Robust Kullback-Leibler Divergence and Universal Hypothesis Testing for Continuous Distributions

Pengfei Yang and Biao Chen

Abstract

Universal hypothesis testing refers to the problem of deciding whether samples come from a nominal distribution or an unknown distribution that is different from the nominal distribution. Hoeffding’s test, whose test statistic is equivalent to the empirical Kullback-Leibler divergence (KLD), is known to be asymptotically optimal for distributions defined on finite alphabets. With continuous observations, however, the discontinuity of the KLD in the distribution functions results in significant complications for universal hypothesis testing. This paper introduces a robust version of the classical KLD, defined as the KLD from a distribution to the Lévy ball of a known distribution. This robust KLD is shown to be continuous in the underlying distribution function with respect to the weak convergence. The continuity property enables the development of a universal hypothesis test for continuous observations that is shown to be asymptotically optimal for continuous distributions in the same sense as that of the Hoeffding’s test for discrete distributions.

Index Terms

Kullback-Leibler divergence, universal hypothesis testing, Lévy metric.

I. INTRODUCTION

The Kullback-Leibler divergence (KLD), also known as the relative entropy, is one of the most fundamental metrics in information theory and statistics [1], [2]. The KLD has a number of operational meanings and finds applications in a diverse range of disciplines. For example, the mutual information, which is a special case of the KLD, is a fundamental quantity in both channel coding and data compression [2]. In hypothesis testing, the KLD is known to be the decay rates of error probabilities (e.g., see Stein’s lemma [2] and Sanov’s theorem [3]).

An important application of the KLD is in the so-called universal hypothesis testing: given a nominal distribution \( P_0 \), the objective is to decide, upon observing a sample sequence, whether the underlying distribution that generates the sequence is \( P_0 \) or a distribution different from \( P_0 \). This problem was first formulated by Hoeffding [4]; with finite alphabet, Hoeffding developed a detector that is shown to be optimal according to the generalized Neyman-Pearson (NP) criterion, i.e., it achieves optimal type II error exponent subject to a constraint on the type-I error exponent [4]. The test statistic of Hoeffding’s detector is equivalent to the KLD between the empirical distribution and \( P_0 \).

Hoeffding’s result, however, does not generalize to the universal hypothesis testing with continuous alphabet. Clearly, computing empirical KLD for continuous distributions is meaningless as the empirical distribution, which is discrete, and the nominal distribution \( P_0 \), which is continuous, have different support sets. Additionally, the asymptotic optimality of Hoeffding’s test was established using a combinatorial argument [4] and thus is inapplicable to the continuous case. Attempts to reconstruct a similar decision

This work was supported in part by the Air Force Office of Scientific Research under Grant FA9550-16-1-0077.

P. Yang was with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, 13244 USA. Email: ypengf@gmail.com. He is now with Point72 Asset Management, LP., New York, NY. B. Chen is with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, 13244 USA. Email: bichen@syr.edu.
rule for continuous observations have been largely fruitless with the only exception of the work by Zeitouni and Gutman [5] where large deviation bounds were used in lieu of combinatorial bounds. The results in [5], however, are obtained at the cost of a weaker optimality with a rather complicated detector.

The difficulty in dealing with continuous observations for universal hypothesis testing stems from the subtle but important distinction on the continuity property of the KLD with respect to the underlying distributions. With finite alphabet distributions, the KLD defined between two distributions is known to be continuous in the distribution functions. This is not the case for the KLD defined between two distributions on the real line, i.e., those with continuous observations [6]. Specifically, weak convergence (i.e., convergence of distribution functions) does not imply convergence of the KLD. As such, even when two distributions who are arbitrarily close in terms of distribution functions, the KLD between them can be arbitrarily large.

This paper defines a robust version of the classical KLD that utilizes the Lévy metric which, unlike the KLD, is a true distance metric for distributions. The robust KLD, defined as the KLD from a distribution to a Lévy ball of another distribution, is shown to be continuous in the first distribution. This continuity property enables the development of a test for the universal hypothesis testing that is similar in its form to Hoeffding’s detector and attains the desired asymptotic NP optimality. Not only is the optimality in a stronger sense than that of [5], but the test statistic is also much more intuitive and easier to compute.

The rest of the paper is organized as follows. Section II defines the KLD between two (sets of) distributions; introduces the concepts of weak convergence and the Lévy metric along with their connections; and review the universal hypothesis testing for the finite alphabet case. Section III defines the robust KLD and establishes the continuity property with respect to weak convergence. In Section IV, the large deviation approach by Zeitouni and Gutman [5] to the universal hypothesis testing for continuous distributions is first reviewed; a robust version of the universal hypothesis testing problem is then introduced and the asymptotically NP optimal test using the robust KLD is derived. Section V concludes this paper.

II. PRELIMINARIES

A. The KL divergence

The KLD was first introduced in [7]. It is often used as a metric to quantify the distance between two probability distributions. For finite alphabets, the KLD between a probability distribution \( \mu = (\mu_1, \mu_2, \cdots, \mu_n) \) and another distribution \( P = (p_1, p_2, \cdots, p_n) \) is

\[
D(\mu \| P) = \sum_{i=1}^{n} \mu_i \log \frac{\mu_i}{p_i}. 
\]  

(1)

For distributions defined on the real line \( \mathcal{R} \), the KL divergence between \( \mu \) and \( P \) is defined as

\[
D(\mu \| P) = \int_{\mathcal{R}} d\mu \log \frac{d\mu}{dP}. 
\]  

(2)

The KLD \( D(\mu \| P) \) is jointly convex for both discrete and continuous distributions. For two sets of probability distributions, say \( \Gamma_1 \) and \( \Gamma_2 \), defined on the same probability space, the KLD between the two sets is defined to be the infimum of the KLD of all possible pairs of distributions, i.e.,

\[
D(\Gamma_1 \| \Gamma_2) := \inf_{\gamma_1 \in \Gamma_1, \gamma_2 \in \Gamma_2} D(\gamma_1 \| \gamma_2). 
\]  

(3)
B. Weak convergence and the Lévy metric

Denote the space of probability distributions on $(\mathbb{R}, \mathcal{F})$ as $\mathcal{P}$, where $\mathbb{R}$ is the real line and $\mathcal{F}$ is the sigma-algebra that contains all the Borel sets of $\mathbb{R}$. 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 [8].

Weak convergence is defined to be the convergence of the distribution functions as given below.

Definition 1. (Weak convergence [8], [9]) For $P_n, P \in \mathcal{P}$, we say $P_n$ weakly converges to $P$ and write $P_n \xrightarrow{w} P$, if $P_n(x) \rightarrow P(x)$ for all $x$ such that $P$ is continuous at $x$.

The Lévy metric $d_L$ between distributions $F \in \mathcal{P}$ and $G \in \mathcal{P}$ is defined as

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

The Lévy metric makes $(\mathcal{P}, d_L)$ a metric space [3], i.e., we have, for $\mu, P, Q \in \mathcal{P}$,

$$d_L(\mu, P) = 0 \iff \mu = P, \quad d_L(\mu, P) = d_L(P, \mu), \quad d_L(\mu, P) \leq d_L(\mu, Q) + d_L(Q, P).$$

The Lévy ball centered at $P_0 \in \mathcal{P}$ with radius $\delta$ is denoted as

$$B_L(P_0, \delta) = \{P \in \mathcal{P} : d_L(P, P_0) \leq \delta\}. \quad (4)$$

Fig. 1 plots the CDF of the standard normal distribution and its Lévy 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.

The Lévy metric is strongly related to the concept of the weak convergence of probability measures.

Lemma II.1. [8], [9] 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 $d_L(P_n, P) \rightarrow 0$.

C. Universal hypothesis testing

Let a sequence of independent and identically distributed (i.i.d.) observations $(x_0, \cdots, x_{n-1}) = x^n$ be the output of a source $P$. Consider the following hypothesis test

$$\mathcal{H}_0 : P = P_0, \quad \mathcal{H}_1 : P = P', \quad (5)$$
where $P_0$ is a known distribution while $P' \neq P_0$ is defined on the same probability space as $P_0$ but is otherwise unknown. The fact that $P'$ can be an arbitrary distribution gives rise to the name universal hypothesis testing. Clearly, while any decision rule will be independent of $P'$, the performance of the decision rule depends on $P'$.

The universal hypothesis testing was first studied by Hoeffding [4] who considered distributions with finite alphabet. Hoeffding’s detector is equivalent to the following threshold test of the empirical KLD:

$$D(\mu_n||P_0) \geq \eta. \quad (6)$$

Resorting to combinatorial bounds, Hoeffding successfully established the asymptotical Neyman-Pearson (NP) optimality of the above test. Specifically, let $\phi$ be the sequence of detectors $\{\phi^n(x_0, \ldots, x_{n-1}), n \geq 1\}$. Define the error exponents for the two types of error probabilities respectively as follows,

$$I^P(\phi) := \liminf_{n \to \infty} -\frac{1}{n} \log P^n(\phi^n(x^n) = 0),$$

$$J^P(\phi) := \liminf_{n \to \infty} -\frac{1}{n} \log P^n_0(\phi^n(x^n) = 1).$$

Zeitouni and Gutman [5] have shown that to achieve the best trade-off between $I^P(\phi)$ and $J^P(\phi)$, the test can depend on $x^n$ only through the empirical measure $\hat{\mu}_n$, defined to be

$$\hat{\mu}_n(t) = \frac{\sum I(x_i \leq t)}{n}. \quad (7)$$

The sequence of detectors can be equivalently expressed using $\Omega$ which is a sequence of partitions $(\Omega_0(n), \Omega_1(n)) (n = 1, 2, \ldots)$ of which $\Omega_0(n) \cap \Omega_1(n) = \emptyset$ and $\mathcal{P} = \Omega_0(n) \cup \Omega_1(n)$. The decision rule is made in favor of $H_i$ if $\hat{\mu}_n \in \Omega_i(n)$, $i = 0, 1$. Therefore $I^P(\phi)$ and $J^P(\phi)$ can be written as

$$I^P(\Omega) = \liminf_{n \to \infty} -\frac{1}{n} \log P^n(\hat{\mu}_n \in \Omega_0(n)),$$

$$J^P(\Omega) = \liminf_{n \to \infty} -\frac{1}{n} \log P^n_0(\hat{\mu}_n \in \Omega_1(n)).$$

The generalized NP criterion maximizes the exponent of type II probability of error under a constraint on the minimal rate of decrease in type I probability of error:

$$\max_{\Omega} I^P(\Omega) \quad \text{s.t.} \quad J^P(\Omega) \geq \eta. \quad (8)$$

Thus, among all sequences of detectors that satisfy the constraint on the type I error exponent in (8), Hoeffding’s test in (6) maximizes the type II error exponent.

### III. Robust KLD

#### A. Continuity property of the KLD

Let the nominal distribution be $P_0$. Let $\mu$ and $\mu_k$, $k = 1, 2, \ldots$, be distributions with the same sample space as $P_0$. Suppose the sequence of distributions $\mu_k$ converge weakly to $\mu$. It is of interest to study whether the corresponding KLD between $\mu_k$ and $P_0$ also converge to the KLD between $\mu$ and $P_0$. That is, does $\mu_k \xrightarrow{w} \mu$ imply $D(\mu_k||P_0) \rightarrow D(\mu||P_0)$?

The statement is true if the distributions involved are defined on a finite alphabet. With finite elements in the sample space of $P_0$, $D(\mu||P_0)$ is continuous in $\mu$ that has the same sample support as $P_0$. Convergence in distribution implies convergence in the corresponding KLD.
This, however, is not the case for $P_0 \in \mathcal{P}$. Indeed, it was established in [6] that the KLD is only lower semicontinuous with respect to the weak convergence for the continuous case, i.e.,

$$D(\mu||P_0) \leq \liminf_{k \to \infty} D(\mu_k||P_0).$$

However, the KLD is not upper semicontinuous, i.e., the following result is not necessarily true:

$$D(\mu||P_0) \geq \limsup_{k \to \infty} D(\mu_k||P_0),$$

thus the KLD is not continuous in $\mu$ for continuous observations. To see this, let $\mu = P_0$ thus $D(\mu||P_0) = 0$. Choose a distribution, say $P$, that is not absolutely continuous with respect to $P_0$. Let

$$P_\epsilon = (1 - \epsilon)P_0 + \epsilon P,$$

then $P_\epsilon$ is also not absolutely continuous with respect to $P_0$, thus $D(P_\epsilon||P_0)$ is unbounded for any given $\epsilon > 0$. However, $P_\epsilon$ weakly converges to $P_0$ as $\epsilon \to 0$. While this construction takes advantage of the distributions that are not absolutely continuous with respect to $P_0$, the same is true even if one is constrained to a sequence of distributions that is absolutely continuous with respect to $P_0$.

This lack of continuity for the KLD for the continuous case is the primary reason for the difficulty in generalizing Hoeffding’s result to the universal hypothesis testing with continuous observations. A direct consequence of the lack of continuity is that the superlevel set defined by the KLD is not closed [5]. The superlevel set is given by

$$\{\mu \in \mathcal{P} : D(\mu||P_0) \geq \eta_1\}. \quad (9)$$

The fact that the KLD is not continuous in $\mu$ leads to the unexpected property that the closure of the above set encompasses the entire probability space, i.e., any $P$ that does not belong to the above superlevel set has a sequence of distributions in the set that weakly converge to $P$. As such, a test in a similar form as (6) can not be used for continuous distributions.

**B. Robust KLD and its continuity property**

Given a pair of distributions $(\mu, P_0)$, the robust KLD is defined to be the KLD between $\mu$ and the Lévy ball centered at $P_0$. Using the definition (3), the robust KLD is the KLD defined between the two sets $\{\mu\}$ and $B_L(P_0, \delta_0)$ where $\delta_0 > 0$ is the radius of the Lévy ball. We denote it simply as $D(\mu||B_L(P_0, \delta_0))$.

The following theorem establishes its continuity property in $\mu$ under some mild assumptions.

**Theorem III.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 $\mu$ with respect to the weak convergence.

The non-trivial part of the proof is to show that $D(\mu||B_L(P_0, \delta_0))$ is upper semicontinuous in $\mu$ (Lemma A.6). Lemma A.7 proves $D(\mu||B_L(P_0, \delta_0))$ is lower semicontinuous in $\mu$. Therefore, $D(\mu||B_L(P_0, \delta_0))$ is continuous in $\mu$. The complete proof is lengthy and is included in Appendix A. Important intermediate steps are summarized below.

1) We first partition (quantize) the real line into a set of finite intervals. The robust KLD corresponding to the quantized distributions converge to the true robust KLD as the quantization becomes finer. The proof is in essence proving that a max-min inequality is in fact an equality (Lemma A.1).
2) The robust KLD is defined as the infimum over a Lévy ball and it is established that there exists a distribution inside or on the surface of the Lévy ball that achieves the infimum (see proof of Lemma A.1).
3) The robust KLD is continuous in the radius of the Lévy ball (Lemma A.2).

4) The robust KLD and the quantized robust KLD are convex functions of the respective distributions (Lemma A.3).

5) The supremum of the robust KLD over a Lévy ball centered at the first distribution is achieved by a distribution whose distribution function consists of two parts with a single transition point: the first part (i.e., prior to the transition point) corresponds to the lower bound of the Lévy ball and the second part (i.e., after the transition point) corresponds to the upper bound of the Lévy ball. Thus the class of distributions so defined is determined by the transition point given the Lévy ball. As such, the problem of finding an optimal distribution is reduced to finding an optimal transition point (Lemma A.4).

6) The robust KLD is bounded with (Lemma A.5)
\[ \sup_{\mu, P_0 \in \mathcal{P}} D(\mu || B_L(P_0, \delta_0)) = \log \frac{1}{\delta_0}. \]

7) The supremum of the robust KLD over a Lévy ball converges to the robust KLD as the Lévy ball diminishes, i.e., as its radius goes to 0. Therefore, the robust KLD is upper semicontinuous (Lemma A.6).

8) The robust KLD is lower semicontinuous (Lemma A.7).

The intuition of the continuity property of the robust KLD is the following. The classical KLD is a function of two distributions, and its value may vary arbitrarily large with small perturbation in one of the distributions with respect to the Lévy metric. The reason is because the Lévy metric is strictly weaker than the KLD, i.e., convergence in the KLD necessarily implies convergence in the Lévy metric but not the other way around. For the robust KLD where the KLD is defined between the first distribution and a Lévy ball centered around the second distribution, small perturbations in the first distribution can now be tolerated by the Lévy ball around the second distribution, thanks again to the fact that the Lévy metric is strictly weaker than the KLD.

C. Discussions

The continuity property in Theorem III.1 does not hold if the distribution ball is constructed using some other measures, including the total variation and the KLD.

**Definition 2.** [10] The total variation between \( P \in \mathcal{P} \) and \( Q \in \mathcal{P} \) is
\[ d_{TV}(P, Q) := \sup_{S \in \mathcal{F}} |P(S) - Q(S)|. \]

Let \( P_0(t) \) be continuous in \( t \) and denote by \( B_{TV}(P_0, \delta_0) \) and \( B_{KL}(P_0, \delta_0) \) distribution balls defined using the total variation and the KLD, respectively, in a manner similar to that of the Lévy ball in [4]. We show that there exist a sequence \( P_n \) that weakly converge to \( P_0 \), yet neither \( D(P_n || B_{TV}(P_0, \delta_0)) \) nor \( D(P_n || B_{KL}(P_0, \delta_0)) \) converges to 0. For any \( n > 0 \), choose a \( P_n \in B_L(P_0, 1/n) \) such that \( P_n(t) \) is a step function, that is, the distribution function is a staircase function throughout the entire real line. 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, we have
\[
D(P_n || B_{TV}(P_0, \delta_0)) = \inf_{\{P \in B_{TV}(P_0, \delta_0)\}} D(P_n || P)
\]
\[
\geq \inf_{\{P \in B_{TV}(P_0, \delta_0)\}} P_n(S_n) \log \frac{P_n(S_n)}{P(S_n)} + P_n(S_n^c) \log \frac{P_n(S_n^c)}{P(S_n^c)}
\]
\[
= \inf_{\{P \in B_{TV}(P_0, \delta_0)\}} 1 \log \frac{1}{P(S_n)} + 0 \log \frac{0}{P(S_n^c)}
\]
\[
\geq 1 \log \frac{1}{P_0(S_n) + \delta_0}
\]
\[
= \log \frac{1}{\delta_0},
\]
where the first inequality is due to the data processing inequality of the KLD and the second inequality comes from the definition of $B_{TV}(P_0, \delta_0)$. Therefore,
\[
\lim_{n \to \infty} \inf D(P_n || B_{TV}(P_0, \delta_0)) \geq \log \frac{1}{\delta_0}
\]
\[
> D(P_0 || B_{TV}(P_0, \delta_0))
\]
\[
= 0.
\]

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

As for the KLD case, for any $P$ such that $D(P || P_0) \leq \delta_0$, $D(P_n || P) = \infty$ where $P_n$ is constructed in the same manner as above. Therefore, $D(P_n || B_{KL}(P_0, \delta_0)) \to D(P_0 || B_{KL}(P_0, \delta_0)) = 0$.

The assumption that $P_0(t)$ is continuous in $t$ is also necessary for the continuous property of the robust KLD to 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 $\mu_i$ be the distribution such that $\mu_i(t) = 0$ for $t < 0.5 + \frac{1}{i}$ and $\mu_i(t) = 1$ for $t \geq 0.5 + \frac{1}{i}$. Thus $\mu_i \xrightarrow{w} \mu$ as $i \to \infty$, where $\mu(t) = 0$ for $t < 0.5$ and $\mu(t) = 1$ for $t \geq 0.5$. We can see that $D(\mu || B_L(P_0, 0.5)) = 0$ since $\mu \in B_L(P_0, 0.5)$. As $\mu_i \xrightarrow{w} \mu$,
\[
\lim_{i \to \infty} D(\mu_i || B_L(P_0, 0.5)) = \lim_{i \to \infty} \log \frac{1}{0.5}
\]
\[
> D(\mu || B_L(P_0, 0.5)),
\]
and the distribution in $B_L(P_0, 0.5)$ achieving the KLD value of $\log 2$ is a degenerate one: it takes values of the two points 0.5 and 0.5 + $1/i$ with equal probability.

The proof of Theorem III.1 sheds some light on the dynamics of the KLD of continuous distributions. Furthermore, the established continuity property of the robust KLD provides is key to solving the robust version of the universal hypothesis testing problem for the continuous case. This will be elaborated in the following section.

IV. ROBUST UNIVERSAL HYPOTHESIS TESTING

A. Review of the large deviation approach

While Hoeffding’s test do not apply to distributions with continuous observations, Zeitouni and Gutman [5] developed a universal hypothesis test for distributions defined on the real line under a strictly weaker notion of optimality. Their approach relies on the large deviation theory, specifically, the general Sanov’s theorem. For a given set $\Gamma \subset \mathcal{P}$, denote the closure and interior sets of $\Gamma$ as $cl\Gamma$ and $int\Gamma$. 
Theorem IV.1 (General Sanov’s Theorem). [3] Given a probability set $\Gamma \subseteq \mathcal{P}$, for a probability measure $Q \notin \Gamma$,

$$\inf_{P \in \text{cl} \Gamma} D(P||Q) \leq \lim_{n \to \infty} \inf \frac{1}{n} \log Q(\{x^n : \hat{\mu}_n \in \Gamma\}) \leq \lim_{n \to \infty} \sup \frac{1}{n} \log Q(\{x^n : \hat{\mu}_n \in \Gamma\}) \leq \inf_{P \in \text{int} \Gamma} D(P||Q),$$

where $\hat{\mu}_n$ is the empirical distribution defined in [7].

The general Sanov’s Theorem illustrates the large deviation principle for the empirical measures and is used extensively in the proof of Theorems IV.2 and IV.3. For any set $\Gamma \subseteq \mathcal{P}$, define its $\delta$–smooth set to be

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

The major contribution in [5] is summarized in the Theorem below.

Theorem IV.2. [5] Define $\Lambda$ as,

$$\Lambda_1(n) = \Lambda_1 := \{\mu : D(B_L(\mu, 2\delta)||P_0) \geq \eta\}^\delta, \quad \Lambda_0 := \mathcal{P} \setminus \Lambda_1.$$  

$\Lambda$ is $\delta$–optimal, i.e.,

1) $J^{P_0}(\Lambda) \geq \eta$.

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

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

Theorem IV.2 applies to both discrete and $\mathcal{R}$-valued random variables. However, for the finite alphabet case, the corresponding detector as in (10) yields weaker results than Hoeffding’s detector [4], a price paid for its generality.

With continuous alphabet, one has to be content with “$\delta$–optimal” rather than “optimal” for universal hypothesis testing if there is no restriction on the detector $\Omega$ while using the general Sanov’s Theorem. For a test $\Omega$ defined by the partition of the probability space $(\Omega_1, \Omega_2)$, either $\Omega_1$ or $\Omega_2$, may consist of only empirical distributions - it was established that the optimal test can depend on the observations only through the empirical distributions [5]. Suppose $\Omega_1$ consists of only empirical distributions. Then $\text{int} \Omega_1$ is empty and $\text{cl} \Omega_2$ equals to $\mathcal{P}$. It is also possible that the interior set and closure are too abstract or complicated to describe. In these cases, one can not take advantage of the general Sanov’s Theorem to analyze the error exponents. That is why in Theorem IV.2 for an arbitrary test $\Omega$, we need to first perform $\delta$–smooth operation on it before comparing its error exponents to those of the test $\Lambda$. As such, without any restriction on the detector $\Omega$, one has to settle with “$\delta$–optimality” instead of “optimality” for the continuous case.

Detector (10) has a complicated form. Given the empirical distribution $\hat{\mu}_n$, it is difficult to determine whether $\hat{\mu}_n \in \Lambda_1$ or $\Lambda_2$ due to the following two reasons.

- Computing $D(B_L(\hat{\mu}_n, 2\delta)||P_0)$ is an infinite dimension optimization problem. This is illustrated in Fig. 2 where one needs to find a continuous $\mu^*$ inside the shaded region such that

$$D(\mu^*||P_0) = \inf_{\mu \in B_L(\hat{\mu}_n, 2\delta)} D(\mu||P_0).$$

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

Fig. 2: The shaded region is $B_L(\hat{\mu}_n, 2\delta)$ and the solid line is $P_0$.

One of the difficulties to directly generalize the discrete case to the continuous case, as mentioned in [5], is that the superlevel set $\{\mu \in \mathcal{P} : D(\mu||P_0) \geq \eta\}$ is not closed in $\mathcal{P}$, which we have discussed in detail in Section III-A. In the next section, rather than “$\delta$–smoothing” the detector as one does in (10), we generalize the hypothesis $\mathcal{H}_0$ from $P_0$ to $B_L(P_0, \epsilon_0)$. Then, we show that under the minimax criterion, the empirical likelihood ratio test is optimal.

B. Robust Universal Hypothesis Testing

Let $\mathcal{P}_0 := B_L(P_0, \epsilon_0)$. The robust version of the universal hypothesis testing amounts to the testing of the following two hypotheses.

$$\mathcal{H}_0 : P \in \mathcal{P}_0, \quad \mathcal{H}_1 : P = P'. \quad (12)$$

Here $P_0$ is assumed to be a known continuous distribution and $\epsilon_0 > 0$. $P' \notin \mathcal{P}_0$ and is unknown.

Compared to the universal hypothesis test (5), the robust version (12) replaces a single $P_0$ with $B_L(P_0, \epsilon_0)$ for the null hypothesis. With this robust setting, the asymptotic NP criterion (8) is replaced by the following minimax asymptotic NP criterion:

$$\max_{\Omega} I^{P'}(\Omega) \quad \text{s.t.} \quad J^{P_0}(\Omega) \geq \eta, \quad (13)$$

where

$$J^{P_0}(\Omega) := \inf_{P \in \mathcal{P}_0} J^P(\Omega). \quad (14)$$

Thus type II error exponent is maximized subject to a constraint on the worst type I error exponent. The reason that the Lévy metric is used to define $\mathcal{P}_0$ is that the Lévy metric is the weakest hence the most general one [10]. 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. An additional advantage is that the resulting optimal detector is rather intuitive and straightforward to implement. Theorem IV.3 below describes the optimal solution to the robust universal hypothesis testing problem.

**Theorem IV.3.** For the robust universal hypothesis testing problem, the detector $\Lambda = \{\Lambda_0, \Lambda_1\}$ defined by, for some $\eta > 0$,

$$\Lambda_1(n) = \Lambda_1 := \{\mu : D(\mu||B_L(P_0, \epsilon_0)) > \eta\}, \quad \Lambda_0(n) = \mathcal{P} \setminus \Lambda_1,$$

satisfies the following properties:
1) \( J^{P_0}(\Lambda) = \eta \).
2) \( I^{P'}(\Lambda) = D(\Lambda_0||P') \).
3) For any detector \( \Omega \) with \( \Omega_1(n) = \Omega_1 \) with \( \Omega_1 \) open, if
   \[
   J^{P_0}(\Omega) > \eta,
   \]
   then for any \( P' \notin B_L(P_0, \varepsilon_0) \),
   \[
   I^{P'}(\Omega) \leq I^{P'}(\Lambda).
   \]

Theorem IV.3 states that the detector
   \[
   D(\hat{\mu}_n||P_0) \overset{H_1}{\geq} \eta,
   \]
   is optimal in type II error decay rate among all detectors \( \Omega = \{ \Omega_0(n), \Omega_1(n) \} \) that have the same worst case type I error decay rate as \( \Lambda \). In particular, the optimal type II error decay rate is precisely \( D(\Lambda_0||P') \) when \( P' \) is the true distribution under \( H_1 \).

Proof. The three parts of the Theorem are proved below.

1) From the general Sanov’s theorem, we have
   \[
   \inf_{P \in \mathcal{P}_0} J^P(\Lambda) = \inf_{P \in \mathcal{P}_0} \lim_{n \to \infty} \inf \frac{1}{n} \log P^n(\{x^n : \hat{\mu}_n \in \Lambda_1\})
   \geq \inf_{\mu \in c\Lambda_1} D(\mu||P)
   = \inf_{\mu \in c\Lambda_1} D(\mu||\mathcal{P}_0)
   = \eta,
   \]
   the last equality holds since \( D(\mu||\mathcal{P}_0) \) is continuous in \( \mu \) thus \( c\Lambda_1 \subseteq \{ \mu : D(\mu||\mathcal{P}_0) \geq \eta \} \). On the other hand,
   \[
   \inf_{P \in \mathcal{P}_0} J^P(\Lambda) \leq \inf_{P \in \mathcal{P}_0} \lim_{n \to \infty} \sup \frac{1}{n} \log P^n(\{x^n : \hat{\mu}_n \in \Lambda_1\})
   \leq \inf_{\mu \in c\Lambda_1} D(\mu||P)
   = \eta.
   \]
   The last equality holds since \( int\Lambda_1 = \Lambda_1 \).

2) Again from the general Sanov’s theorem, we have
   \[
   I^{P'}(\Lambda) = \lim_{n \to \infty} \inf \frac{1}{n} \log P^n(\{x^n : \hat{\mu}_n \in \Lambda_0\})
   \geq \inf_{\mu \in c\Lambda_0} D(\mu||P')
   = D(\Lambda_0||P').
   \]
   The last equality holds since \( c\Lambda_0 = \Lambda_0 \). On the other hand, \( \{ \mu : D(\mu||\mathcal{P}_0) < \eta \} \subseteq int\Lambda_0 \), thus,
   \[
   I^{P'}(\Lambda) \leq \lim_{n \to \infty} \sup \frac{1}{n} \log P^n(\{x^n : \hat{\mu}_n \in \Lambda_0\})
   \leq \inf_{\mu \in int\Lambda_0} D(\mu||P')
   \leq \inf_{\mu : D(\mu||\mathcal{P}_1) < \eta} D(\mu||P')
   \leq D(\Lambda_0||P').
   \]
Inequality (19) holds because of the following. There exists a distribution \( P \in \mathcal{P}_0 \) such that \( D(P||P') < \infty \). For any \( P_c \in \Lambda_0 \) and \( 0 < \lambda < 1 \), we have \( (1 - \lambda)P_c + \lambda P \in \{ \mu : D(\mu||\mathcal{P}_0) < \eta \} \) since
\[
D((1 - \lambda)P_c + \lambda P||\mathcal{P}_0) \leq (1 - \lambda)D(P_c||\mathcal{P}_0) + \lambda D(P||\mathcal{P}_0)) < (1 - \lambda)\eta + 0 < \eta,
\]
where (20) comes from the fact that \( D(\mu||\mathcal{P}_0) \) is convex in \( \mu \) while inequality (21) is due to the fact \( P \in \mathcal{P}_0 \). Thus,
\[
\inf_{\mu \in \{ \mu : D(\mu||\mathcal{P}_1) < \eta \}} D(\mu||P') \leq \lim_{\lambda \to 0^+} D((1 - \lambda)P_c + \lambda P||P') \leq \lim_{\lambda \to 0^+} (1 - \lambda)D(P_c||P') + \lambda D(P||P') \leq D(P_c||P'),
\]
the last inequality holds since \( D(P||P') < \infty \). The above inequalities hold for any \( P_c \in \Lambda_0 \), thus we have
\[
\inf_{\mu \in \{ \mu : D(\mu||\mathcal{P}_0) < \eta \}} D(\mu||P') \leq D(\Lambda_1||P').
\]
3) We have
\[
\inf_{P \in \mathcal{P}_0} D(\Omega_1||P) = \inf_{P \in \mathcal{P}_0} D(int \Omega_1||P) \geq \inf_{P \in \mathcal{P}_0} \liminf_{n \to \infty} -\frac{1}{n} \log P(x^n : \hat{\mu}_n \in \Omega_1) > \eta.
\]
Therefore, \( \Omega_1 \subseteq \Lambda_1 \), or equivalently, \( \Lambda_0 \subseteq \Omega_0 \). Next,
\[
I^{\pi'}(\Omega) = \lim_{n \to \infty} -\frac{1}{n} \log P^n(x^n : \hat{\mu}_n \in \Omega_0) \leq \liminf_{n \to \infty} -\frac{1}{n} \log P^n(x^n : \hat{\mu}_n \in \Lambda_0) = I^{\pi'}(\Lambda).
\]

Compared to detector (10) in Theorem IV.2, detector (17) has three main differences.
- Computing \( D(\hat{\mu}_n||B_1(P_0, \delta)) \) is a finite dimension optimization problem, which is in essence finding a step function inside the shaded area that achieves the minimum KLD to \( \hat{\mu}_n \) (see Fig. 3). This can be shown to be a convex optimization problem with linear constraints thus can be readily solved via standard convex programs [11].
- From Theorem IV.2 the detector developed in [5] can not be compared directly to an arbitrary detector \( \Omega \); instead, \( \Omega^\delta \) is used in establishing the optimality of the proposed detector. This ensures that \( \Omega^\delta \) is open and \( \Omega^\delta \) is closed yet this leads to a weaker sense of optimality, i.e., \( \delta \)-optimality.
- In Theorem IV.3 by restricting \( \Omega_1 \) to be independent of \( n \) and assuming \( \Omega_1 \) is open, asymptotic NP optimality is established which is stronger than \( \delta \)-optimality.
- Theorem IV.2 only provides the lower bound for error exponents, while Theorem IV.3 characterizes the exact values of error exponents. Furthermore, while the error exponents \( I \) and \( J \) are defined using limit infimum, from the proof it can be seen that \( I \) and \( J \) remain unchanged if one uses limit to define error exponents. Therefore, Theorem IV.3 gives an exact characterization of error exponents.

Summarizing, by considering the universal hypothesis testing in the robust setting, the generalized empirical likelihood ratio test becomes optimal, and the construction of the detector and the proof of optimality are much simplified.
Fig. 3: The shaded region is a Lévy ball of the normal distribution and the step function is an example of $\hat{\mu}_n$.

V. CONCLUSION

The Kullback-Leibler divergence (KLD) between a pair of distributions is only lower semicontinuous in the distribution functions for continuous observations. This is in contrast to the case with finite alphabet in which KLD is known to be continuous. As such, while simple and optimal solution may exist for some hypothesis testing problems involving finite alphabet observations, these results often do not generalize to the continuous case as the continuity of KLD plays a crucial role in obtaining the optimal test.

The problem considered in the present paper is the universal hypothesis testing where the null hypothesis is specified by a nominal distribution whereas the alternative hypothesis is specified by a different but otherwise unknown distribution. With finite alphabet, Hoeffding’s test, which is in essence a threshold test of the empirical KLD, is known to be asymptotically Neyman-Pearson (NP) optimal for the finite-alphabet case. For continuous observations, however, existing results have to resort to a weaker notion of optimality with a much more complicated detector compared with Hoeffding’s detector.

This paper introduced the notion of the robust KLD, defined as the KLD between a distribution to the Lévy ball of another distribution. In contrast to the classical KLD, this robust KLD was shown to be continuous in the first distribution function. Subsequently, by formulating a robust version of the universal hypothesis testing where the null hypothesis is specified by a Lévy ball centered around the nominal distribution, it was established that the generalized empirical likelihood ratio test is optimal under the asymptotic minimax NP criterion whose error exponents were characterized precisely. Additionally, the test itself is also much more intuitive and easier to evaluate compared with existing approaches in the literature.

APPENDIX A

PROOF OF THEOREM [III.1]

The set of all partitions $\mathcal{A} = (A_1, \ldots, A_{|\mathcal{A}|})$ of $\mathcal{R}$ into a finite number of sets $A_i$ is denoted by $\Pi$. For a given partition $\mathcal{A}$, denote by $P^\mathcal{A}$ the quantized (discrete) probability over $\mathcal{A}$ of a probability distribution $P \in \mathcal{P}$. Thus $P^\mathcal{A}$ is a $|\mathcal{A}|$ dimensional vector $(P(A_1), P(A_2), \ldots, P(A_{|\mathcal{A}|})) \in \mathcal{R}^{|\mathcal{A}|}$. We introduce a new definition of KLD utilizing partitions, which is equivalent to the classical definition using the Radon-Nikodym derivative in (2).

Definition 3 (The KLD [12]). The KLD between $P \in \mathcal{P}$ and $Q \in \mathcal{P}$ is defined as,

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

(22)
where
\[
D(P^A||Q^A) = \sum_{i=1}^{|A|} P(A_i) \log \frac{P(A_i)}{Q(A_i)}.
\]

The following lemma generalizes [22] from the classical KLD to the robust KLD. For set $\Gamma \subseteq \mathcal{P}$, we define $\Gamma^A := \{P^A : P \in \Gamma\}$.

**Lemma A.1.** For $\mu, P_0 \in \mathcal{P}$ and $\delta_0 > 0$, $D(\mu||B_L(P_0, \delta_0)) = \sup_{\mathcal{A} \in \Pi} D(\mu^A||B_L^A(P_0, \delta_0))$.

**Proof.** The probability space $\mathcal{P}$ defined on $(\mathcal{R}, \mathcal{F})$, where $\mathcal{R}$ is the real line and $\mathcal{F}$ is the sigma-algebra that contains all the Borel sets of $\mathcal{R}$, is not compact with respect to the weak convergence [8]. 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. As a consequence, $\mathcal{P} \subseteq \mathcal{M}$ and $\mathcal{M}$ is compact with respect to the weak convergence [13]. 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 $P \in \mathcal{P}$ [8], 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 Lévy metric $d_L$ and the KLD extend unchanged to $F \in \mathcal{M}$ and $G \in \mathcal{M}$ [8, 13]. The following three steps constitute the proof of the lemma,

\[
D(\mu||B_L(P_0, \delta_0)) = D(\mu||\bar{B}_L(P_0, \delta_0)),
\]

\[
D(\mu||\bar{B}_L(P_0, \delta_0)) = \sup_{\mathcal{A} \in \Pi} D(\mu^A||\bar{B}_L^A(P_0, \delta_0)),
\]

\[
\sup_{\mathcal{A} \in \Pi} D(\mu^A||\bar{B}_L^A(P_0, \delta_0)) = \sup_{\mathcal{A} \in \Pi} D(\mu^A||\tilde{B}_L^A(P_0, \delta_0)),
\]

where $\bar{B}_L(P_0, \delta_0) := \{P \in \mathcal{M} : d_L(P, P_0) \leq \delta_0\}$. We now prove (23)-(25). We first prove (23). Note that $\bar{B}_L(P_0, \delta_0)$ is closed with respect to the weak convergence, thus is compact since $\mathcal{M}$ is compact. Let

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

the existence of $P_\mu$ 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 assume $P_\mu(\infty) = \delta$ and $P_\mu(\infty) = 1$; other cases can be proved in a similar manner. Let $s$ denote the minimum $t$ such that $P_0(t - \delta_0) \geq \delta_0$, we construct $P_\mu'$ 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 $t < s$,

\[
\frac{(P_\mu(t) - \delta)P_\mu(s)}{P_\mu(s) - \delta} \leq P_\mu(t) \iff P_\mu(t) \leq P_\mu(s),
\]
then \(d_L(P'_\mu, P_0) \leq \delta_0\) because

\[
P_0(t - \delta_0) - \delta_0 < 0 \leq P'_\mu(t) \leq P_\mu(t) \leq P_0(t + \delta_0) + \delta_0.
\]

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

\[
D(\mu||P'_\mu) = \int_{-\infty}^{s-} d\mu(t) \log \frac{d\mu(t)}{dP'_\mu(t)} + \int_{s}^{\infty} d\mu(t) \log \frac{d\mu(t)}{dP'_\mu(t)} + \int_{s}^{\infty} d\mu(t) \log \frac{d\mu(t)}{dP_\mu(t)}
\]

\[
= \mu(s-) \log \frac{P_\mu(s) - \delta}{P'_\mu(s)} + \int_{s}^{\infty} d\mu(t) \log \frac{d\mu(t)}{dP_\mu(t)}
\]

\[
\leq D(\mu||P_\mu).
\]

Therefore in either case, there exists \(P'_\mu \in B_L(P_0, \delta_0)\) such that

\[
P'_\mu = \arg \inf_{\{P \in B_L(P_0, \delta_0)\}} D(\mu||P).
\]

(26)

To prove (24), note that Lemma 2.4 in \[14\] shows that

\[
D(B_L(P_0, \delta_0)||\mu) = \sup_{A \in \Pi} D(B^A_L(P_0, \delta_0)||\mu^A).
\]

Using a parallel proof, one can establish that

\[
D(\mu||\bar{B}_L(P_0, \delta_0)) = \sup_{A \in \Pi} D(\mu^A||\bar{B}^A_L(P_0, \delta_0)),
\]

i.e., (24) holds.

Finally, (25) holds since for any \(A \in \Pi, \bar{B}^A_L(P_0, \delta_0) = B^A_L(P_0, \delta_0)\).

The robust KLD is continuous in the radius of the Lévy ball under a mild assumption, as stated in the lemma below.

Lemma A.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 \(\delta_0\).

Proof. Let \(\delta \in (0, \delta_0)\). \(D(\mu||B_L(P_0, \delta_0))\) is left continuity in \(\delta_0\) if \(D(\mu||\{P \in \mathcal{P} : d_L(P, P_0) < \delta_0\}) = D(\mu||B_L(P_0, \delta_0))\). It is easy to see that

\[
D(\mu||\{P \in \mathcal{P} : d_L(P, P_0) < \delta_0\}) \geq D(\mu||B_L(P_0, \delta_0)).
\]

Thus we only need to show the other direction. Denote

\[
P_\delta = \arg \inf_{\{P \in B_L(P_0, \delta)\}} D(\mu||P),
\]

\[
P_{\delta_0} = \arg \inf_{\{P \in B_L(P_0, \delta_o)\}} D(\mu||P).
\]

The existence of \(P_\delta\) and \(P_{\delta_0}\) is guaranteed by (26). 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,
\[
D(\mu||\{P \in \mathcal{P} : d_L(P, P_0) < \delta_0\}) \leq \lim_{\lambda \to 0^+} D(\mu||\lambda P_0 + (1 - \lambda)P_0) \\
\leq \lim_{\lambda \to 0^+} \lambda D(\mu||P_0) + (1 - \lambda)D(\mu||P_0) \\
= D(\mu||P_0) \\
= D(\mu||B_L(P_0, \delta_0)).
\]
Therefore $D(\mu||B_L(P_0, \delta_0))$ is left continuous in $\delta_0$.

The rest is to show $D(\mu||B_L(P_0, \delta_0))$ is right continuous in $\delta_0$. Since $D(\mu||B_L(P_0, \delta_0))$ is decreasing in $\delta_0$, we only need to show:
\[
\lim_{n \to \infty} D\left(\mu||B_L(P_0, \delta_0 + \frac{1}{n}\right) \geq D(\mu||B_L(P_0, \delta_0)).
\]
From (26), there exists $P_n \in B_L(P_0, \delta_0 + \frac{1}{n})$ such that $D(\mu||P_n) = D(\mu||B_L(P_0, \delta_0 + \frac{1}{n})$. $\mathcal{M}$ 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 $n$, $P^* \in B_L(P_0, \delta_0)$. We have,
\[
\lim_{n \to \infty} D\left(\mu||B_L(P_0, \delta_0 + \frac{1}{n}\right) = \lim_{n \to \infty} D(\mu||P_n) \\
\geq D(\mu||P^*) \\
\geq D(\mu||B_L(P_0, \delta_0)) \\
= D(\mu||B_L(P_0, \delta_0)),
\]
where (28) comes from the fact that the KLD is lower semicontinuous and the last equality was proved in (23). Therefore $D(\mu||B_L(P_0, \delta_0))$ is right continuous in $\delta_0$. 

**Lemma A.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 $\mu$. In addition, for any partition $A$ of the real line, $D(\mu^A||B^A_L(P_0, \delta_0))$ is convex in $\mu^A$.

**Proof.** Let $P_i = \arg \inf_{P \in B_L(P_0, \delta_0)} D(\mu_i||P)$, $i = 1, 2$. For any $0 < \lambda < 1$, $\lambda P_1 + (1 - \lambda)P_2 \in B_L(P_0, \delta_0)$, thus,
\[
D(\lambda P_1 + (1 - \lambda)P_2||B_L(P_0, \delta_0)) \leq D(\lambda P_1 + (1 - \lambda)P_2||\lambda P_1 + (1 - \lambda)P_2) \\
\leq \lambda D(\mu_1||P_1) + (1 - \lambda)D(\mu_2||P_2) \\
= \lambda D(\mu_1||B_L(P_0, \delta_0)) + (1 - \lambda)D(\mu_2||B_L(P_0, \delta_0)).
\]
Therefore, $D(\mu||B_L(P_0, \delta_0))$ is a convex function of $\mu$. That $D(\mu^A||B^A_L(P_0, \delta_0))$ is convex in $\mu^A$ follows a similar argument. 

**Lemma A.4.** Given $\mu_0, P_0 \in \mathcal{P}$ and $\delta, \delta_0 > 0$, we have
\[
\sup_{\mu \in B_L(\mu_0, \delta)} D(\mu||B_L(P_0, \delta_0)) = \sup_{x \in \mathbb{R}} D(\mu_x^\delta||B_L(P_0, \delta_0)),
\]
where
\[
\mu_x^\delta(t) = \begin{cases} 
\max(0, \mu_0(t - \delta) - \delta) & \text{if } t < x, \\
\min(1, \mu_0(t + \delta) + \delta) & \text{if } t \geq x.
\end{cases}
\]
Proof. We have

\[
\sup_{\mu \in B_L(\mu_0, \delta)} D(\mu || B_L(P_0, \delta_0)) = \sup_{\mu \in B_L(\mu_0, \delta)} \sup_{A \in \Pi} D(\mu^A || B_L^A(P_0, \delta_0))
\]

\[
= \sup_{A \in \Pi} \sup_{\mu \in B_L(\mu_0, \delta)} D(\mu^A || B_L^A(P_0, \delta_0))
\]

\[
= \sup_{A \in \Pi} \sup_{\mu^A \in B_L^A(\mu_0, \delta)} D(\mu^A || B_L^A(P_0, \delta_0)).
\]

Equality (28) comes from Lemma A.1.

Fix a finite partition \( \mathcal{A} \) of the real line. Without loss of generality we can assume \( |\mathcal{A}| = n \) and

\[ \mathcal{A} = \{(-\infty, a_1], (a_1, a_2], \ldots, (a_{n-2}, a_{n-1}], (a_{n-1}, \infty]\}. \]

The partition \( \mathcal{A} \) over the probability space \( \mathcal{P} \) can be represented as an \( n \)-dimensional polytope. Denote the \( n \)-dimensional point \( x = (x_1, x_2, \ldots, x_n) \),

\[ \mathcal{P}^A = \{ x \in R^n : \sum_i x_i = 1 \text{ and } \forall i, 0 \leq x_i \leq 1 \}. \]

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

\[ B_L^A(\mu_0, \delta) = \{ x \in \mathcal{P}^A : \forall 1 \leq j \leq n-1, L_j \leq \sum_{i=1}^j x_i \leq U_j \}, \]

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 point \( x \) is a vertex of \( B_L^A(\mu_0, \delta) \) if and only if \( \sum_{i=1}^n x_i \) 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 \( x \), 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 \( x^1, \ldots, x^n \) of \( B_L^A(\mu_0, \delta) \) 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 A.3, \( D(\cdot || B_L^A(P_0, \delta_0)) \) is a convex function, thus the supremum on the polytope \( B_L^A(\mu_0, \delta) \) is achieved at its vertices. Let

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

Then any \( x^k \) is a quantization of \( \mu_x^\delta \) over the partition \( \mathcal{A} \) for some \( x \).

\[
\sup_{\mu^A \in B_L^A(\mu_0, \delta)} D(\mu^A || B_L^A(P_0, \delta_0)) = \max_k D(x^k || B_L^A(P_0, \delta_0))
\]

\[
\leq \sup_x D(\mu_x^\delta || B_L^A(P_0, \delta_0))
\]

\[
\leq \sup_{A \in \Pi} \sup_{x \in \mathcal{R}} D((\mu_x^\delta)^A || B_L^A(P_0, \delta_0)),
\]

\[
= \sup_{x \in \mathcal{R}} D((\mu_x^\delta)^A || B_L^A(P_0, \delta_0)),
\]

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

(30)
the last equality comes from Lemma [A.1]. From (29) and (30), we have
\[ \sup_{\mu \in B_L(\mu_0, \delta)} D(\mu || B_L(P_0, \delta_0)) \leq \sup_{x \in \mathcal{R}} D(\mu^\delta_x || B_L(P_0, \delta_0)). \]

For the other direction, since \( \mu^\delta_x \in B_L(\mu_0, \delta) \),
\[ \sup_{\mu \in B_L(\mu_0, \delta)} D(\mu || B_L(P_0, \delta_0)) \geq \sup_{x \in \mathcal{R}} D(\mu^\delta_x || B_L(P_0, \delta_0)). \]

Therefore,
\[ \sup_{\mu \in B_L(\mu_0, \delta)} D(\mu || B_L(P_0, \delta_0)) = \sup_{x \in \mathcal{R}} D(\mu^\delta_x || B_L(P_0, \delta_0)). \]

A direct result of the above lemma is the boundedness of the robust KLD, which is stated below.

**Lemma A.5.** \( \sup_{\mu, P_0 \in \mathcal{P}} D(\mu || B_L(P_0, \delta_0)) = \log \frac{1}{\delta_0} \).

**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 [A.4],
\[ \sup_{\mu \in B_L(S_0, 1)} D(\mu || B_L(P_0, \delta_0)) = \sup_{x \in \mathcal{R}} D(\mu^\delta_x || B_L(P_0, \delta_0)), \]
where \( \mu^1_x(t) = 0 \) for \( t < x \), and \( \mu^1_x(t) = 1 \) for \( t \geq x \). Therefore,
\[ \sup_{x \in \mathcal{R}} D(\mu^\delta_x || B_L(P_0, \delta_0)) = \sup_{x \in \mathcal{R}} \log \frac{1}{\min(1, P_0(x + \delta_0) + \delta_0) - \max(0, P_0(x - \delta_0) - \delta_0)} \]
\[ = \log \frac{1}{\delta_0}, \]
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 that belongs to \( \mathcal{M} \setminus \mathcal{P} \) can always achieve the supremum for any \( P_0 \).

With all the previous lemmas, we now show that robust KLD is upper semicontinuous.

**Lemma A.6.** 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 \( \mu \) with respect to the weak convergence.

**Proof.** For any fixed \( \mu_0 \in \mathcal{P} \), the statement is equivalent to proving that when \( \delta \to 0 \),
\[ \lim_{\delta \to 0} \sup_{\mu \in B_L(\mu_0, \delta)} D(\mu || B_L(P_0, \delta_0)) \leq D(\mu_0 || B_L(P_0, \delta_0)). \] (31)

From Lemma [A.4] it is equivalent to proving
\[ \lim_{\delta \to 0} \sup_{x \in \mathcal{R}} D(\mu^\delta_x || B_L(P_0, \delta_0)) \leq D(\mu_0 || B_L(P_0, \delta_0)). \]
Denote $u_{-\delta}$ as the left boundary of support set of distribution $\mu(t + \delta)$ and $u_{-\delta}^\delta$ as the infimum $x$ such that $\mu(x + \delta) = 1 - \delta$. Similarly, denote $u_\delta$ as the left boundary of distribution $\mu(t - \delta)$ and $u_\delta^\delta$ as the infimum $x$ such that $\mu(x - \delta) = 1$. Fig. 4 illustrates these locations. Note that these boundary points may not be finite. We will first prove for any $\delta_1 \in (0, \delta_0)$,

$$
\limsup_{\delta \to 0} D(\mu_x^\delta || B_L(P_0, \delta_0)) \leq D(\mu_0 || B_L(P_0, \delta_0 - \delta_1)).
$$

Now fix $\delta_1$, we then establish that $D(\mu_x^\delta || B_L(P_0, \delta_0))$ can be uniformly bounded as $x$ varies. Denote

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

For fixed $\delta < \delta_1$, let

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

and

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

To get $P_{\delta_0 - \delta_1}^{\delta, u}(t)$, we first shift $P_{\delta_0 - \delta_1}(t)$ to the left by $\delta$, then scale it by $(1 - \delta_1)$ and shift it up by $\delta_1$; similarly to get $P_{\delta_0 - \delta_1}^{\delta, l}(t)$, we shift $P_{\delta_0 - \delta_1}(t)$ to the right by $\delta$, then scale it by $(1 - \delta_1)$. Clearly

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

For any $x$, construct $P_{\delta_0 - \delta_1}^x$ in a similar way as $\mu_x^\delta$,

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

$P_{\delta_0 - \delta_1}^x \in B_L(P_0, \delta_0)$ since

$$
d_L(P_{\delta_0 - \delta_1}^x, P_0) \leq d_L(P_{\delta_0 - \delta_1}^x, P_{\delta_0 - \delta_1}) + d_L(P_{\delta_0 - \delta_1}, P_0) \leq \delta_1 + (\delta_0 - \delta_1) = \delta_0,
$$

where the first inequality holds because $(P, d_L)$ is a metric space (i.e., $d_L$ satisfies the triangle inequality); the second inequality comes from (33) and the definition of $P_{\delta_0 - \delta_1}$. 

---

**Fig. 4:** Illustration of $u_{-\delta}, u_\delta, u_{-\delta}^\delta$ and $u_\delta^\delta$. The solid line represents $\mu_0$ and shaded region represents $B_L(\mu_0, \delta)$. 
From Lemma A.5, \( D(\mu_0||P_{\delta_0-\delta_1}) = D(\mu_0||B_L(P_0, \delta_0 - \delta_1)) < \infty \), therefore \( \mu_0 \) is absolutely continuous with respect to \( P_{\delta_0-\delta_1} \). From the construction of \( \mu^x_\delta \) and \( P^x_{\delta_0-\delta_1} \), we can see that \( \mu^x_\delta \) is absolutely continuous with respect to \( P^x_{\delta_0-\delta_1} \) as well. Therefore, we have

\[
\limsup_{\delta \to 0} D(\mu^x_\delta||B_L(P_0, \delta_0)) = \limsup_{\delta \to 0} \sup_{x \in \mathcal{R}} D(\mu^x_\delta||P_{\delta_0-\delta_1}) 
\leq \limsup_{\delta \to 0} D(\mu^x_\delta||P^x_{\delta_0-\delta_1}),
\]

establishing (32). We now prove \( D(\mu^x_\delta||P^x_{\delta_0-\delta_1}) \) can be uniformly bounded as \( x \) varies. Inequalities appear in cases 1)-3) are due to the log sum inequality unless otherwise stated.

1) For \( x < u_{-\delta} \),

\[
D(\mu^x_\delta||P^x_{\delta_0-\delta_1}) \leq \delta \log \frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u_{-\delta}} d(\mu_0(t + \delta)) \log \frac{d(\mu_0(t + \delta) + \delta)}{d((1 - \delta_1)P_{\delta_0-\delta_1}(t + \delta) + \delta_1)}
= \delta \log \frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u_{-\delta}} d(\mu_0(t + \delta)) \log \frac{d(\mu_0(t + \delta))}{(1 - \delta_1)d(P_{\delta_0-\delta_1}(t + \delta))}
= \delta \log \frac{\delta}{\delta_1} + (1 - \delta) \log \frac{1}{(1 - \delta_1)} + \int_{u_{-\delta}}^{u_{-\delta}+\delta} d(\mu_0(t)) \log \frac{d(\mu_0(t))}{d(P_{\delta_0-\delta_1}(t))},
\]

when \( \delta \to 0 \), the above converges to

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

2) For \( u_{-\delta} \leq x \leq u_\delta \),

\[
D(\mu^x_\delta||P^x_{\delta_0-\delta_1}) = (u_0(x + \delta) + \delta) \log \frac{(u_0(x + \delta) + \delta)}{(1 - \delta_1)P_{\delta_0-\delta_1}(x + \delta) + \delta_1}
+ \int_{x^+}^{u_{-\delta}} d(\mu_0(t + \delta)) \log \frac{d(\mu_0(t + \delta))}{d((1 - \delta_1)P_{\delta_0-\delta_1}(t + \delta) + \delta_1)}
\leq \delta \log \frac{\delta}{\delta_1} + (u_0(x + \delta)) \log \frac{d(\mu_0(t + \delta))}{(1 - \delta_1)d(P_{\delta_0-\delta_1}(x + \delta))}
+ \int_{x^+}^{u_{-\delta}} d(\mu_0(t + \delta)) \log \frac{d(\mu_0(t + \delta))}{(1 - \delta_1)d(P_{\delta_0-\delta_1}(t + \delta))}
= \delta \log \frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{x} d(\mu_0(t + \delta)) \log \frac{d(\mu_0(t + \delta))}{(1 - \delta_1)d(P_{\delta_0-\delta_1}(t + \delta))},
\]

which degenerates to the case of \( x < u_{-\delta} \) since (34) is the same as (33).

3) For \( u_\delta < x \leq u_{-\delta} \),

\[
D(\mu^x_\delta||P^x_{\delta_0-\delta_1}) = \int_{x^-}^{x^-} d(\mu_0(t - \delta) - \delta) \log \frac{d(\mu_0(t - \delta) - \delta)}{d((1 - \delta_1)P_{\delta_0-\delta_1}(t - \delta))}
+ (\mu_0(x + \delta) + \delta - (\mu_0(x - \delta) - \delta)) \log \frac{(\mu_0(x + \delta) + \delta - (\mu_0(x - \delta) - \delta))}{(1 - \delta_1)P_{\delta_0-\delta_1}(t + \delta) + \delta_1 - (1 - \delta_1)P_{\delta_0-\delta_1}(t - \delta)}
+ \int_{x^+}^{u_{-\delta}} d(\mu_0(t + \delta)) \log \frac{d(\mu_0(t + \delta))}{d((1 - \delta_1)P_{\delta_0-\delta_1}(t + \delta) + \delta_1)}
\]
Assume the weak convergence. Given Lemma A.7. From the above arguments, we have 

\[
\begin{align*}
\int_{u_1}^{u_2} d(\mu_0(t - \delta)) \log \frac{d(\mu_0(t - \delta))}{(1 - \delta_1)dP_{\delta_0 - \delta_1}(t - \delta)} + (2\delta + \mu_0(x + \delta) - \mu_0(x - \delta)) \log \frac{2\delta + \mu_0(x + \delta) - \mu_0(x - \delta)}{\delta_1 + (1 - \delta_1)(P_{\delta_0 - \delta_1}(t + \delta) - P_{\delta_0 - \delta_1}(t - \delta))} \\
+ \int_{x+}^{u_2} d(\mu_0(t + \delta)) \log \frac{d(\mu_0(t + \delta))}{(1 - \delta_1)dP_{\delta_0 - \delta_1}(t + \delta)} \\
\leq \int_{u_1}^{u_2} d(\mu_0(t - \delta)) \log \frac{d(\mu_0(t - \delta))}{(1 - \delta_1)dP_{\delta_0 - \delta_1}(t - \delta)} + 2\delta \log \frac{\delta}{\delta_1} + \int_{x-}^{u_2} d\mu_0(t) \log \frac{d\mu_0(t)}{(1 - \delta_1)dP_{\delta_0 - \delta_1}(t)} \\
+ \int_{x+}^{u_2} d(\mu_0(t + \delta)) \log \frac{d(\mu_0(t + \delta))}{(1 - \delta_1)dP_{\delta_0 - \delta_1}(t + \delta)} \\
= 2\delta \log \frac{\delta}{\delta_1} + \int_{u_1 - \delta}^{u_2 + \delta} d\mu_0(t) \log \frac{d\mu_0(t)}{(1 - \delta_1)dP_{\delta_0 - \delta_1}(t)} \\
= 2\delta \log \frac{\delta}{\delta_1} + (1 - 2\delta) \log \frac{1}{1 - \delta_1} + \int_{u_1 - \delta}^{u_2 + \delta} d\mu_0(t) \log \frac{d\mu_0(t)}{dP_{\delta_0 - \delta_1}(t)}.
\end{align*}
\]

when \( \delta \to 0 \), the above converges to 

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

4) Other symmetric cases can be solved similarly.

From the above arguments, we have 

\[
\limsup_{\delta \to 0} \sup_{x \in \mathbb{R}} D(\mu_0^\delta||B_L(P_0, \delta_0)) \leq \log \frac{1}{1 - \delta_1} + D(\mu_0||B_L(P_0, \delta_0 - \delta_1)).
\]

Notice that this is true for any \( \delta_1 \). Letting \( \delta_1 \to 0 \), we have 

\[
\limsup_{\delta \to 0} \sup_{x \in \mathbb{R}} D(\mu_0^\delta||B_L(P_0, \delta_0)) \leq \lim_{\delta_1 \to 0} \left( \log \frac{1}{1 - \delta_1} + D(\mu_0||B_L(P_0, \delta_0 - \delta_1)) \right) = \lim_{\delta_1 \to 0} D(\mu_0||B_L(P_0, \delta_0 - \delta_1)) = D(\mu_0||B_L(P_0, \delta_0))
\]

the last equality comes from Lemma A.2 \( D(\mu_0||B_L(P_0, \delta_0)) \) is left continuous in \( \delta_0 \) if \( P_0(t) \) is continuous in \( t \). \( \square \)

**Lemma A.7.** Given \( P_0 \in \mathcal{P} \) and \( \delta_0 > 0 \), \( D(\mu||B_L(P_0, \delta_0)) \) is lower semicontinuous in \( \mu \) with respect to the weak convergence.

**Proof.** Assume \( \mu_n \xrightarrow{w} \mu_0 \). From \( (26) \), we know there exists \( P_n \in B_L(P_0, \delta_0) \) such that \( D(\mu_n||P_n) = D(\mu_0||B_L(P_0, \delta_0)) \). Since \( B_L(P_0, \delta_0) \) is compact, there exists a subsequence of \( P_n \) (which we again denote by \( P_n \)) that converge to \( P_{\mu_0} \in B_L(P_0, \delta_0) \). \( D(\mu||P_{\mu_0}) \leq \liminf_{n \to \infty} D(\mu_n||P_n) \) because \( (\mu_n, P_n) \to (\mu_0, P_{\mu_0}) \) and the KLD is lower semi-continuous. Therefore we have 

\[
D(\mu_0||B_L(P_0, \delta_0)) = D(\mu_0||B_L(P_0, \delta_0)) \leq D(\mu_0||P_{\mu_0}) \leq \liminf_{n \to \infty} D(\mu_n||P_n) = \liminf_{n \to \infty} D(\mu_n||B_L(P_0, \delta_0))
\]

where \( (35) \) comes from \( (23) \). \( \square \)
It is straightforward to prove $D(\mu || B_L(P_0, \delta_0))$ is lower semicontinuous in $\mu$; proving that it is also upper semicontinuous is tricky. The key step of the long proof in this section is Lemma \[A.4\] which is explained below.

For a fixed $P_0$, with small perturbation on $\mu$, $D(\mu || P_0)$ 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 $\mu$, since the Lévy metric is the weakest among other metrics. For all perturbations on $\mu$ 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 $B_L(\mu, \delta)$. Such shifts can be tolerated by $B_L(P_0, \delta_0)$, so as the level of perturbation on $\mu$ 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 semicontinuous and lower semicontinuous in Lemma \[A.6\] and \[A.7\] we know that if $P_0(t)$ is continuous in $t$, $D(\mu || B_L(P_0, \delta_0))$ is continuous in $\mu$ with respect to the weak convergence.

REFERENCES

[1] I. Csiszár and P. C. Shields, “Information theory and statistics: a tutorial,” Commun. Inf. Theory, vol. 1, no. 4, pp. 417-528, 2004.
[2] T. M. Cover and J. A. Tomas, Elements of information theory, Wiley-Interscience, 1991.
[3] A. Dembo and O. Zeitouni, Large deviation techniques and applications, second edition, Springer, 1998.
[4] W. Hoeffding, “Asymptotically optimal tests for multinomial distributions,” Ann. Math. Statist., vol. 36, no. 2, pp. 369-401, Apr. 1965.
[5] 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.
[6] E. Posner, “Random coding strategies for minimum entropy,” IEEE Trans. Inf. Theory., vol. 21, no. 4, pp. 388-391, Jul. 1975.
[7] S. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Statistics, 22(1): 79-86, 3 1951.
[8] M. Loève, Probability theory, second edition, Van Nostrand, 1960.
[9] P. Billingsley, Convergence of probability measures, Wiley, Jul. 1999.e
[10] 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.
[11] P. Yang, “Robust Kullback-Leibler divergence and its applications in universal hypothesis testing and deviation detection,” Ph.D. dissertation, Dept. Elect. and Comp. Eng., Syracuse Univ., Syracuse, NY, 2016.
[12] M.S. Pinsker, Information and Information Stability of Random Variables and Processes (Transl.), San Francisco:Holden Day, 1964.
[13] I. Csizsár, “A simple proof of Sanov’s theorem,” Bulletin of the Brazilian Mathematical Society, vol. 37, no. 4, pp. 453-459, 2006.
[14] 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.