_{n-1}), n \geq 1\}$. Define the generalized error exponents for the two types of error probabilities respectively as follows,

$$I^{P_2}(\phi) := \underline{\lim}_{n\to\infty} -\frac{1}{n} \log P_2^n(x^n : \phi^n(x_0, \cdots, x_{n-1}) = 1),$$
$$J^{P_1}(\phi) := \underline{\lim}_{n\to\infty} -\frac{1}{n} \log P_1^n(x^n : \phi^n(x_0, \cdots, x_{n-1}) = 2).$$

Zeitouni and Gutman [14] have shown that to achieve the best trade-off between I^{P_2} and J^{P_1}, the test depends on x^n only through the empirical measure $\hat{\mu}_n$, which is defined

by

$$\hat{\mu}_n(t) = \frac{\sum_i I_{\{x_i \leq t\}}}{n}. \tag{3.3}$$

Let $\hat{\mathcal{P}}_n$ denote the set of all possible empirical distribution functions for n samples, $\cup_n \hat{\mathcal{P}}_n$ belongs to $\mathcal{P}$. Therefore $I^{P_2}(\phi)$ and $J^{P_1}(\phi)$ can be written as

$$I^{P_2}(\Omega) = \underline{\lim}_{n\to\infty} -\frac{1}{n} \log P_2^n(\hat{\mu}_n \in \Omega_1(n)), \tag{3.4}$$

$$J^{P_1}(\Omega) = \underline{\lim}_{n\to\infty} -\frac{1}{n} \log P_1^n(\hat{\mu}_n \in \Omega_2(n)). \tag{3.5}$$

Here Ω is a sequence of partitions $(\Omega_1(n), \Omega_2(n))$ $(n = 1, 2, \cdots)$ of which $\Omega_1(n) \cap \Omega_2(n) = \emptyset$ and $\mathcal{P} = \Omega_1(n) \cup \Omega_2(n)$. The decision rule is made in favor of H_i if $\hat{\mu}_n \in \Omega_i(n)$, $i = 1, 2$.

Similar to the fixed sample size problem, under a worst case constraint, i.e., a constraint on the minimal rate of decrease in type I probability of error, we want to maximize the exponent of the worst case type II probability of error. This is referred to as the minimax asymptotic NP hypothesis testing problem, defined as:

$$\sup_{\Omega} \inf_{P_2 \in \mathcal{P}_2} I^{P_2}(\Omega) \quad \text{s.t.} \quad \inf_{P_1 \in \mathcal{P}_1} J^{P_1}(\Omega) \geq \eta. \tag{3.6}$$

Throughout this Chapter, we will find the optimal detectors with respect to the minimax asymptotic NP criterion, for different sets of $(\mathcal{P}_1, \mathcal{P}_2)$.

3.2 An Overlooked Fact

Recall that the Lévy metric d_L makes $(\mathcal{P}, d_L)$ a metric space which is compatible with the weak topology on $\mathcal{P}$ [6]. The following result is well known and will be used throughout this chapter.

Lemma 3.2.1. *[15]For probability measures P_n, $P \in \mathcal{P}$, P_n weakly converges to P is equivalent to any of the following statements.*

1. *$E_{P_n} f \to E_P f$ for all the bounded and lipschitz continuous functions,*
2. *$P_n(\mathcal{A}) \to P(\mathcal{A})$ for any continuity set $\mathcal{A}$ of P.*
3. *$d_L(P_n, P) \to 0$.*

For any set $S_n \subseteq \mathcal{R}^n$, the boundary set of S_n is denoted as ∂S_n. For a set $\Gamma \subset \mathcal{P}$, the closure and interior of Γ in $\mathcal{P}$ with respect to the Lévy metric is denoted as $cl\Gamma$ and $int\Gamma$. A direct result of the above theorem is the following.

Theorem 3.2.2. *Consider the binary hypothesis testing problem,*

$$\mathcal{H}_1 : P \in \mathcal{P}_1, \quad \mathcal{H}_2 : P \in \mathcal{P}_2. \tag{3.7}$$

Under the minimax NP criterion (3.2), if $cl\mathcal{P}_1 \cap cl\mathcal{P}_2$ contains any P such that P is absolutely continuous on $\mathcal{R}$, then any set $S_n \subseteq \mathcal{R}^n$ with ∂S_n having measure 0 is no better than random guess.

Proof. Assume $P \in cl\mathcal{P}_1 \cap cl\mathcal{P}_2$ and P is absolutely continuous on $\mathcal{R}$, there exists sequences $\{P_k^i \in \mathcal{P}_i\}$ that weakly converge to P for $i = 1, 2$. For any $S_n \subseteq \mathcal{R}^n$ with ∂S_n having measure 0, $P(\partial S_n) = 0$ because P is absolutely continuous on $\mathcal{R}$, that is, S_n is a continuity set of P. Then from the second equivalent condition in Lemma 3.2.1 we have $P(S_n) = \lim_{k\to\infty} P_k^i(S_n)$ for $i = 1, 2$.

Suppose $\sup_{P_2\in\mathcal{P}_2} P_2(\phi^n = 1) \leq \alpha$, then

$$\begin{aligned}
\sup_{P_1\in\mathcal{P}_1} P_1(S_n^c) &\geq \lim_k P_k^1(S_n^c) && (3.8)\\
&= P(S_n^c) && (3.9)\\
&= \lim_k P_k^2(S_n^c) && (3.10)\\
&\geq \inf_{P_2\in\mathcal{P}_2} P_2(S_n^c) && (3.11)\\
&= 1 - \sup_{P_2\in\mathcal{P}_2} P_2(S_n), && (3.12)\\
&\geq 1-\alpha. && (3.13)
\end{aligned}$$

Therefore, we can use a random guess independent of the observations to achieve the optimality in problem (3.2). □

Remark 1. *The above statement also holds for the asymptotic minimax NP criterion (3.6).*

Loosely speaking, Theorem 3.2.2 states that the two uncertainty sets need to be separated, that is, the two sets can not be arbitrarily close to each other with respect to the weak convergence. Otherwise the minimax hypothesis testing problem becomes degenerate. Theorem 3.2.2 may seem trivial, yet the result is quite subtle and is often overlooked. In the discrete case with finite, say, m elements, $\mathcal{P}$ is a compact subspace of the m-dimensional Euclidean space. The uncertainty sets under $\mathcal{H}_1$ and $\mathcal{H}_2$ are usually characterized by continuous functions, in this case "disjoint" usually implies "separated". In general, for the finite alphabet case, as long as the two uncertainty sets are disjoint, Theorem 3.2.2 is redundant. However, this is not the case for continuous distributions; in fact some well defined problems for the discrete case do not generalize to the continuous case because of Theorem 3.2.2. An example is the moment constrained testing problem as elaborated below.

In moment constrained testing problems, the uncertainty sets $\mathcal{P}_1$ and $\mathcal{P}_2$ are specified

by the union or intersection of the sets

$$\{\mu \in \mathcal{P} : E_\mu f \leq c\},$$

where f is a real-valued function. The motivation for considering moment classes comes from the simple observation that moments, as mean or correlation, are often much easier to handle and characterize than the complete statistical distribution. In the finite alphabets case, the moment constrained testing problem is well defined and has a long and rich history [31]. However, in the continuous case, the moment constrained testing problem may turn out to be meaningless. This can be illustrated using the following observation, which simply states that any probability distribution, including those whose moments do not exist, is arbitrarily close to the set of distributions with mean equal to 0.

Lemma 3.2.3. *cl*$\{\mu \in \mathcal{P} : E_\mu X = 0\} = \mathcal{P}$.

Proof. For any distribution $P \in \mathcal{P}$ whose expected value $E_P X$ may or may not exist, e.g., P may be the Cauchy distribution. Let P_n be the truncated and appropriately normalized version of P on the interval $[-n, n]$, thus $E_{P_n} X$ always exists. Define

$$\mu_n = \left(1 - \frac{1}{n}\right) P_n + \frac{1}{n} I_n, \tag{3.14}$$

where

$$I_n(t) = \begin{cases} 0 & \text{for } t < -(n-1)E_{P_n} X, \\ 1 & \text{for } t \geq -(n-1)E_{P_n} X. \end{cases} \tag{3.15}$$

I_n is simply a degenerate probability measure with probability 1 at $-(n-1)E_{P_n} X$. It is easy to see that $\mu_n \xrightarrow{w} P$ and $E_{\mu_n} X = 0$. That is, there exist a sequence of distributions weakly converging to P but the sequence belongs to the set $\{\mu \in \mathcal{P} : E_\mu X = 0\}$. □

Similarly, one can show that for any $m > 0$, cl$\{\mu \in \mathcal{P} : E_\mu X^2 \geq m\} = \mathcal{P}$. As we can see in the above proof, the underlying reason is that if f is unbounded, $E_\mu f$ is sensitive to small disturbance at the tail of the distribution μ, while the Lévy metric is not.

Therefore, in the moment constrained robust detection problem defined on the real line, one needs to make sure that the two uncertainty sets are not arbitrarily close to each other with respect to the weak convergence, otherwise the problem becomes degenerate. This subtle presumption was neglected in the previous literatures, such as [33].

If the moment constraints are all characterized using the bounded and Lipschitz continuous functions, from the first equivalent condition in Lemma 3.2.1, it is obvious to see that the uncertainty sets are closed. Thus the moment constrained robust detection problems are well defined.

In Section 3.3 and Section 3.4, two robust detection problems are discussed. One is the robust universal hypothesis testing and the other is deviation detection. Similar to moment constrained detection, the issue that two uncertainty sets in the continuous case are not separated needs to be resolved. The detailed observation and solution will be elaborated in Section 3.4.

3.3 Robust Universal Hypothesis Testing

3.3.1 Related Work

In [14], the problem of deciding whether an i.i.d. sequence of random variables originate from a known continuous source P_0 or an unknown source P' different from P_0 is considered. This can be modeled as in (3.1) with

$$\mathcal{P}_1 = \{P_0\}, \quad \mathcal{P}_2 = \{P'\}$$

where $P' \neq P_0$. The fact that P' is unknown gives rise to the name the universal hypothesis testing, and the goal is to find an optimal detector under the asymptotic minimax criterion (3.6).

The problem was first formulated by Hoeffding [2], where the alphabet of the i.i.d. source is finite. By using combinatorial bounds, he successfully constructed an optimal detector that takes the following simple form,

$$D(\hat{\mu}_n||P_0) \underset{H_1}{\overset{H_2}{\gtrless}} \eta. \tag{3.16}$$

Unfortunately, when the distribution is on the real line, the detector described in (3.16) fails to be the optimal and Hoeffding's approach cannot be directly extended as the combinatorial approach does not apply to the case with continuous alphabet.

However, under a weaker notion of optimality, Zeitouni and Gutman [14] extended the classical work by Hoeffding from the discrete case to the continuous case. In lieu of the combinatorial approach, they resorted to the large deviation theory. Their proof primarily was based on the following general Sanov's Theorem.

Theorem 3.3.1 (General Sanov's Theorem)**.** *[6] Given a probability set $\Gamma \subseteq \mathcal{P}$, for a probability measure $Q \notin \Gamma$,*

$$\begin{aligned} \inf_{P \in cl\Gamma} D(P||Q) &\leq \underline{\lim}_{n\to\infty} -\frac{1}{n} \log Q(\{x^n : \hat{\mu}_n \in \Gamma\}) \\ &\leq \overline{\lim}_{n\to\infty} -\frac{1}{n} \log Q(\{x^n : \hat{\mu}_n \in \Gamma\}) \\ &\leq \inf_{P \in int\Gamma} D(P||Q). \end{aligned}$$

The general Sanov's Theorem illustrates the large deviation principle for the empirical measures and will be used extensively in the proof of Theorems 3.3.2 and Theorem 3.3.3.

For any set $\Gamma \subseteq \mathcal{P}$, define its $\delta-$smooth set to be

$$\Gamma^{\delta} := \cup_{\mu \in \Gamma} \{P \in \mathcal{P} : d_L(P, \mu) < \delta\}.$$

The major contribution in [14] is the following.

Theorem 3.3.2. *[14] Define Λ as,*

$$\Lambda_2(n) = \Lambda_2 := \{\mu : D(B_L(\mu, 2\delta)||P_0) \geq \eta\}^{\delta} \;\; \Lambda_1 := \mathcal{P}/\Lambda_2. \tag{3.17}$$

Λ is $\delta-$optimal, i.e.,

1. *$J^{P_0}(\Lambda) \geq \eta$.*
2. *If Ω is a test such that $J^{P_0}(\Omega^{6\delta}) \geq \eta$, then for any $P' \neq P_0$,*

$$I^{P'}(\Omega^{\delta}) \leq I^{P'}(\Lambda). \tag{3.18}$$

In the finite alphabets case, the corresponding detector as in (3.17) yields weaker results than Hoeffding's detector [2]. This is the price paid for its generality - Theorem 3.3.2. applies to both discrete and of $\mathcal{R}$-valued random variables.

3.3.2 Some observations

For discrete distributions, the above detector (3.17) yields weaker result compared with Hoeffding's detector [2]. However, one has to be content with "$\delta-$optimality" rather than "optimality" in the continuous case, if there is no restriction on detector Ω and the general Sanov's Theorem is used. It is plausible that for a test Ω, either Ω_1 or Ω_2, say Ω_1, consists of only empirical distributions, since the test can depend on the observations only through the empirical distributions. Then the interior point set of Ω_1 is empty and the closure of Ω_2 equals to $\mathcal{P}$. For such a test, one can not take advantage of the general Sanov's Theorem to

analyze the error exponents. That is why in Theorem 3.3.2, for an arbitrary test Ω, we need to first perform $\delta-$smooth operation on it before comparing its error exponents to those of the test Λ. As such, if there is no restriction on detector Ω, we have "$\delta-$optimality" rather than "optimality" adopted in the continuous case.

Detector (3.17) has a complicated form. Given the empirical distribution $\hat{\mu}_n$, it is hard to determine whether $\hat{\mu}_n \in \Lambda_1$ or Λ_2 due to the following two reasons.

- First, one has to compute $D(B_L(\hat{\mu}_n, 2\delta)||P_0)$. From Fig. 3.1, we can see that computing $D(B_L(\hat{\mu}_n, 2\delta)||P_0)$ is an infinite dimension optimization problem. This is, in essence, equivalent to finding a continuous μ^* inside the shaded region such that $D(\mu^*||P_0) = \inf_{\mu \in B_L(\hat{\mu}_n, 2\delta)} D(\mu||P_0)$.

- Secondly, assume one can compute $D(B_L(\hat{\mu}_n, 2\delta)||P_0)$. If $D(B_L(\hat{\mu}_n, 2\delta)||P_0) \geq \eta$ then $\hat{\mu}_n \in \Lambda_1$. But if $D(B_L(\hat{\mu}_n, 2\delta)||P_0) < \eta$, one needs to further check if $\hat{\mu}_n$ belongs to the δ-smooth set of $\{\mu : D(B_L(\mu, 2\delta)||P_0) \geq \eta\}$.

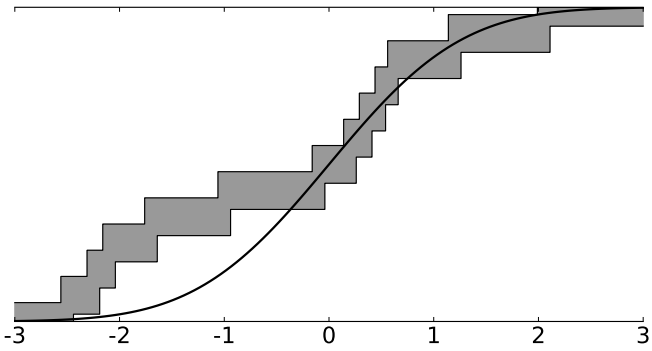


Fig. 3.1: The shaded region is $B_L(\mu_n, 2\delta)$ and the solid line is P_0.

One of the difficulties to directly generalize the discrete case to the continuous case, as mentioned in [14], is that $\{P \in \mathcal{P} : D(P||P_0) \geq \eta\}$ is not closed in $\mathcal{P}$. Actually, it will be shown in Section 3.4 that

$$\{P \in \mathcal{P} : D(P||P_0) \geq \eta\} \neq cl\{P \in \mathcal{P} : D(P||P_0) \geq \eta\} = \mathcal{P}. \tag{3.19}$$

In Section 3.3.3, rather than "$\delta-$smoothing" the detector as one does in (3.17), we "$\delta-$smooth" P_0 to be a Lévy ball centered at P_0. Then, under the minimax criterion, the empirical likelihood ratio test which is simple and intuitive, is shown to be optimal.

Besides, Theorem 3.3.2 can not be directly extended to the case where $\mathcal{P}_1$ is an arbitrary set of distributions. For the proof in [14] to hold, $\{P \in \mathcal{P} : D(P||\mathcal{P}_1) \leq \eta\}$ should be compact. However, if we let $\mathcal{P}_1 = \cup_{n>0} P_n$ where P_n is $\mathcal{N}(n, 1)$. Then the set $\{P \in \mathcal{P} : D(P||\mathcal{P}_1) \leq \eta\}$ is not compact, since the sequence $\{P_n\}$ belongs to this set but does not have a subsequence that has a limit point in $\mathcal{P}$.

3.3.3 Robust Universal Hypothesis Testing

The robust universal hypothesis testing is the generalization of the universal hypothesis testing to the robust setting, which is modeled as in (3.1) with,

$$\begin{aligned} \mathcal{P}_1 &= B_L(P_0, \epsilon_0), \\ \mathcal{P}_2 &= \{P'\}. \end{aligned}$$

Here P_0 is assumed to be a known continuous distribution and $\epsilon_0 > 0$. $P' \notin B_L(P_0, \epsilon_0)$ and is unknown just as in the universal hypothesis testing.

The reason to use the L$\acute{e}$vy metric among numerous distance metrics between distributions is that the L$\acute{e}$vy metric is the weakest [18], in another word, $B_L(P_0, \epsilon_0)$ contains all distributions that are close enough to P_0 as measured using any other metrics. In addition, the optimal solution will be shown to be rather straightforward. Theorem 3.3.3 below describes the optimal solution to the above problem.

Theorem 3.3.3. *For the robust universal hypothesis testing problem, for any given $\delta > 0$, detector*

$$D(\hat{\mu}_n||B_L(P_0, \epsilon_0)) \underset{H_1}{\overset{H_2}{\gtrless}} \eta, \tag{3.20}$$

is optimal among all detector Ω such that $\Omega_2(n) = \Omega_2$ and Ω_2 is open, i.e., let

$$\Lambda_2(n) = \Lambda_2 := \{\mu : D(\mu||B_L(P_0, \epsilon_0)) > \eta\},\ \Lambda_1(n) := \Lambda_1 = \{\mu : D(\mu||B_L(P_0, \epsilon_0)) \leq \eta\},$$

then,

1. $\inf_{P_1 \in \mathcal{P}_1} J^{P_1}(\Lambda) = \eta$.
2. $I^{P'}(\Lambda) = D(\Lambda_1||P')$.
3. *For detector Ω with $\Omega_2(n) = \Omega_2$ and Ω_2 open, if*

$$\inf_{P_1 \in \mathcal{P}_1} J^{P_1}(\Omega) > \eta, \tag{3.21}$$

then for any $P' \notin B_L(P_0, \epsilon_0)$,

$$I^{P'}(\Omega) \leq I^{P'}(\Lambda). \tag{3.22}$$

Proof. 1. By Sanov's theorem, we have

$$\begin{aligned}
\inf_{P_1 \in \mathcal{P}_1} J^{P_1}(\Lambda) &= \inf_{P \in \mathcal{P}_1} \varliminf_{n \to \infty} -\frac{1}{n} \log P_1^n(\{x^n : \hat{\mu}_n \in \Lambda_2\}) && (3.23)\\
&\geq \inf_{P \in \mathcal{P}_1} \inf_{\mu \in cl\Lambda_2} D(\mu||P) && (3.24)\\
&= \inf_{\mu \in cl\Lambda_2} D(\mu||\mathcal{P}_1) && (3.25)\\
&= \eta, && (3.26)
\end{aligned}$$

the last equality holds since $D(\mu||\mathcal{P}_1)$ is continuous in μ thus $cl\Lambda_2 \subseteq \{\mu : D(\mu||\mathcal{P}_1) \geq$

$\eta\}$. On the other hand,

$$\begin{aligned}
\inf_{P_1\in\mathcal{P}_1} J^{P_1}(\Lambda) &\leq \inf_{P\in\mathcal{P}_1} \overline{\lim_{n\to\infty}} -\frac{1}{n}\log P_1^n(\{x^n : \hat{\mu}_n \in \Lambda_2\}) && (3.27)\\
&\leq \inf_{P\in\mathcal{P}_1} \inf_{\mu\in int\Lambda_2} D(\mu||P) && (3.28)\\
&= \eta. && (3.29)
\end{aligned}$$

The last equality holds since $int\Lambda_2 = \Lambda_2$.

2. Again from Sanov's theorem, we have

$$\begin{aligned}
I^{P'}(\Lambda) &= \underline{\lim_{n\to\infty}} -\frac{1}{n}\log P'^n(\{x^n : \hat{\mu}_n \in \Lambda_1\}) && (3.30)\\
&\geq \inf_{\mu\in cl\Lambda_1} D(\mu||P') && (3.31)\\
&= D(\Lambda_1||P'). && (3.32)
\end{aligned}$$

The last equality holds since $cl\Lambda_1 = \Lambda_1$. On the other hand, $\{\mu : D(\mu||\mathcal{P}_1) < \eta\} \subseteq int\Lambda_1$, thus,

$$\begin{aligned}
I^{P'}(\Lambda) &\leq \overline{\lim_{n\to\infty}} -\frac{1}{n}\log P'^n(\{x^n : \hat{\mu}_n \in \Lambda_1\}) && (3.33)\\
&\leq \inf_{\mu\in int\Lambda_1} D(\mu||P') && (3.34)\\
&\leq \inf_{\mu\in\{\mu:D(\mu||\mathcal{P}_1)<\eta\}} D(\mu||P') && (3.35)\\
&\leq D(\Lambda_1||P'). && (3.36)
\end{aligned}$$

Equation (3.36) holds because of the following. There exists a distribution $P \in \mathcal{P}_1$ such that $D(P||P') < \infty$. For any $P_c \in \Lambda_1$ and $0 < \lambda < 1$, we have $(1-\lambda)P_c+\lambda P \in$

$\{\mu : D(\mu||\mathcal{P}_1) < \eta\}$ since

$$\begin{aligned} & D((1-\lambda)P_c + \lambda P||\mathcal{P}_1) \\ \leq\ & (1-\lambda)D(P_c||\mathcal{P}_1) + \lambda D(P||\mathcal{P}_1)) && (3.37) \\ <\ & (1-\lambda)\eta + 0 && (3.38) \\ <\ & \eta, && (3.39) \end{aligned}$$

where (3.37) comes from the fact that $D(\mu||\mathcal{P}_1)$ is convex in μ, which is proved in Lemma 2.3.3. Since $P \in \mathcal{P}_1$, we also have (3.38). Then,

$$\begin{aligned} \inf_{\mu\in\{\mu:D(\mu||\mathcal{P}_1)<\eta\}} D(\mu||P') &\leq \lim_{\lambda\to 0^+} D((1-\lambda)P_c + \lambda P||P') && (3.40) \\ &\leq \lim_{\lambda\to 0^+} (1-\lambda)D(P_c||P') + \lambda D(P||P') && (3.41) \\ &\leq D(P_c||P'), && (3.42) \end{aligned}$$

the last inequality holds since $D(P||P') < \infty$. The above inequalities hold for any $P_c \in \Lambda_1$, thus we have

$$\inf_{\mu\in\{\mu:D(\mu||\mathcal{P}_1)<\eta\}} D(\mu||P') \leq D(\Lambda_1||P'). \tag{3.43}$$

3. We have

$$\begin{aligned} \inf_{P_1\in\mathcal{P}_1} D(\Omega_2||P_1) &= \inf_{P_1\in\mathcal{P}_1} D(int\Omega_2||P_1) && (3.44) \\ &\geq \inf_{P_1\in\mathcal{P}_1} \varliminf_{n\to\infty} -\frac{1}{n}\log P_1(x^n : \hat{\mu}_n \in \Omega_2) && (3.45) \\ &> \eta. && (3.46) \end{aligned}$$

Therefore, $\Omega_2 \subseteq \Lambda_2$, or equivalently, $\Lambda_1 \subseteq \Omega_1$. Next,

$$\begin{aligned} I^{P'}(\Omega_1) &= \underline{\lim}_{n\to\infty} -\frac{1}{n}\log P'^n(x^n : \hat{\mu}_n \in \Omega_1) && (3.47)\\ &\leq \underline{\lim}_{n\to\infty} -\frac{1}{n}\log P'^n(x^n : \hat{\mu}_n \in \Lambda_1) && (3.48)\\ &= I^{P'}(\Lambda). && (3.49) \end{aligned}$$

□

One reason that detector (3.17) is complicated is due to the fact that $\{\mu \in \mathcal{P} : D(\mu||P_0) \geq \eta\}$ is not closed, as shown in (3.19). However, $\{\mu \in \mathcal{P} : D(\mu||\mathcal{P}_1) \geq \eta\}$ is closed since $D(\mu||\mathcal{P}_1)$ is continuous in μ (Theorem 2.2.1). This is a very important step in proving Theorem 3.3.3.

Compared to detector (3.17) in Theorem 3.3.2, detector (3.20) has three main differences.

- Fig. 3.2 shows that computing $D(\hat{\mu}_n||B_L(P_0, \delta_0))$ is a finite dimension optimization problem, which is in essence finding a step function inside the shaded area that achieves the minimum KL divergence to $\hat{\mu}_n$. This can be shown to be a convex optimization problem whose solution can be computed efficiently.

- As mentioned in the previous section, without any restrictions on Ω, we can only get $\delta-$optimality. The reason is that to characterize the asymptotic performance of an detector Ω, Sanov's Theorem will inevitably be used, which relies on the interior set or the closure of Ω. However, for an arbitrary detector Ω, its interior set could be empty and its closure could be $\mathcal{P}$, or its interior set and closure are too abstract or complicated to describe. In these cases, one can hardly draw any conclusion using the Sanov's Theorem.

  Zeitouni and Gutman proposed a way in Theorem 3.3.2 to get around the above difficulty, by comparing Ω^δ instead of Ω. The advantage to doing so is that Ω_2^δ is

open and Ω_1^δ is closed, the price paid is the weaker optimality, i.e., one has to settle with the δ-optimality.

In Theorem 3.3.3, by restricting Ω to be independent of n and assuming Ω_2 is open, we extend the $\delta-$optimality to optimality, with a much simplified proof compared to that of Theorem 3.3.2.

- Theorem 3.3.2 only gives the lower bound of the error exponents, while Theorem 3.3.3 specifies the exact value of the error exponents. Furthermore, I and J are defined using limit infimum, yet from the proof it can be seen that I and J remain unchanged if one uses limit to define the worse case error exponents. Therefore Theorem 3.3.3 gives an exact characterization of the error exponents.

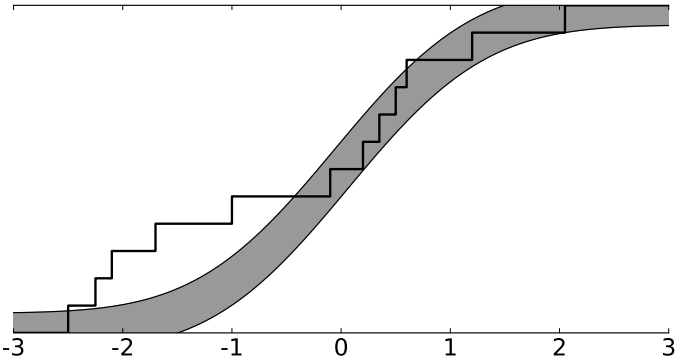


Fig. 3.2: The shaded region is a L$\acute{e}$vy ball of the normal distribution and the step function is an example of $\hat{\mu}_n$.

After generalizing the universal hypothesis testing to the robust setting, the generalized empirical likelihood ratio test becomes optimal, the construction of detector and the proof of optimality are much simplified. In the next section, we will explore the deviation detection problem.

3.4 Deviation Detection

3.4.1 Introduction

We consider in this section the so-called deviation detection problem. The normal state, i.e., the null hypothesis, is characterized by a known nominal distribution; any *significant* departure from this nominal distribution constitutes the alternative hypothesis. The most sensible formulation is thus the following composite hypothesis testing problem: under the null hypothesis, samples follow a distribution from a suitably defined proximity distribution set close to the nominal distribution; under the alternative hypothesis, samples follow a distribution that is significantly different from the nominal distribution.

Deviation detection has numerous engineering and societal applications, including network intrusion detection, fraud detection, quality control, vacant channel detection in cognitive wireless networks [21–24]. What is common among those applications is that the normal operating state is often characterized by a known distribution of some observables, e.g., learned from past history. Anomaly, in statistical distribution of the observables, occurs when the system deviates from the normal operating state and such anomaly is often not known *a priori* due to the unpredictable nature of the cause of the anomaly.

Closely related to the deviation detection problem is a class of the robust detection problems [25–29], in which *each* uncertainty set is populated by probability distributions within a proximity set defined using a certain metric, e.g., the KL divergence, with respect to the respective nominal distributions. For such a robust detection problem, to minimize the worst case performance over the uncertainty classes, the solution typically involves identifying a pair of least favorable distributions (LFDs), and subsequently designing a simple hypothesis test between the LFDs. However, as elaborated in Section 3.1, the existence of LFDs requires the joint stochastic boundedness property. Such a property, however, does not hold for the deviation detection problem, for the same reason as that of the universal

hypothesis testing problem.

The key distinction between the deviation detection and the robust detection described above is that the former has only a single nominal distribution (under the null hypothesis) while the latter has a nominal distribution under each hypothesis. This distinction turns out to be crucial. In fact, with continuous valued random sequences, the KL divergence is not a suitable metric for the deviation detection problem in defining the proximity set for the nominal distribution. Specifically, while for discrete random variables, it is guaranteed that the complementary set of any open ball defined using the KL divergence is closed, such is not the case with continuous valued random variables. As such, defining the distribution set for the alternative hypothesis using the KL divergence would encounter significant issues that render the deviation detection problem meaningless.

We thus turn into the L$\acute{e}$vy metric as it is a true metric that metrizes the weak convergence of probability measures. To facilitate the analysis, instead of solving the fixed sample size problem, we follow the generalized NP criterion, which evaluates the asymptotic efficiency by considering the error exponents instead of the error probabilities.

In Section 3.4.2, we formulate the problem and establish that the deviation detection problem characterized by the KL divergence is a degenerate one. Subsequently, we show that by defining the proximity set using the L$\acute{e}$vy metric makes the problem meaningful. We then establish that, the generalized empirical likelihood ratio test turns out to be asymptotically δ-optimal for the worst case performance.

3.4.2 Problem Formulation and Solution

We now formulate the deviation detection problem as follows. Let P_0 be a continuous distribution representing the normal state. Under the null hypothesis the samples follow a distribution from the set $\mathcal{P}_1$ which contains distributions close to P_0, i.e., $\mathcal{P}_1$ is the proximity set of P_0; under the alternative hypothesis the sample distribution belongs to the deviation set $\mathcal{P}_2$, which contains all the distributions that are significantly different from P_0. This can

be written as in (3.1) with

$$\mathcal{P}_1 = \{P \in \mathcal{P} : d(P, P_0) \leq \lambda_1\} \text{ (proximity set)},$$
$$\mathcal{P}_2 = \{P \in \mathcal{P} : d(P, P_0) \geq \lambda_2\} \text{ (deviation set)},$$

where $\lambda_1 < \lambda_2$ and d represents a suitably defined measure of distance between two distributions. Throughout this subsection, we use $\mathcal{P}_1$ and $\mathcal{P}_2$ as defined above.

For robust detection, a wide range of measures have been adopted to define the uncertainty sets that are typically proximity sets of two nominal distributions under the two hypotheses [25–29]. These include the total variation, the Kolmogorov distance, the L$\acute{e}$vy distance, the Hellinger divergence and the KL divergence. However, in the deviation detection problem, the fact that there is only one nominal distribution for both hypotheses precludes the use of many of those measures. This subtle but important observation will be elaborated in Proposition 2 where we demonstrate that deviation from the nominal distribution (i.e., under the alternative hypothesis) is poorly characterized using the KL divergence when the distributions are defined on the real line (i.e., continuous valued).

When the sample space is finite, the set of all probabilities is a compact subset of the Euclidean space. Any d as previously mentioned can be used to define the deviation detection problem. This is because in the discrete case, in general the distance metric of any probability distribution to P_0 is a continuous function which results in $\mathcal{P}_1$ and $\mathcal{P}_2$ being two disjoint compact sets as long as $\lambda_2 > \lambda_1$.

However, if the sample space is $\mathcal{R}$, it is possible that $d(\cdot, P_0)$ is not continuous on $\mathcal{P}$ and the corresponding $cl\mathcal{P}_1 \cap cl\mathcal{P}_2$ is not empty even if $\lambda_2 > \lambda_1$, leading to the degeneration of the hypothesis testing problem. This is true despite the fact that, $\mathcal{P}_1$ and $\mathcal{P}_2$ are themselves two disjoint sets. Proposition 2 illustrates this point using the KL divergence.

Proposition 2. *Assume P_0 is the standard normal distribution. For any given $\lambda > 0$, let*

$$\mathcal{P}_\lambda := \{P \in \mathcal{P} : D(P||P_0) = \lambda\},$$

then

$$cl\mathcal{P}_\lambda = \{P \in \mathcal{P} : D(P||P_0) \leq \lambda\}.$$

Proposition 3 states that the closure of the surface defined by distributions with constant KL divergence to the nominal distribution is the entire sphere, i.e., includes all distributions whose KL divergence is smaller than or equal to the radius of the surface.

Proof. It was shown in [6, 17] that $\{P \in \mathcal{P} : D(P||P_0) \leq \lambda\}$ is a compact set, which implies that $cl\mathcal{P}_\lambda \subseteq \{P \in \mathcal{P} : D(P||P_0) \leq \lambda\}$. For the other direction, we only need to show that for any $P \in \mathcal{P}$ such that $D(P||P_0) < \lambda$, P is a limit point of $\mathcal{P}_\lambda$. We show this by constructing a sequence $\{P_n\}$ that belongs to the set $\mathcal{P}_\lambda$ while the sequence $\{P_n\}$ weakly converges to P.

Let $P_n(t) = (1 - \frac{1}{n})P + \frac{1}{n}Q_n(t)$, where $Q_n(t)$ is Gaussian with mean zero and variance t, then $D(P_n(t)||P_0)$ is a continuous function of t on $(0, 1]$ for any n. Let $t = 1$ then $Q_n(1)$ reduces to P_0 and we have

$$\begin{aligned} D(P_n(1)||P_0) &\leq \left(1 - \frac{1}{n}\right) D(P||P_0) + \frac{1}{n} D(Q_n(1)||P_0) \\ &= \left(1 - \frac{1}{n}\right) D(P||P_0) \\ &\leq D(P||P_0) \\ &< \lambda, \end{aligned} \tag{3.50}$$

where the first inequality comes from the convexity property of KL divergence.

Denote by $S_\epsilon = [-\epsilon, \epsilon]$. From the data processing inequality of the KL divergence, if the observations are quantized to S_ϵ and $\mathcal{R} \setminus S_\epsilon$, then the corresponding quantized KL

divergence will be no greater than the original KL divergence. Denote the resulting KL divergence by

$$D^b(\alpha||\beta) := \alpha \log \frac{\alpha}{\beta} + (1-\alpha) \log \frac{1-\alpha}{1-\beta},$$

then

$$\begin{aligned}
\underline{\lim}_{t\to 0} D(P_n(t)||P_0) &\geq \underline{\lim}_{t\to 0} D^b\left((1-\frac{1}{n})P(x \in S_\epsilon) + \frac{1}{n}Q_n(t)(S_\epsilon)||P_0(S_\epsilon)\right) \\
&= \lim_{t\to 0} D^b\left((1-\frac{1}{n})P(x \in S_\epsilon) + \frac{1}{n}Q_n(t)(S_\epsilon)||P_0(S_\epsilon)\right) \\
&= D^b\left((1-\frac{1}{n})P(S_\epsilon) + \lim_{t\to 0}\frac{1}{n}Q_n(t)(S_\epsilon)||P_0(S_\epsilon)\right) \\
&= D^b\left((1-\frac{1}{n})P(S_\epsilon) + \frac{1}{n}||P_0(S_\epsilon)\right). \qquad (3.51)
\end{aligned}$$

The first inequality comes from the data processing inequality. The first and second equalities holds since $D^b(\cdot||P_0(S_\epsilon))$ is a continuous function because $0 < P_0(S_\epsilon) < 1$. When $t \to 0$ the quantity $Q_n(t)(S_\epsilon) \to 1$ as $Q_n(t)$ is zero mean Gaussian distribution with variance t. Therefore we have the last equality. Since the above relation holds for any $\epsilon > 0$, we can take the supremum over all $\epsilon > 0$ in (3.51) and we get

$$\begin{aligned}
\underline{\lim}_{t\to 0} D(P_n(t)||P_0) &\geq \sup_{\epsilon>0} D^b((1-\frac{1}{n})P(x \in S_\epsilon) + \frac{1}{n}||P_0(x \in S_\epsilon)) \\
&= \infty. \qquad (3.52)
\end{aligned}$$

Thus from relations (3.50), (3.52) and the continuity of $D(P_n(t)||P_0)$ on $0 < t \leq 1$, there exists a $t_n \in (0, 1]$ such that for each n

$$D(P_n(t_n)||P_0) = \lambda.$$

Next we will show that $P_n(t_n)$ converges to P in the Lévy metric. Assume that a function

f is both bounded by $[m, M]$ and lipschitz continuous on $\mathcal{R}$, then

$$\begin{aligned} E_{P_n(t_n)}f &= (1-\frac{1}{n})E_P f + \frac{1}{n}E_{Q_n(t_n)}f \\ &\leq (1-\frac{1}{n})E_P f + \frac{1}{n}M. \end{aligned}$$

Similarly,

$$E_{P_n(t_n)}f \geq (1-\frac{1}{n})E_P f + \frac{1}{n}m.$$

Thus

$$\lim_{n\to\infty} E_{P_n(t_n)}f = E_P f. \tag{3.53}$$

From Lemma 3.2.1, (3.53) is equivalent to $d_L(P_n(t_n), P) \to 0$, so P is a limit point of $\mathcal{P}_\lambda$. □

Remark 2. *This result can be generalized to any arbitrary* P_0.

If we let $\mathcal{P}_2$ be $\{P \in \mathcal{P} : D(P||P_0) > \lambda_2\}$, then according to Proposition 2,

$$\mathcal{P}_1 \subseteq cl\mathcal{P}_2,$$

which violates Theorem 3.2.2 and the corresponding minimax detection problem would be degenerate.

The L$\acute{e}$vy metric d_L is suitable in our problem because of the following reasons. Firstly, the proximity set measured by the L$\acute{e}$vy metric is more general, hence more inclusive compared to that defined using the contamination model, the total variation, the Kolmogorov distance and the KL divergence models. Take the KL divergence for instance, for any

$\lambda_1 > 0$, there exists a λ_1' such that

$$\{P \in \mathcal{P} : D(P||P_0) \leq \lambda_1\} \subseteq \{P \in \mathcal{P} : d_L(P, P_0) \leq \lambda_1'\}.$$

However, the reverse statement is not true, i.e., for any λ_1, λ_1',

$$\{P \in \mathcal{P} : d_L(P, P_0) \leq \lambda_1'\} \nsubseteq \{P \in \mathcal{P} : D(P||P_0) \leq \lambda_1\}.$$

The reason is that, in general, convergence in d_L is strictly weaker compared to convergence in any other measures [15]. That means any proximity set constructed by d_L includes all the distributions which are close enough in any other d to P_0. On the other hand, if the proximity set is constructed by metrics other than d_L, then it will exclude some distributions close to P_0 in terms of d_L.

Another reason, which is a consequence of the fact that d_L is the weakest true distance metric, is that the corresponding $\mathcal{P}_i = cl\mathcal{P}_i, i = 1, 2$, and $\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset$ as long as $\lambda_2 > \lambda_1$.

Therefore the deviation detection problem to be considered here is now formulated using the L$\acute{e}$vy metric with the two probability sets under the two hypotheses defined as,

$$\mathcal{P}_1 = \{P \in \mathcal{P} : d_L(P, P_0) \leq \lambda_1\} \text{ (proximity set)},$$
$$\mathcal{P}_2 = \{P \in \mathcal{P} : d_L(P, P_0) \geq \lambda_2\} \text{ (deviation set)}.$$

Here we notice that the set $\mathcal{P}_1$ is identical to that of the robust universal hypothesis testing. Indeed, the detector developed for the robust universal hypothesis testing applies to the deviation detection problem under the asymptotic NP criterion. Proposition 3 shows that the generalized empirical likelihood ratio test is also optimal for the deviation detection problem, whose proof parallels that of 3.3.3.

Proposition 3. *In the deviation detection problem, for any given $\delta > 0$, detector*

$$D(\hat{\mu}_n||\mathcal{P}_1) \underset{H_1}{\overset{H_2}{\gtrless}} \eta, \tag{3.54}$$

is optimal among all Ω such that $\Omega_2(n) = \Omega_2$ and Ω_2 is open, i.e., let

$$\Lambda_2(n) = \Lambda_2 := \{\mu : D(\mu||\mathcal{P}_1) \geq \eta\}, \;\; \Lambda_1(n) := \Lambda_1 = \{\mu : D(\mu||\mathcal{P}_1) < \eta\},$$

then,

1. $\inf_{P_1 \in \mathcal{P}_1} J^{P_1}(\Lambda) = \eta$.
2. $\inf_{P_2 \in \mathcal{P}_2} I^{P_2}(\Lambda) = D(\Lambda_1||\mathcal{P}_2)$,
3. *For Ω with $\Omega_2(n) = \Omega_2$ and Ω_2 open, Λ is optimal, i.e., if*

$$\inf_{P_1 \in \mathcal{P}_1} J^{P_1}(\Omega) > \eta, \tag{3.55}$$

then,

$$\inf_{P_2 \in \mathcal{P}_2} I^{P_2}(\Omega) \leq \inf_{P_2 \in \mathcal{P}_2} I^{P_2}(\Lambda). \tag{3.56}$$

3.5 Summary

This chapter deals with two binary composite hypothesis testing problem: the robust universal detection problem and the deviation detection problem. For robust detection under the minimax NP criterion, the uncertainty sets under the two hypotheses are not allowed to be arbitrarily close to each other as measured by the L$\acute{e}$vy metric. Otherwise, the problem becomes degenerate. This is illustrated using the moment constrained hypothesis testing, where a well-defined robust detection problem in the discrete case is shown to be degener-

ate for the continuous case.

The choice of the L$\acute{e}$vy metric in defining the proximity for both universal detection and deviation detection problems is due to the following. The Lévy metric is the weakest distance metric hence the uncertainty set thus defined is the largest, in the sense that it includes distributions that are close to the nominal distribution using any other metrics. For example, with the classical KL divergence, it is shown that the closure of the surface of a distribution defined by a KL divergence ball is equivalent to the entire KL divergence ball, making it unsuitable to define the proximity set in both detection problems as well as the deviation set in the deviation detection problem.

The root reason that the KL divergence is unsuitable is because of the discontinuity of the KL divergence for continuous valued random variables where the continuity is defined with respect to the weak convergence. This is the motivation for defining a robust version of the KL divergence in Chapter 2 where the KL divergence is with respect to a L$\acute{e}$vy ball instead of a single distribution. Its continuity with respect to the weak convergence helps the development of the generalized empirical likelihood ratio test that is shown to be optimal, for both the robust universal hypothesis testing and deviation detection. We demonstrate the advantages of the generalized empirical likelihood ratio test over the existing approach developed by Zeitouni and Gutman.

In the next chapter, we will discuss the computation and estimation of the robust KL divergence.

CHAPTER 4
COMPUTATION AND ESTIMATION OF THE ROBUST KL DIVERGENCE

4.1 Computation of the Robust KL Divergence

Chapter 3 established that the empirical robust KL divergence $D(\hat{\mu}_n||B_L(P_0,\delta_0))$ is the optimal statistic, for both the robust universal hypothesis testing and the deviation detection problem, under the asymptotic minimax NP criterion. The proposed detector is also much easier to implement and attains a stronger optimality compared with that proposed in [14]. This statement is made concrete in the present chapter where we develop an efficient procedure for evaluating the empirical robust KL divergence.

Given n samples from a distribution μ, the detection statistic amounts to evaluating the empirical robust KL divergence

$$D(\hat{\mu}_n||B_L(P_0,\delta_0)) = \inf_{P\in B_L(P_0,\delta_0)} D(\hat{\mu}_n||P) \tag{4.1}$$

Without loss of generality, we order the n samples in ascending order and denote the ordered samples as $(x_0, x_1, \cdots, x_{n-1})$. Furthermore, for $P \in B_L(P_0,\delta_0)$ denote by $y_i =$

$P(x_i)$ and $y_{-1} = 0$. As such, searching the optimal P in problem (4.1) is reduced to searching the optimal $\mathbf{y}$ in the following optimization problem,

$$\underset{\mathbf{y}}{\text{minimize}} \quad \sum_{i=0}^{n-1} \frac{1}{n} \log \frac{1/n}{y_i - \max(y_{i-1}, l_i)} \tag{4.2a}$$

$$\text{s.t.} \quad \mathbf{l}' \preceq \mathbf{y} \preceq \mathbf{u}', \tag{4.2b}$$

where $l_i' = \max(P_0(x_i - \delta_0) - \epsilon_0, 0), u_i' = \min(P_0(x_i + \delta_0) + \delta_0, 1)$.

Problem (4.2) is not a convex optimization problem. To transform (4.2) to a convex optimization problem, we introduce y_{n+i} and the condition $y_{n+i} \leq y_i - \max(y_{i-1}, l_i')$, which is equivalent to

$$y_i - y_{n+i} - y_{i-1} \geq 0 \quad \text{and} \quad y_i - y_{n+i} \geq l_i'. \tag{4.3}$$

Notice that to achieve the minimum, y_{n+i} has to equal $y_i - \max(y_{i-1}, l_i)$. Therefore we can modify the problem (4.2) to the following convex optimization problem,

$$\underset{\mathbf{y}}{\text{minimize}} \quad \sum_{i=n}^{2n-1} \frac{1}{n} \log \frac{1/n}{y_i} \tag{4.4a}$$

$$\text{s.t.} \quad \mathbf{l} \preceq \mathbf{y} \preceq \mathbf{u}, \tag{4.4b}$$

$$\mathbf{l}_c \preceq A\mathbf{y}, \tag{4.4c}$$

where

$$\mathbf{l}^t = (l_0', \cdots, l_{n-1}', 0, \cdots, 0), \mathbf{u}^t = (u_0', \cdots, u_{n-1}', \infty, \cdots, \infty),$$

$$A = \begin{bmatrix} 1 & & & & -1 & & & \\ -1 & 1 & & & & -1 & & \\ & \ddots & \ddots & & & & \ddots & \\ & & -1 & 1 & & & & -1 \\ 1 & & & & -1 & & & \\ & 1 & & & & -1 & & \\ & & \ddots & & & & \ddots & \\ & & & 1 & & & & -1 \end{bmatrix}, \mathbf{l}_c = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ l'_0 \\ l'_1 \\ \vdots \\ l'_{n-1} \end{bmatrix}.$$

This is a convex optimization problem with separable convex objective functions and linear constraints, thus numerical solutions can be readily obtained via standard convex program.

4.2 Estimation of the Robust KL divergence

The KL divergence is hard to estimate for the continuous case. Estimating the KL divergence includes as a special case the problem of estimating entropy and mutual information. Numerous methods exist [45–49], most of them estimating the densities or likelihood ratio first and then computing the KL divergence using the estimated distributions. In recent years, direct KL divergence estimation has been developed including methods based on empirical CDF, k-nearest neighbors density estimation or variational characterization of divergences. The robust KL divergence provides a natural way to estimate the KL divergence. We will discuss two cases, one with P_0 known, the other P_0 unknown.

4.2.1 P_0 is known

Naturally, $D(\hat{\mu}_n||B_L(P_0, \delta_0))$ can be viewed as an estimate of $D(\mu||B_L(P_0, \delta_0))$. The question is whether $D(\hat{\mu}_n||B_L(P_0, \delta_0))$ converges to $D(\hat{\mu}||B_L(P_0, \delta_0))$ as the sample size increases. The answer is stated below.

Proposition 4. *Given $P_0 \in \mathcal{P}$, if $P_0(t)$ is continuous in t, $D(\hat{\mu}_n||B_L(P_0,\delta_0)) \xrightarrow{a.s.} D(\mu||B_L(P_0,\delta_0))$.*

Proof. Denote $C_\mu \subseteq \mathcal{R}$ as the continuity set of $\mu(t)$. $D(\mu||B_L(P_0,\delta_0))$ is continuous in μ. If $\mu_n \xrightarrow{w} \mu$, then $\lim_{n\to\infty} D(\hat{\mu}_n||B_L(P_0,\delta_0)) = D(\mu||B_L(P_0,\delta_0))$, therefore,

$$\begin{aligned}
& \Pr\left(\lim_{n\to\infty} D(\hat{\mu}_n||B_L(P_0,\delta_0)) = D(\mu||B_L(P_0,\delta_0))\right) \\
\geq\ & \Pr\left(\hat{\mu}_n \xrightarrow{w} \mu\right) && (4.5)\\
=\ & \Pr\left(\lim_{n\to\infty} \hat{\mu}_n(t) = \mu(t),\ \text{for all } t \in C_\mu\right) && (4.6)\\
=\ & 1 - \Pr\left(\lim_{n\to\infty} \hat{\mu}_n(t) \neq \mu(t),\ \text{for some } t \in C_\mu\right) && (4.7)\\
\geq\ & 1 - \sum_{t\in C_\mu} \Pr\left(\lim_{n\to\infty} \hat{\mu}_n(t) \neq \mu(t)\right) && (4.8)\\
=\ & 1. && (4.9)
\end{aligned}$$

The last equality comes from the fact that for any $t \in C_\mu$, $\mu_n(t) \xrightarrow{a.s.} \mu(t)$. □

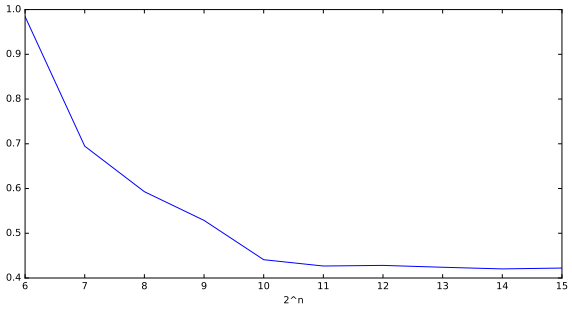
In the discrete case, $D(\hat{\mu}_n||P_0)$ converges to $D(\mu||P_0)$ almost surely. This is, however, not the case for the continuous case. A remedy is to replace P_0 with $B_L(P_0,\delta_0)$, which ensures convergence of the estimate of the robust version of the KL divergence. Some simulation of $D(\hat{\mu}_n||B_L(P_0,\delta_0))$ as n increases is given in the Fig.4.1.

We notice that $D(\mu||B_L(P_0,\delta_0)) \to D(\mu||P_0)$ as $\delta_0 \to 0$. For a fixed δ_0, the KL divergence and the robust KL divergence are related by the following equation.

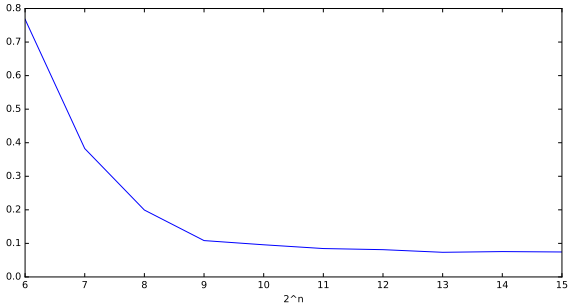
$$D(\mu||P_0) = D(\mu||B_L(P_0,\delta_0)) + [D(\mu||P_0) - D(\mu||B_L(P_0,\delta_0))]. \quad (4.10)$$

The first part of the right hand side is bounded and continuous; the second part is always positive.

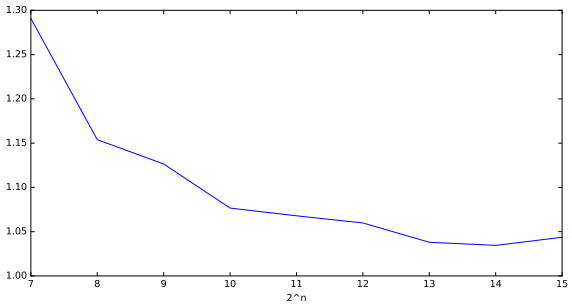
Loosely speaking, if μ and P_0 are both "smooth", for example, they belong to the exponential family distributions, then for small δ_0, $D(\mu||B_L(P_0,\delta_0))$ is close to $D(\mu||P_0)$. With small perturbation on μ, $D(\mu||P_0)$ may increase by an arbitrarily large value. In this case the number of samples needed to accurately estimate the resulting KL divergence



(a) Estimate of $D(\mathcal{N}(0,1)||B_L(\mathcal{N}(1,1),0.01))$.



(b) Estimate of $D(\mathcal{N}(0,2)||B_L(\mathcal{N}(0,1),0.01))$.



(c) Estimate of $D(exp(1)||B_L(\mathcal{N}(3,4),0.01))$.

Fig. 4.1: Estimate of the robust KL.

also increases tremendously. One such example can be found in [49]. The robust KL divergence is insensitive to such perturbation and is able to capture the "smooth" part of the KL divergence.

4.2.2 P_0 is unknown

We now discuss the two samples problem, i.e., estimating $D(\mu||B_L(P_0, \delta_0))$ given the samples generated from both μ and P_0 in the absence of the knowledge of the actual distributions. First, we have the following Theorem.

Theorem 4.2.1. *Given μ, $D(\mu||B_L(P_0, \delta_0))$ is continuous at P_0 with respect to weak convergence, provided $P_0(t)$ is continuous in t,*

Proof. Assume $P_m \xrightarrow{w} P_0$ as $m \to \infty$, we will show that $D(\mu||B_L(P_m, \delta_0)) \to D(\mu||B_L(P_0, \delta_0))$. For any m, there exists a $P_m^* \in B_L(P_m, \delta_0)$ such that $D(\mu||P_m^*) = D(\mu||B_L(P_m, \delta_0))$. For any $\delta > 0$, there exists an M such that $d_L(P_m, P_0) \leq \delta$ for $m \geq M$. We have,

$$\begin{aligned} \underline{\lim}_{m\to\infty} D(\mu||B_L(P_m, \delta_0)) &= \underline{\lim}_{m\to\infty} D(\mu||P_m^*) && (4.11) \\ &\geq D(\mu||B_L(P_0, \delta_0 + \delta)). && (4.12) \end{aligned}$$

The last inequality holds since

$$\begin{aligned} d_L(P_m^*, P_0) &\leq d_L(P_m^*, P_m) + d_L(P_m, P_0) && (4.13) \\ &\leq \delta_0 + \delta. && (4.14) \end{aligned}$$

Notice that (4.12) is true for any δ, then from Lemma 2.3.2, $D(\mu||B_L(P_0, \delta_0))$ is continuous at δ_0, therefore,

$$\underline{\lim}_{m\to\infty} D(\mu||B_L(P_m, \delta_0)) \geq D(\mu||B_L(P_0, \delta_0)). \quad (4.15)$$

On the other hand,

$$\overline{\lim_{m\to\infty}} D(\mu||B_L(P_m,\delta_0)) \le D(\mu||B_L(P_0,\delta_0-\delta)), \tag{4.16}$$

where the inequality holds because $B_L(P_0,\delta_0-\delta) \subseteq B_L(P_m,\delta_0)$ for $m \ge M$. Equation (4.16) holds for any δ. Since $D(\mu||B_L(P_0,\delta_0))$ is continuous at δ_0, we have,

$$\overline{\lim_{m\to\infty}} D(\mu||B_L(P_m,\delta_0)) \;\le\; D(\mu||B_L(P_0,\delta_0)), \tag{4.17}$$

From (4.15) and (4.17), we have

$$\lim_{m\to\infty} D(\mu||B_L(P_m,\delta_0)) = D(\mu||B_L(P_0,\delta_0)). \tag{4.18}$$

□

A direct result of the above theorem is the parallel statement of Proposition 4 for P_0, i.e., for the empirical distribution $\hat{P}_m \sim P_0$, $D(\mu||B_L(\hat{P}_m,\delta_0))$ converges to $D(\mu||B_L(P_0,\delta_0))$ almost surely.

Previously we have proved $D(\mu||B_L(P_0,\delta_0))$ is continuous at μ if $P_0(t)$ is continuous in t. Combine with the above Theorem, we have the following Proposition.

Proposition 5. *Assume $\mu_n \xrightarrow{w} \mu$, $P_m \xrightarrow{w} P_0$ and $P_0(t)$ is continuous in t. Then*

$$\lim_{n\to\infty}\lim_{m\to\infty} D(\mu_n||B_L(P_m,\delta_0)) = D(\mu||B_L(P_0,\delta_0)).$$

Notice the above is different from $\lim_{n,m\to\infty} D(\mu_n||B_L(P_m,\delta_0)) = D(\mu||B_L(P_0,\delta_0))$. The latter is a much stronger condition and if true, then given the empirical distributions $\hat{\mu}_n$ and $\hat{P}_m$, the estimate $D(\hat{\mu}_n||B_L(\hat{P}_m,\delta_0))$ will converge to $D(\mu||B_L(P_0,\delta_0))$ almost surely.

4.3 Summary

This chapter first investigates the computation of the robust KL divergence between an empirical distribution and a L$\acute{e}$vy ball, which is the asymptotic optimal detection statistic for both robust universal hypothesis testing and deviation detection. The problem is converted to a convex optimization problem and can be readily solved via standard convex programs. Furthermore, estimation of the robust KL divergence is considered and the constructed estimate is shown to converge almost surely, when either one of the two distributions is known. For the case that two sequences of samples are given for each distributions, i.e., when both distributions are unknown, stronger convergence results are desired which will be left as future work.

CHAPTER 5
TO LISTEN OR NOT: DISTRIBUTED DETECTION WITH ASYNCHRONOUS TRANSMISSIONS

5.1 Introduction

Distributed detection has been a well studied topic in the past few decades [10,35,36]. Most existing results assume either a parallel structure where all sensors propagate their local data/decisions to a fusion center (FC), or a tandem network where sensors are connected in a serial manner and the last node becomes the FC. Noteworthy exceptions include that of tree structures and directed acycylic topologies as studied in [37–40].

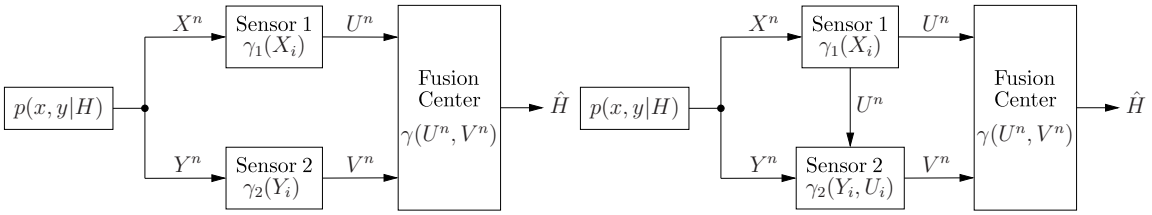
We consider a variation of the parallel fusion system that is largely motivated by the broadcast nature of the wireless transmission. While it is typically assumed that the fusion center implements a mapping (decision rule) that takes inputs from all the sensors, communications from sensors to the FC often occur asynchronously (e.g., in a traditional TDM - time division multiplexing system or in an Aloha type of random access system). As such, it bears the question of whether sensors should perhaps take advantage of the asynchronism

by listening to transmissions from other sensors before deciding what to transmit. Such an ability of overhearing other sensors' transmissions is made possible given the broadcast nature of wireless transmission. Using a two sensor system as an illustration, our task is to compare the performance between the two systems schematically shown in Fig. 1. At each time, local sensors make decisions based on current observations with or without aid of other sensors' output, where (a) is the classical parallel system and (b) is one where overhearing occurs.

It is clear in Fig. 1 that (b) should perform no worse than (a). One may also view (b) as subsuming a two-sensor serial system, which is well known to perform no worse than the parallel system with identical observations [35]. To see that this is true, we note that the system in (b) can reproduce the final decision of a serial system when sensor 2 serves as a fusion center, i.e., let sensor 2 make the final decision and send it to the fusion center as its final decision, thus the input U from sensor 1 is essentially ignored in the FC.

This paper is interested in comparing the two systems in Fig. 1 for the large sample regime, i.e., when the number of samples become large. Notice that the asymptotics is with respect to the number of observations as opposed to the number of sensors (i.e., network size) as studied in, for example, [41]. This will become clear in Section 5.2. While it can be easily shown that with fixed sample size, overhearing may strictly outperform the classical parallel system (c.f. Appendix), we show in this letter that for the large sample regime, the performance comparison largely depends on the observation model. For conditionally independent observations, we show that there is no difference in asymptotic detection performance between the two systems. However, for conditionally dependent observations, it is possible that strict performance improvement is attained through the overhearing scheme.

The overhearing system allows one of the sensors to have access to side information (in the form of the other sensor's output) in addition to its own observation. There have been other forms of side information studied in the literature in decentralized detection. The unlucky broker problem, considered in [42], involves decision making where some



(a) Parallel topology (b) With overhearing

Fig. 5.1: Distributed network

sensors have access to the decision output of an initial fixed detector. Separately in [43], an interactive decentralized detection scheme was considered where a peripheral sensor takes input from the other sensor (which also serves as a fusion center) before sending its decision back to the fusion center where a final decision is made. A key distinction between the overhearing scheme and the interactive detection is in the system model: the fusion center in the overhearing scheme has quantized decision output from the two sensors, with one of them making decision based on its observation as well as the same output from the other sensor; for the interactive scheme, the fusion center has access to its own observation as well as the output of the peripheral sensor.

5.2 Problem Statement

We consider throughout this letter a two-sensor system and the asymptotics is taken in the time domain, i.e., when the two sensors observe a sequence of observations. Generalization to a system involving multiple sensors will be briefly discussed wherever applicable.

To be more specific, in reference to Fig. 1, let the sensor observations be $\{(X_i, Y_i) : i = 1, \cdots, n\}$. The observations are assumed to be independent and identically distributed (i.i.d.) in time i, i.e.,

$$P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} P_{XY}(x_i, y_i),$$

where $P_{XY}(x, y)$ is the joint distribution function of (X, Y) on the sample space $(\mathcal{X}, \mathcal{Y})$.

The two hypotheses under test are defined by the joint distribution of (X, Y):

$$P_{XY}(x,y) = \begin{cases} P^0_{XY}(x,y) & \text{if } H_0 \text{ is true;} \\ P^1_{XY}(x,y) & \text{if } H_1 \text{ is true.} \end{cases}$$

The FC, however, does not have direct access to the entire observation. Instead, at time i, the observations (X_i, Y_i) are quantized to (U_i, V_i) where $U_i = \gamma_1(X_i)$, and

$$V_i = \begin{cases} \gamma_2(Y_i) & \text{for system (a);} \\ \gamma_2(Y_i, U_i) & \text{for system (b).} \end{cases}$$

Thus in system (b), the output of sensor 2 not only depends on its observation Y_i, but also depends on U_i, the output of sensor 1 at time i. The fusion center takes the sequence (U_i, V_i), $i = 1, \cdots, n$ and makes the final decision

$$\hat{H} = \gamma_0(\mathbf{U}, \mathbf{V})$$

where $\mathbf{U} = (U_1, \cdots, U_n)$ and $\mathbf{V} = (V_1, \cdots, V_n)$. The objective is to determine if the overhearing scheme is superior in detection performance as the number of samples n grows to infinity. As per Chernoff-Stein Lemma [5], we use the KL divergence as the asymptotic performance metric in our study. For probability measures P and Q defined over the sample space Ω, KL divergence between them is given by

$$D(P||Q) = \int_\Omega \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P.$$

We assume for ease of presentation that all sensor outputs are binary; our result can be easily extended to multi-bit quantization.

Therefore, the objective is to find, for both systems in Fig.1, γ_1^* and γ_2^* that maximize $D(P^0_{UV}||P^1_{UV})$, and to compare the maximum achievable KL divergence between the two

systems.

For the parallel topology in Fig. 1(a), (γ_1, γ_2) can be equivalently characterized using two binary partitions

$$R_x = \{x : \gamma_1(x) = 0\},\ R_y = \{y : \gamma_2(y) = 0\}.$$

Thus, the optimum KL divergence for the parallel system can be equivalently defined as

$$Q_{xy} := \max_{R_x, R_y} D(P^0_{UV} || P^1_{UV}). \tag{5.1}$$

With overhearing, i.e., the system described in Fig. 1(b), where γ_2 has both Y and output of quantizer γ_1 as its input, an equivalent characterization of (γ_1, γ_2) is

$$R_x = \{x : \gamma_1(x) = 0\},$$

$$R_{0y} := \{y : \gamma_2(y, 0) = 0\},\ \ R_{1y} := \{y : \gamma_2(y, 1) = 0\}.$$

The corresponding optimum KL divergence for the overhearing system is therefore

$$\tilde{Q}_{xy} := \max_{R_x, R_{0y}, R_{1y}} D(P^0_{UV} || P^1_{UV}). \tag{5.2}$$

It is clear that $\tilde{Q}_{xy} \geq Q_{xy}$. The question we attempt to answer is whether strict improvement in asymptotic performance is possible, i.e., if $\tilde{Q}_{xy} > Q_{xy}$ can be true for some observation models.

We show in the next two sections that the answer to the above problem depends on the observation model. For conditionally independent observations, i.e., for (X, Y) that satisfy

$$P^i_{XY}(x, y) = P^i_X(x) P^i_Y(y), \quad i = 0, 1,$$

overhearing does not provide asymptotic detection performance improvement. For conditionally dependent observations, however, we show through a simple example that asymptotic detection performance improvement in terms of KL divergence is indeed possible.

We now introduce a lemma [44] that shows, for a binary hypothesis test, a monotone likelihood ratio (LR) quantizer maximizes KL divergence of the quantizer output.

Lemma 5.2.1. *[44] For a random variable W with distribution P_W^i under H_i, $i = 1, 2$, maximum of $D(P^0_{\gamma(W)}||P^1_{\gamma(W)})$ over 1-bit quantizer $\gamma(\cdot)$ is achieved only by a single threshold LR quantizer. A quantizer $\gamma(\cdot)$ is said to be a single threshold LR quantizer if*

$$\gamma(w_1) = \gamma(w_2) \Leftrightarrow (L_W(w_1) - \tau)(L_W(w_2) - \tau) \geq 0 \tag{5.3}$$

for some τ, where $L_W(\cdot)$ is log LR function of W.

Notation used in this letter: For simplicity we use P_A^i and $P_{A|B}^i$ to denote distribution function of A and conditional distribution function of A given B under H_i. Furthermore, define

$$d_x(R_x) = P^0(X \in R_x) \log_2 \frac{P^0(X \in R_x)}{P^1(X \in R_x)} + P^0(X \notin R_x) \log_2 \frac{P^0(X \notin R_x)}{P^1(X \notin R_x)}, \tag{5.4}$$

$$d_y(R_y) = P^0(Y \in R_y) \log_2 \frac{P^0(Y \in R_y)}{P^1(Y \in R_y)} + P^0(Y \notin R_y) \log_2 \frac{P^0(Y \notin R_y)}{P^1(Y \notin R_y)}, \tag{5.5}$$

$$\begin{aligned} & d_{y|x}(R_y, R_x) \qquad (5.6) \\ = \;& P^0(Y \in R_y|X \in R_x) \log_2 \frac{P^0(Y \in R_y|X \in R_x)}{P^1(Y \in R_y|X \in R_x)} + P^0(Y \notin R_y|X \in R_x) \log_2 \frac{P^0(Y \notin R_y|X \in R_x)}{P^1(Y \notin R_y|X \in R_x)}. \end{aligned}$$

These quantities will be used in evaluating KL divergence under different distribution models.

5.3 Distributed Detection with Asynchronous Transmissions

5.3.1 Conditionally independent observations

For the parallel system, conditional independence between X and Y leads directly to conditional independence between U and V as they are respectively independent functions of X and Y. Therefore,

$$\begin{aligned} Q_{xy} &= \max_{R_x,R_y} D(P^0_{UV}||P^1_{UV}) && (5.7)\\ &= \max_{R_x} D(P^0_U||P^1_U) + \max_{R_y} D(P^0_V||P^1_V) \\ &= d_x(R^*_x) + d_y(R^*_y). && (5.8) \end{aligned}$$

From Lemma 5.2.1, the optimal R^*_x, R^*_y are given by

$$R^*_x = \{x : L_X(x) > \tau^*_X\}, \tag{5.9}$$

$$R^*_y = \{y : L_Y(y) > \tau^*_Y\}, \tag{5.10}$$

for some τ^*_X, τ^*_Y.

For the model described in Fig. 1(b) where sensor 2 makes its own decision based on both Y and U, it may appear that there is potential improvement in terms of KL divergence as compared with the parallel case. In particular, it is apparent that the U and V are no longer conditionally independent as V explicitly depends on U. We show, however, in Proposition 2 that such an overhearing scheme does not improve the asymptotic detection performance, i.e., it attains the same maximum KL divergence as with the parallel case.

Proposition 6. *If (X, Y) are conditionally independent, then $Q_{xy} = \tilde{Q}_{xy}$.*

Proof. Denote by R_x^c complement set of R_x,

$$\begin{aligned}
\tilde{Q}_{xy} &= \max_{R_x,R_{0y},R_{1y}} D(P_{UV}^0||P_{UV}^1) && (5.11)\\
&= \max_{R_x,R_{0y},R_{1y}} [D(P_U^0||P_U^1) + D(P_{V|U}^0||P_{V|U}^1)] \\
&= \max_{R_x,R_{0y},R_{1y}} [d_x(R_x) + P^0(X \in R_x) d_{y|x}(R_{0y}, R_x) + \\
&\qquad P^0(X \in R_x^c) d_{y|x}(R_{1y}, R_x^c)]. && (5.12)
\end{aligned}$$

Note that for $i \in \{0, 1\}$,

$$P^i(Y \in R_{0y}|X \in R_x) = P^i(Y \in R_{0y}),$$

$$P^i(Y \in R_{1y}|X \in R_x^c) = P^i(Y \in R_{1y}),$$

due to the conditional independence. From (5.5) and (5.6), we have

$$d_{y|x}(R_{0y}, R_x) = d_y(R_{0y}),$$

$$d_{y|x}(R_{1y}, R_x^c) = d_y(R_{1y}).$$

Therefore,

$$\begin{aligned}
\tilde{Q}_{xy} &= \max_{R_x,R_{0y},R_{1y}} [d_x(R_x) + P^0(X \in R_x) d_y(R_{0y}) + P^0(X \in R_x^c) d_y(R_{1y})] && (5.13)\\
&\leq \max_{R_x} [d_x(R_x) + P^0(X \in R_x) d_y(R_y^*) + P^0(X \in R_x^c) d_y(R_y^*)] && (5.14)\\
&= \max_{R_x} [d_x(R_x) + d_y(R_y^*)] && (5.15)\\
&= Q_{xy}. && (5.16)
\end{aligned}$$

Combined with $\tilde{Q}_{xy} \geq Q_{xy}$, we have $\tilde{Q}_{xy} = Q_{xy}$. □

Remark 3. *The above result can be generalized to k sensors with each having $D > 2$ quantization levels. The ith sensor takes as input its own observation and output of sensors*

$1, 2, \ldots, i-1$. *The result also holds if we allow random quantizers for the same reason as that of [44] for the centralized case.*

The above result about the asymptotic performance is in contrast to that of fixed sample size test where, as shown in the Appendix, the overhearing scheme may strictly outperform the parallel system for conditionally independent observations.

5.3.2 Conditionally dependent observations

Consider the following example where both X and Y are ternary random variables with sample space $\{1, 2, 3\}$. Let

$$A = \begin{pmatrix} (1-\epsilon)/9 & (1+\epsilon)/9 & 1/9 \\ (1+\epsilon)/9 & (1-\epsilon)/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix},$$

with $0 < |\epsilon| \leq 1$, represent joint probability distribution of X and Y. The two hypotheses under test are respectively

$$H_0 : \epsilon = \epsilon_0, \;\; H_1 : \epsilon = \epsilon_1.$$

Under each hypothesis, (X, Y) are dependent of each other. Let $(\epsilon_0, \epsilon_1) = (-1, d)$, $D(P^0_{XY}||P^1_{XY}) = 4/9 - (4/9)\log_2(1-d)$ is an increasing function of d.

It is easy to show that local LR quantizers are degenerate (i.e., the output is a constant), which can never achieve optimal KL divergence. To see this, we note that the marginal distributions under H_0 and H_1 are identical to uniform distribution, hence marginal LR is constant 1. The fact that local LR quantizer is no longer optimum for maximum KL divergence is due to the conditional dependence between X and Y under each hypothesis. For this discrete example, however, one can compute the maximum KL divergence for both systems through exhaustive search. For example , for the parallel system, each quantizer amounts to a mapping from the ternary alphabets to a binary one.

The result is plotted in Fig. 5.2 where Q_{xy}, $\tilde{Q}_{xy}$ are plotted together with the difference of the two (scaled by a factor of 10) as a function of d. It is apparent from the figure that strict improvement in asymptotic detection performance is possible when d exceeds a certain threshold.

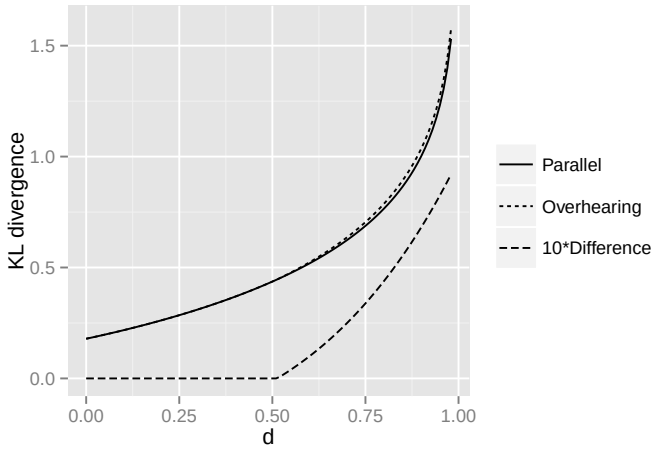


Fig. 5.2: Comparision betweent Q_{xy} and $\tilde{Q}_{xy}$ of conditionally dependent example in Section 5.3.2.

5.4 Examples for fixed sample size test

We use two examples, one discrete, one continuous, to show that strict detection performance improvement is possible for the overhearing scheme for the finite sample size test. We use the receiver operating characteristic (ROC) curve for performance comparison. Binary quantizers are assumed throughout the examples. For independent observations under each hypothesis, the optimal binary quantizers are local LR single threshold quantizers for both parallel and overhearing systems [35].

5.4.1 A discrete example

Let X and Y be independent ternary random variables with identical sample space $\{1, 2, 3\}$. The distributions of the pair under the two hypotheses are given in Tab. 5.1.

	1	2	3
P_X^0	0.73	0.02	0.25
P_Y^0	0.50	0.17	0.33

	1	2	3
P_X^1	0.67	0.05	0.28
P_Y^1	0.75	0.12	0.13

Table 5.1: The distributions of X and Y.

Since X and Y are independent under each hypothesis, for fusion rules AND and OR, we can get their respective ROC curves by exhausting all local binary LR quantizers. Fig. 5.3 shows ROC curves under fusion rules AND and OR. Similarly, the ROC curve of the overhearing system can be attained via an exhaustive search. It turns out that for this example, the overhearing scheme has a ROC curve that achieves the same detection performance as the parallel system provided that the parallel system uses the better of the two fusion rules, i.e., the ROC curve of the overhearing scheme is the same as the concave envelope of the ROC curves for the parallel system with fusion center implementing AND and OR rules.

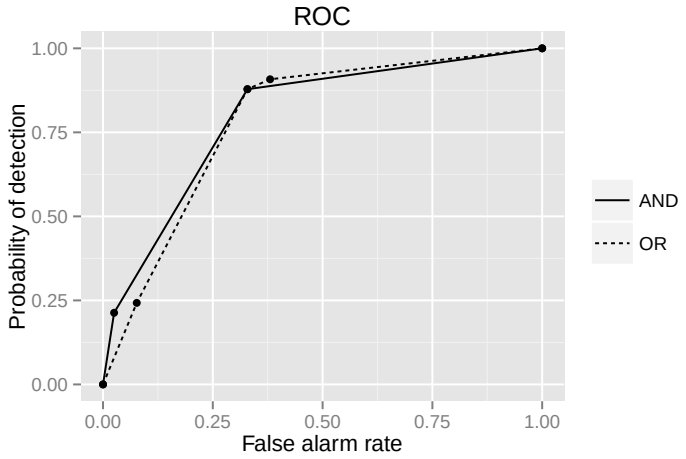


Fig. 5.3: The ROC curves for the discrete example.

5.4.2 A continuous example

Consider the detection of the shift in the mean value of Gaussian observations where the two hypotheses are specified by

$$H_0 : (X, Y) \sim N(0, 0, 1, 1, 0), \quad H_1 : (X, Y) \sim N(a, b, 1, 1, 0).$$

The observations X and Y are independent under each hypothesis, hence, for the parallel system, we only need to determine the optimal thresholds of local LR quantizers, which are given in [35].

As for the overhearing system in Fig. 1(b), let t^1 be the threshold for quantizer 1, t_i^2 be the threshold for quantizer 2 if the output of quantizer 1 is i, $i \in \{0, 1\}$. Define

$$\alpha^1 = Pr(u = 0|H = 1), \alpha_i^2 = Pr(v = 0|u = i, H = 1),$$

$$\beta^1 = Pr(u = 1|H = 0), \beta_i^2 = Pr(v = 1|u = i, H = 0),$$

which are all functions of (t^1, t_0^2, t_1^2). For a given false alarm probability P_f, we can get the optimal (t^1, t_0^2, t_1^2) by solving the following nonlinear equations,

$$\frac{t_0^2\alpha^1}{1-\beta^1} = \frac{t_1^2(1-\alpha^1)}{\beta^1} = \frac{t^1(\alpha_0^2 - \alpha_1^2)}{\beta_1^2 - \beta_0^2}, \tag{5.17}$$

$$P_f = (1-\beta^1)\beta_0^2 + \beta^1\beta_1^2. \tag{5.18}$$

Eqs. (5.17) and (5.18) are derived using the Lagrange multiplier method similar to that of the parallel case [35]. Fig. 5.4 shows the result of three ROC curves. The ROC curve of overhearing system is strictly above the ROC curve of parallel system, the latter is the concave envelope of ROC curves with FC implementing AND or OR. Notice that in this case, the parallel system needs to use dependent randomization to achieve the detection

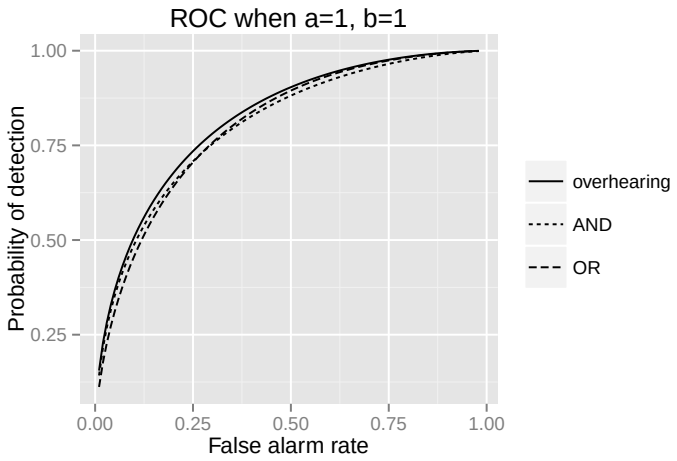


Fig. 5.4: The ROC curves for the continuous example.

performance determined by the concave envelope of that using AND and OR fusion rules.

5.5 Summary

This chapter investigates a variation of the parallel fusion system: sensors may take advantage of the asynchronism by overhearing other sensors' transmissions in the hope of achieving a better detection performance. Using a two sensor system as an illustration, we show that while overhearing may strictly outperform the classical parallel system for the fixed sample size test, in the large sample regime there is no performance gain, as measured by the KL divergence achievable at the fusion center, provided that the observations are conditionally independent. However, for conditionally dependent observations, it is demonstrated that strict asymptotic detection performance improvement can be attained through the overhearing scheme.

CHAPTER 6
CONCLUSION AND FUTURE RESEARCH

6.1 Conclusion

The probability space defined on discrete alphabets, is a compact subspace of the Euclidean space. The distance metrics such as the KL divergence and the total variation are continuous functions of distributions for the discrete case. However, if the probability space is defined on continuous alphabets such as the real line, it becomes much more complicated. Many distance metrics, most notably the KL divergence, become only lower semicontinuous. The consequence of such a lack of continuity is that many of the well established results in hypothesis testing for the discrete case no longer applies to the continuous case. This thesis makes progress toward bridging the gap on two of the hypothesis testing problems: the universal hypothesis testing and deviation detection.

In Chapter 2, we provided a robust version of the KL divergence, which is defined to be the KL divergence between a distribution and the Lévy ball of a known distribution. This robust KL divergence is shown to be continuous with respect to the weak convergence. The use of Lévy metric in the robust KL divergence is due to the fact that the Lévy metric is a true distance metric and is also the most general one: closeness in the Lévy metric implies closeness in every other known metrics. In other words, a Lévy ball centered at a

nominal distribution encompasses the largest set of probability distributions that are close to the nominal distribution. Some of the important properties of the robust KL divergence are identified as follows.

- The robust KL divergence of discretized distributions will converge to the robust KL divergence of the original distributions as the quantization level increases.
- The robust KL divergence is defined as the infimum over the L$\acute{e}$vy ball and the infimum is attained by a distribution inside the L$\acute{e}$vy ball.
- The robust KL divergence is continuous in the radius of the L$\acute{e}$vy ball.
- The robust KL divergence is a convex function.
- The supremum of the robust KL divergence over a L$\acute{e}$vy ball can be achieved by a distribution which is the combination of two distributions that correspond to the lower bound and upper bound of the L$\acute{e}$vy ball defined using the cumulative distribution function.
- The robust KL divergence is bounded and the supremum is a function of the radius of the L$\acute{e}$vy ball and this bound is independent of the actual distributions that define the KL divergence..

The robust KL divergence plays an important role in the robust universal hypothesis testing and deviation detection. Specifically, the continuity property is much desired when constructing robust detectors for those detection problems.

In Chapter 3, we examine the robust universal hypothesis testing and deviation detection, both of them can be considered special cases of the general framework of: robust detection. For robust detection under the minimax NP criterion, we have shown that the uncertainty sets under the two hypotheses should not to be arbitrarily close to each other as measured by the L$\acute{e}$vy metric. Otherwise, the detection problem becomes degenerate with respect to the minimax NP criterion. This result is redundant for the discrete case, as

the probability space is simple and the uncertainty sets under two hypotheses are usually characterized by continuous functions. As such, disjoint sets, which is required in specifying the uncertainty sets under the two hypotheses, are naturally separated in terms of the L$\acute{e}$vy metric. However, in the continuous case, this subtle requirement is often overlooked. We illustrate such a situation using a moment constrained testing problem, and we demonstrate that additional assumptions are needed on the moment constraints for the problem to become meaningful in the continuous case.

The above observation led to a more sensible formulation of the two hypothesis testing problems. Specifically, for the robust universal hypothesis testing, the two hypotheses are characterized respectively by a proximity set of the nominal distribution, defined using the L$\acute{e}$vy ball, and an unknown distribution. In deviation detection, the binary hypotheses are characterized by a proximity set and a deviation set of the nominal distribution, again, defined using the L$\acute{e}$vy metric. To be more concrete, we show that the classical KL divergence is not a suitable metric in defining the proximity set for the hypothesis testing problems: the closure of the surface of distributions that are of constant KL divergence to the nominal distribution becomes the entire KL divergence ball.

The generalized empirical likelihood ratio test, of which the statistic is the robust KL divergence between the empirical distribution and the proximity set defined using the L$\acute{e}$vy metric, is shown to be optimal under the asymptotic minimax NP criterion, for both robust universal hypothesis testing and deviation detection. The key that makes the generalized empirical likelihood ratio test optimal, is the continuity of the robust KL divergence. We also demonstrate the advantages of the generalized empirical likelihood ratio test over the existing approach in terms of its implementation..

In Chapter 4, the computation of the robust KL divergence between an empirical distribution and a known L$\acute{e}$vy ball is converted to a convex optimization problem which can be readily solved using standard convex program. In addition, we have shown that with either one of the two distributions known, the estimate of the robust KL divergence, which is

constructed by directly substituting the empirical distribution for the unknown distribution, converges almost surely to the true value.

In Chapter 5, we study a variation of the parallel fusion system: sensors may listen to transmissions from other sensors before deciding what to transmit in order to improve the inference performance. Using a two-sensor system as an illustration, we showed that for conditionally independent observations, while overhearing may strictly outperform the parallel system for the fixed sample size test, it provides no performance gain in the large sample regime, as measured by the KL divergence obtained at the fusion center. For conditionally dependent observations, however, overhearing can attain strict asymptotic performance improvement.

6.2 Future Work

This thesis has made important contributions to the theory of robust hypotheses testing for continuous valued observations. Our investigation also helps identify some challenging research problems. We list two of the most important ones below.

1. If a distribution $P_0(t)$ is continuous in t, $D(\mu||B_L(P_0,\delta_0))$ is shown to be continuous in μ as well as in P_0 with respect to the weak convergence. Therefore for $\mu_n \xrightarrow{w} \mu$ and $P_m \xrightarrow{w} P_0$, we have

$$\lim_{n\to\infty}\lim_{m\to\infty} D(\mu_n||B_L(P_m,\delta_0)) = D(\mu||B_L(P_0,\delta_0)). \tag{6.1}$$

   This convergence result, however, is not strong enough for the purpose of estimating the robust KL divergence when both μ and P_0 are unknown. In this case, the desired one is the following,

$$\lim_{n,m\to\infty} D(\mu_n||B_L(P_m,\delta_0)) = D(\mu||B_L(P_0,\delta_0)), \tag{6.2}$$

which is much stronger than (6.1). We conjecture that (6.2) is true. Two promising approaches to proving (6.2) are as follows.

- Given $P_0(t)$ is continuous in t, establish that $D(\mu||B_L(P_0, \delta_0))$ is uniformly continuous in μ.
- Prove

$$\lim_{m\to\infty} \lim_{n\to\infty} D(\mu_n||B_L(P_m, \delta_0)) = D(\mu||B_L(P_0, \delta_0)), \tag{6.3}$$

  then combine with equation (6.1), one can show that equation (6.2) holds. The difficulty is that $D(\mu_n||B_L(P_m, \delta_0))$ does not converge to $D(\mu||B_L(P_m, \delta_0))$ as $n \to \infty$, since $P_m(t)$ may not be continuous in t. However, the conjecture that (6.2) holds is due to the following intuition. The difference between $\lim_{n\to\infty} D(\mu_n||B_L(P_m, \delta_0))$ and $D(\mu||B_L(P_m, \delta_0))$ is proportional to the largest jump of $P_m(t)$. As $P_m \xrightarrow{w} P_0$ such jump diminishes since $P_0(t)$ is itself continuous, therefore the difference diminishes as well, leading to our conjecture that (6.3) holds.

If equation (6.2) is indeed true, then given the empirical distributions $\hat{\mu}_n$ and $\hat{P}_m$, the estimate $D(\hat{\mu}_n||B_L(\hat{P}_m, \delta_0))$ will converge to $D(\mu||B_L(P_0, \delta_0))$ almost surely.

2. Estimating the classical KL divergence is known to be difficult for the continuous case. This thesis reveals that the KL divergence may vary arbitrarily with small perturbation on the distributions. On the other hand, estimating the robust KL divergence for the continuous case is quite straightforward, at least in the case when one of the two distributions is known. Additionally, we have shown that the robust KL divergence is a lower bound of the KL divergence. A more precise characterization of the relation between the estimate of the KL divergence and that of the robust KL divergence will be needed in order to shed light on the problem of estimating the

classical KL divergence.

REFERENCES

[1] P. J. Huber and V. Strassen, "Minimax tests and the neyman-pearson lemma for capacities," *Ann. Statist.*, vol. 1, no. 2, pp. 251-263, Mar. 1973.

[2] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, no. 2, pp. 369-401, Apr. 1965.

[3] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statistics*, 22(1): 79-86, 3 1951.

[4] I. Csiszár and P. C. Shields, "Information theory and statistics: a tutorial," *Commun. Inf. Theory*, vol. 1, no. 4, pp. 417-528, 2004.

[5] T. M. Cover and J. A. Tomas, *Elements of information theory*, Wiley-Interscience, 1991.

[6] A. Dembo and O. Zeitouni, *Large deviation techniques and applications*, second edition, Springer, 1998.

[7] B. Póczos and J. G. Schneider, "On the estimation of alpha-divergences," *Proc. International Conference on Artificial Intelligence and Statistics*, pp. 609-617, 2011.

[8] J. B. Oliva, B. Póczos and J. Schneider, "Distribution to distribution regression", *Proc. International Conference on Machine Learning*, pp. 1049-1057, 2013.

[9] I. S. Dhillon, S. Mallela and R. Kumar, "A divisive information theoretic feature clustering algorithm for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1265-1287, 2003.

[10] J. N. Tsitsiklis, "Decentralized detection," *In Advances in Statistical Signal Processing*, pp. 297-344, JAI Press, 1993.

[11] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Math. Contr. Signals Syst.*, vol. 1, no. 2, pp. 167-182, 1988.

[12] V. S. S. Nadendla, H. Chen and P. K. Varshney, "Secure distributed detection in the presence of eavesdroppers," *Proc. Conf Signals, Systems and Computers (ASILOMAR)*, the Forty Fourth Asilomar Conf, 2010, pp. 1437-1441

[13] T. Van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory.*, vol. 60, no. 7, pp. 3797-3820, Jul. 2014.

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

[15] P. Billingsley, *Convergence of probability measures*, Wiley, Jul. 1999.

[16] M. Loève, *Probability theory*, second edition, Van Nostrand, Princeton, N.J., 1960. MR 23 #A670.

[17] I. Csiszár, "A simple proof of Sanov's theorem," *Bulletin of the Brazilian Mathematical Society*, vol. 37, no. 4, pp. 453-459, 2006.

[18] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, no. 3, pp. 419-435, Dec. 2002.

[19] P. F. Yang and B. Chen, "Deviation detection with continuous observations," *in Proc. IEEE GlobalSIP*, Orlando, FL, Dec. 2015.

[20] P. Groeneboom, J. Oosterhoff and F. H. Ruymgaart, "Large deviation theorems for empirical probability measures," *Ann. Probab.*, vol. 7, no. 4, pp. 553-586, Aug. 1979.

[21] J. F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Trans. Signal Process.*, vol. 51, pp. 407âĂŞ416, 2003.

[22] R. J. Bolton and D. J. Hand, "Statistical fraud detection: a review," *Statistical Science*, vol. 17, no. 3, pp. 234-255, 2002.

[23] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: a survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1-15:58, Jul. 2009.

[24] L. Lai, H. V. Poor, Y. Xin and G. Georgiadi, "Quickest search over multiple sequences," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5375-5386, Aug. 2011.

[25] P. J. Huber, "Robust confidence limits," *Zeitschrift för Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 10, no. 4, pp. 269-278, 1968.

[26] G. G$\ddot{u}$l and A. M. Zoubir, "Robust hypothesis testing for modeling errors," *Proc IEEE Int Conf Acoust Speech Signal Process (ICASSP)*, pp. 5514-5518, Vancouver, Canada, May. 2013.

[27] G. G$\ddot{u}$l and A. M. Zoubir, "Robust hypothesis testing with α-divergence," *submitted to IEEE Trans. Signal Process.*, [Online]. Available: http://arxiv.org/abs/1501.05019

[28] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 413-421, Jan. 2009.

[29] B. C. Levy, *Principles of signal detection and parameter estimation*, Springer, 2008.

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

[31] C. Pandit, "Robust statistical modeling based on moment classes with applications to admission control, large deviations, and hypothesis testing," *PhD thesis*, University of Illinois at Urbana-Champaign, Urbana, IL, USA, 2004.

[32] W. R. Pestman, "*Mathematical Statistics*," second edition, de Gruyter, Apr. 2009.

[33] Y. Kitamura, "Asymptotic optimality of empirical likelihood for testing moment restrictions," *Econometrica*, vol. 69, no. 6, pp. 1661-1672, 2001.

[34] P. Groeneboom, J. Oosterhoff and F. H. Ruymgaart, "Large deviation theorems for empirical probability measures," *Ann. Probab.*, vol. 7, no. 4, pp. 553-586, Aug. 1979.

[35] P. K. Varshney, "Distributed detection with false alarm rate constraints," *Distributed Detection and Data Fusion*, pp. 179-215, Springer, 1997.

[36] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors I. Fundamentals," *Proc. IEEE*, vol. 85, no. 1, pp. 54-63, Jan. 1997.

[37] Z. -B. Tang, K. R. Pattipati, and D. L. Kleinman, "Optimization of detection networks. II. Tree structures," *IEEE Trans. SMC*, vol. 23, no. 1, pp. 211-221, Jan./Feb. 1993.

[38] W. P. Tay, J. N. Tsitsiklis and M. Z. Win, "Data fusion trees for detection: does architecture matter?" *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4155-4168, Sep. 2008.

[39] O. P. Kreidl and A. S. Willsky, "An efficient message-passing algorithm for optimizing decentralized detection networks," *IEEE Trans. Autom. Control*, vol. 55, no. 3, pp. 563-578, Mar. 2010.

[40] A. Pete, K. R. Pattipati, and D. L. Kleinman, "Optimization of decision networks in structured task environments," *IEEE Trans. SMC*, vol. 26, no. 6, pp. 739-748, Nov. 1996.

[41] J. N. Tsitsiklis, "Decentralized detection by a large number of sensors," *Mathemat. Contr., Signals Syst.*, vol. 1, pp. 167-182, 1988.

[42] S. Marano, V. Matta and F. Mazzarella, "Refining decisions after losing data: the unlucky broker problem," *IEEE Trans. Signal Processing*, vol. 58, no. 4, pp. 1980-1990, Apr. 2010.

[43] E. Akofor and B. Chen, "Interactive distributed detection: architecture and performance analysis," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6456-6473, Oct. 2014.

[44] J. N. Tsitsiklis, "Extremal properties of likelihood-ratio quantizers," *IEEE Trans. Commun.*, vol. 41, no. 4, pp. 550-558, Apr. 1993.

[45] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Trans. Information Theory*, 45(4): 1315-1321, 5 1999.

[46] Y. K. Lee and B. U. Park, "Estimation of Kullback-Leibler divergence by local likelihood," *Annals of the Institute of Statistical Mathematics*, 58(2):327-340, Jun, 2006.

[47] N. N. Leonenko, L. Pronzato and V. Savani, "A class of renyi information estimators for multidimensional densities," *Annals of Statistics*, vol. 36, No. 5, pp. 2153-2182, Oct, 2008.

[48] X. Nguyen, M. J. Wainwright and M. I. Jordan, "Nonparametric estimation of the likelihood ratio and divergence functionals," *IEEE Trans. Inf. Theor.*, vol. 56, no. 11, pp. 5847-5861, Nov. 2010.

[49] Q. Wang S. Kulkarni and S. Verd$\acute{u}$, "Divergence estimation of continuous distributions based on data-dependent partitions," *Proc. IEEE Int. Symp. Information Theory*, vol. 51, no. 9, pp. 3064-3074, Sep. 2005.

VITA

NAME OF AUTHOR: Pengfei Yang

MAJOR: Electrical and Computer Engineering

EDUCATION

Ph.D. Dec.2016 Syracuse University, NY, USA (expected)

B.S. Jun.2010 University of Science and Technology of China, China

PUBLICATIONS

JOURNAL:

P. Yang and B. Chen, "Robust Kullback-Leibler divergence and robust hypothesis testing", *in preparation*.

P. Yang and B. Chen, "To listen or not: distributed detection with asynchronous transmissions", *Signal Processing Letters, IEEE*, vol.22, no.5, pp.628-632, Oct. 2014.

CONFERENCE:

P. Yang and B. Chen, "Deviation detection of continuous distribution", *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Orlando, FL, 2015, pp. 537-541.

P. Yang and B. Chen, "Wyner's common information in Gaussian channels", *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pp. 3112-3116, Honolulu, Hawaii, Jun. 2014.

P. Yang, B. Chen, H. Chen and P. K. Varshney, "Tandem distributed detection with conditionally dependent observations", *2012 15th International Conference on Information Fusion (FUSION)*, pp.1808-1813, Singapore, July. 2012.