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

Pengfei Yang and Biao Chen

***Abstract*— Universal hypothesis testing (UHT) 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 (KL divergence), is known to be asymptotically optimal for distributions defined on finite alphabets. With continuous observations, however, the discontinuity of the KL divergence in the distribution functions results in significant complications for UHT. This paper introduces a robust version of the classical KL divergence, defined as the KL divergence from a distribution to the Lévy ball of a known distribution. This robust KL divergence is shown to be continuous in the underlying distribution function with respect to the weak convergence. The continuity property enables the development of an asymptotically optimal test for the university hypothesis testing problem with continuous observations. The optimality is in the same sense as that of the Hoeffding's test and stronger than that of Zeitouni and Gutman. Perhaps more importantly, the developed test statistic can be computed through convex programs, making it much more meaningful in practice. Numerical experiments are also conducted to evaluate its performance as compared with some kernel based goodness of fit test that has been proposed recently.**

***Index Terms*— Kullback-Leibler divergence, universal hypothesis testing, Lévy metric.**

I. Introduction

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

An important application of the KL divergence is in the so-called universal hypothesis testing (UHT): 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 decay rate subject to a constraint on the type-I error decay rate [4]. The test statistic of Hoeffding's detector is equivalent to the KL divergence between the empirical distribution and P_0 (see 8).

Hoeffding's result, however, does not generalize to the UHT with continuous alphabet. Clearly, computing empirical KL divergence 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 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 weakened optimality with a rather complicated detector.

The difficulty in dealing with continuous observations for UHT stems from the subtle but important distinction on the continuity property of the KL divergence with respect to the underlying distributions. With finite alphabet distributions, the KL divergence defined between two distributions is known to be continuous in the distribution functions. This is not the case for the KL divergence 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 KL divergence. As such, even when two distributions are arbitrarily close in terms of distribution functions, the KL divergence between them can be arbitrarily large, making the KL divergence unsuitable for the UHT problem with continuous observations.

This paper introduces a robust version of the classical KL divergence that utilizes the Lévy metric which, unlike the KL divergence, is a true distance metric for distributions. The robust KL divergence, defined as the KL divergence from a

Manuscript received May 12, 2017; revised October 10, 2018; accepted October 21, 2018. Date of publication November 9, 2018; date of current version March 15, 2019. This work was supported in part by the Air Force Office of Scientific Research under Grant FA9550-16-1-0077 and in part by the National Science Foundation under Grant CNS-1731237.

P. Yang was with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA. He is now with Point72 Asset Management, L.P., New York, NY 10022 USA (e-mail: ypengf@gmail.com).

B. Chen is with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA (e-mail: bichen@syr.edu).

Communicated by G. Moustakides, Associate Editor for Sequential Methods.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2018.2879057

distribution to a Lévy ball of a second distribution, is shown to be continuous in the first distribution - the involvement of the Lévy ball around the second distribution has a smoothing effect that ensures the continuity of the robust KL divergence. This continuity property is crucial in developing a test for the UHT that is similar in its form to Hoeffding's detector while attaining the desired asymptotic NP optimality. Not only is the optimality stronger than that of [5], but the test statistic is also much more intuitive and amenable to numerical evaluation.

The rest of the paper is organized as follows. Section II defines the KL divergence between two (sets of) distributions; introduces the concepts of weak convergence and the Lévy metric along with their connections; and reviews the UHT for the finite alphabet case. Section III defines the robust KL divergence and establishes the continuity property with respect to weak convergence. In Section IV, the large deviation approach by Zeitouni and Gutman [5] to the UHT for continuous distributions is first reviewed; a robust version of the UHT problem is then introduced and the asymptotically NP optimal test using the robust KL divergence is derived. Section V shows how the empirical robust KL divergence can be computed via convex programming and compares the performance with a most recently proposed kernel based goodness of fit test developed in [7], namely the kernel Stein discrepancy test. Section VI concludes this paper.

II. Preliminaries

A. The KL Divergence

The KL divergence was first introduced in [8] to quantify the divergence between two probability distributions. For finite alphabets, the KL divergence between a probability distribution $\mu = (\mu_1, \mu_2, \cdots, \mu_n)$ and another distribution $P = (p_1, p_2, \cdots, p_n)$ is defined as

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

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

$$D(\mu||P) = \int_{\mathcal{R}} d\mu \log \frac{d\mu}{dP}. \tag{2}$$

In the above definition, we have implicitly assumed that the two distributions are absolutely continuous with each other, leading to a bounded KL divergence. The KL divergence $D(\mu||P)$ is jointly convex for both discrete and continuous distributions [9]. For two sets of probability distributions, say Γ_1 and Γ_2, defined on the same probability space, the KL divergence between the two sets is defined to be the infimum of the KL divergence of all possible pairs of distributions, i.e.,

$$D(\Gamma_1||\Gamma_2) := \inf_{\gamma_1 \in \Gamma_1, \gamma_2 \in \Gamma_2} D(\gamma_1||\gamma_2). \tag{3}$$

B. Weak Convergence and the Lévy Metric

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 [10].

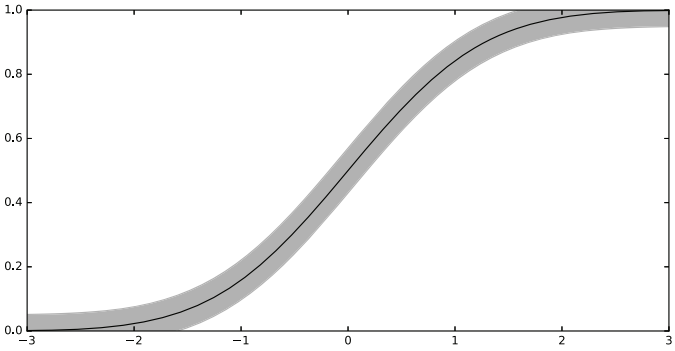

Fig. 1. The Lévy ball centered at standard normal distribution with radius 0.045.

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

Definition 1 (Weak Convergence [10], [11]): 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.

Definition 2 (Lévy Metric [10], [11]): The Lévy metric d_L between distributions $F \in \mathcal{P}$ and $G \in \mathcal{P}$ is denoted as $d_L(F, G) := \inf\{\epsilon : F(x - \epsilon) - \epsilon \leq G(x) \leq F(x + \epsilon) + \epsilon, \forall x \in \mathcal{R}\}$.

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

$$\begin{aligned} d_L(\mu, P) = 0 \Leftrightarrow \mu = P, \\ d_L(\mu, P) &= d_L(P, \mu), \\ d_L(\mu, P) &\leq d_L(\mu, Q) + d_L(Q, P). \end{aligned}$$

Definition 3: The Lévy ball centered at $P_0 \in \mathcal{P}$ with radius δ is denoted as

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

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 Lévy 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 4 [10], [11]: For sequences in $\mathcal{P}$ whose limit in weak convergence is also in $\mathcal{P}$, the weak convergence and convergence in the Lévy metric 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 $d_L(P_n, P) \to 0$.

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 μ. Consider the following hypothesis test

$$\mathcal{H}_0 : \mu = P_0, \quad \mathcal{H}_1 : \mu = Q, \tag{5}$$

where P_0 is a known distribution while $Q \neq P_0$ is defined on the same probability space as P_0 but is otherwise unknown. The fact that Q can be an arbitrary gives rise to the name UHT. Clearly, while any decision rule will be independent of Q, the type II error rate of the decision rule depends on Q.

In this paper, our main goal is to find an optimal detector under the generalized asymptotical Neyman-Pearson (NP) criterion [5]. Under this criterion, the optimal detector seeks the best trade-off between first and second type error decay rates. Specifically, let ϕ be the sequence of detectors $\{\phi^n(x^n), n \geq 1\}$. Define the error decay rates for the two types of error probabilities respectively as follows,

$$I^Q(\phi) := \liminf_{n\to\infty} -\frac{1}{n}\log Q(\phi^n(x^n)) = 0),$$
$$J^{P_0}(\phi) := \liminf_{n\to\infty} -\frac{1}{n}\log P_0(\phi^n(x^n)) = 1).$$

Zeitouni and Gutman [5] have shown that to achieve the best trade-off between I^Q and J^{P_0}, the test can depend on x^n only through the empirical measure $\hat{\mu}_n$, defined to be

$$\hat{\mu}_n(t) = \frac{\sum_i I_{\{x_i \leq t\}}}{n}. \tag{6}$$

Clearly, $\hat{\mu}_n$ belongs to $\mathcal{P}$. A consequence of the above result is that, any detector (which is a partition of the sample space $\mathcal{R}^n$) can now be equivalently characterized through a partition of the probability space $\mathcal{P}$ [5]. As such, a sequence of detectors can be equivalently expressed using $\mathbf{\Omega}$, which is a sequence of partitions $(\Omega_0(n), \Omega_1(n))$ $(n = 1, 2, \cdots)$, 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 $\hat{\mu}_n \in \Omega_i(n)$, $i = 0, 1$. Throughout the paper, we will use $\mathbf{\Omega}$ instead of ϕ to denote a sequence of detectors, which enables the use of the general Sanov's Theorem (Theorem 7) in proving our main result (Theorem 13).

Therefore $I^Q(\phi)$ and $J^{P_0}(\phi)$ can be written as

$$I^Q(\mathbf{\Omega}) = \liminf_{n\to\infty} -\frac{1}{n}\log Q(\hat{\mu}_n \in \Omega_0(n)),$$
$$J^{P_0}(\mathbf{\Omega}) = \liminf_{n\to\infty} -\frac{1}{n}\log P_0(\hat{\mu}_n \in \Omega_1(n)).$$

The generalized NP criterion maximizes the decay rate of type II probability of error under a constraint on the minimal decay rate for the type I probability of error:

$$\max_{\mathbf{\Omega}} I^Q(\mathbf{\Omega}) \quad \text{s.t. } J^{P_0}(\mathbf{\Omega}) \geq \eta. \tag{7}$$

The UHT was originally studied by Hoeffding [4] who considered distributions with finite alphabet. Hoeffding's detector is equivalent to the following threshold test of the empirical KL divergence:

$$D(\hat{\mu}_n||P_0) \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{8}$$

Hoeffding successfully established the asymptotical NP optimality of the above test. However, for continuous sample space such as $\mathcal{R}$, Hoeffding's detector (8) becomes degenerate since computing KL divergence between empirical distribution $\hat{\mu}_n$ to continuous distribution P_0 is meaningless. In addition, Hoeffding resorted to combinatorial bounds in establishing the asymptotic optimality which is not applicable to continuous observations. Zeitouni and Gutman [5] addressed the UHT with continuous observations via a large deviation approach; the obtained test, however, only achieves weakened optimality compared with Hoaffding's test and is numerically challenging to evaluate (see our review in Section IV-A). In the following, we give a formal definition of the robust KL divergence, and establish its continuous property. This allows us to tackle the UHT from a different perspective as described in Section IV-B.

III. Robust KL Divergence

A. Continuity Property of the KL Divergence

Let the nominal distribution be P_0. Let μ and μ_n, $n = 1, 2, \cdots$, be distributions with the same sample space as P_0. Suppose the sequence of distributions μ_n converge weakly to μ. It is of interest to study whether the corresponding KL divergence between μ_n and P_0 also converge to the KL divergence between μ and P_0. That is, does $\mu_n \xrightarrow{w} \mu$ imply $D(\mu_n||P_0) \to D(\mu||P_0)$?

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 μ that has the same sample support as P_0. Convergence in distribution implies convergence in the corresponding KL divergence.

This, however, is not the case for $P_0 \in \mathcal{P}$. Indeed, it was established in [6] that the KL divergence is only lower semicontinuous with respect to the weak convergence for the continuous case, i.e.,

$$D(\mu||P_0) \leq \liminf_{n\to\infty} D(\mu_n||P_0).$$

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

$$D(\mu||P_0) \geq \limsup_{n\to\infty} D(\mu_n||P_0).$$

Thus the KL divergence is not continuous in μ 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

$$\mu_n = \left(1 - \frac{1}{n}\right) P_0 + \frac{1}{n} P.$$

Clearly μ_n is also not absolutely continuous with respect to P_0, thus $D(\mu_n||P_0)$ is unbounded for any given $n > 0$. However, μ_n weakly converges to P_0 as $n \to \infty$. While this construction takes advantage of 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 (see arguments in proof of Theorem II.3 [12]).

This lack of continuity for the KL divergence for the continuous case is the primary reason for the difficulty in generalizing Hoeffding's result to continuous observations. A direct consequence of the lack of continuity is that the superlevel set defined by the KL divergence is not closed as observed in [5]. The superlevel set is given by

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

The fact that the KL divergence is not continuous in μ 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. In other words, while the sequence of distributions may be arbitrarily close to the nominal distribution (in the weak convergence sense), their KL divergence to the nominal distribution is bounded away from 0. As such, a test in a form similar to (8) can not be used for continuous distributions.

B. Robust KL Divergence and Its Continuity Property

Definition 5: The robust KL divergence between μ and P_0 with the radius parameter $\delta_0 > 0$ is denoted as

$$D(\mu||B_L(P_0,\delta_0)) := \inf_{P\in B_L(P_0,\delta_0)} D(\mu||P), \tag{10}$$

where $B_L(P_0,\delta_0)$ is the Lévy ball centered at P_0 with radius δ_0.

The following theorem establishes its continuity property in μ under some mild assumptions.

Theorem 6: 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 complete proof is lengthy and deferred to Appendix A. The non-trivial part of the proof is to show that $D(\mu||B_L(P_0,\delta_0))$ is upper semicontinuous in μ (Lemma 25). Lemma 26 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. Important intermediate steps are summarized below.

1) We first partition (quantize) the real line into a set of finite intervals. The robust KL divergence corresponding to the quantized distributions converge to the true 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 20).
2) The infimum defined in (10) is attained by a distribution either inside or on the surface of the Lévy ball (see Eq. (27) Lemma 20).
3) The robust KL divergence is continuous in the radius of the Lévy ball (Lemma 21).
4) The robust KL divergence and the quantized robust KL divergence are convex functions of the respective distributions (Lemma 22).
5) The supremum of the robust KL divergence over a Lévy ball centered at the first distribution μ 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 23).
6) The robust KL divergence is upper bounded by (Lemma 24)
$$\sup_{\mu,P_0\in\mathcal{P}} D(\mu||B_L(P_0,\delta_0)) = \log\frac{1}{\delta_0}.$$
7) The supremum of the robust KL divergence over a Lévy ball around μ converges to the robust KL divergence as the Lévy ball diminishes, i.e., as its radius goes to 0. Therefore, the robust KL divergence is upper semicontinuous in μ (Lemma 25).
8) The robust KL divergence is lower semicontinuous (Lemma 26).

The intuition of the continuity property of the robust KL divergence is as follows. The classical KL divergence 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 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 where the KL divergence 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 KL divergence.

The continuity property in Theorem 6 does not hold if the distribution ball is constructed using some other measures, including the total variation and the KL divergence (see [14, Sec. 2.2]). The assumption that $P_0(t)$ is continuous in t is also necessary for the continuous property of the robust KL divergence 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 μ_n be the distribution such that $\mu_n(t) = 0$ for $t < 0.5 + \frac{1}{n}$ and $\mu_n(t) = 1$ for $t \geq 0.5 + \frac{1}{n}$. Thus $\mu_n \xrightarrow{w} \mu$ as $n \to \infty$, 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 $\mu_i \xrightarrow{w} \mu$,

$$\lim_{n\to\infty} D(\mu_n||B_L(P_0,0.5)) = \lim_{n\to\infty} \log\frac{1}{0.5} > D(\mu||B_L(P_0,0.5)),$$

and the distribution in $B_L(P_0, 0.5)$ achieving the KL divergence value of $\log 2$ is a degenerate one: it takes values of the two points 0.5 and $0.5 + 1/n$ with equal probability.

The proof of Theorem 6 sheds some light on the dynamics of the KL divergence of continuous distributions. Furthermore, the established continuity property of the robust KL divergence is key to solving the robust version of the UHT problem for the continuous case.

IV. Robust Universal Hypothesis Testing

A. Review of the Large Deviation Approach

Motivated by the fact that Hoeffding's test does 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 Γ as $cl\Gamma$ and $int\Gamma$, respectively.

Theorem 7 (General Sanov's Theorem [3, Th. 6.2.10]): Given a probability set $\Gamma \subseteq \mathcal{P}$, for a probability measure $Q \notin \Gamma$,

$$\begin{aligned}\inf_{P\in cl\Gamma} D(P||Q) &\le \liminf_{n\to\infty} -\frac{1}{n}\log Q(\{x^n : \hat{\mu}_n \in \Gamma\}) \\ &\le \limsup_{n\to\infty} -\frac{1}{n}\log Q(\{x^n : \hat{\mu}_n \in \Gamma\}) \\ &\le \inf_{P\in int\Gamma} D(P||Q),\end{aligned}$$

where $\hat{\mu}_n$ is the empirical distribution defined in (6).

The general Sanov's Theorem illustrates the large deviation principle for the empirical measures and was used extensively in the proof of Theorems 8 [5]. For any set $\Gamma \subseteq \mathcal{P}$, define its $\delta-$smooth set to be

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

It is apparent that a δ-smoothed set is an open set. The major contribution of [5] is summarized in the Theorem below.

Theorem 8 [5]: Define Λ as,

$$\Lambda_1 = \{\mu : D(B_L(\mu, 2\delta)||P_0) \ge \eta\}^\delta, \tag{11}$$

$$\Lambda_0 = \mathcal{P} \setminus \Lambda_1. \tag{12}$$

Λ is $\delta-$optimal, i.e.,

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

$$I^Q(\Omega^\delta) \le I^Q(\Lambda). \tag{13}$$

The detector in Theorem 8 has weakened optimality compared to Hoeffding's test; it is also numerically challenging to construct such a test. See our detailed remarks below.

Remark 9: Theorem 8 applies to both discrete and $\mathcal{R}$-valued random variables. However, for the finite alphabet case, the corresponding detector in (11) yields a weaker result than Hoeffding's detector [4], a price paid for its generality.

Remark 10: In addition to weakened optimality, perhaps the most important drawback in Theorem 8 is the complexity of the detector. Computing $D(\mu||P_0)$ is meaningless if μ is an empirical distribution (hence discrete) and P_0 a continuous distribution, Theorem 8 works around this issue by computing the KL divergence from $B_L(\mu, 2\delta)$ to P_0 where the Lévy ball around the empirical distribution include continuous distributions when make the computation meaningful. However, finding a continuous distribution within the Lévy ball around μ that minimizes the KL divergence to P_0 is an infinite dimension minimization problem that is numerically prohibitive to evaluate. This is illustrated in Fig. 2 where one needs to find a continuous μ^* inside the shaded region such that

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

Remark 11: Even if evaluating (or approximating) $D(B_L(\hat{\mu}_n, 2\delta)||P_0)$ can be accomplished, the detector described in Theorem 8 may still encounter more complications. In particular, in the case when $D(B_L(\hat{\mu}_n, 2\delta)||P_0) < \eta$, one needs to further check if $\hat{\mu}_n$ belongs to the δ-smooth set of $\{\mu : D(B_L(\mu, 2\delta)||P_0) \ge \eta\}$, which is also computationally challenging.

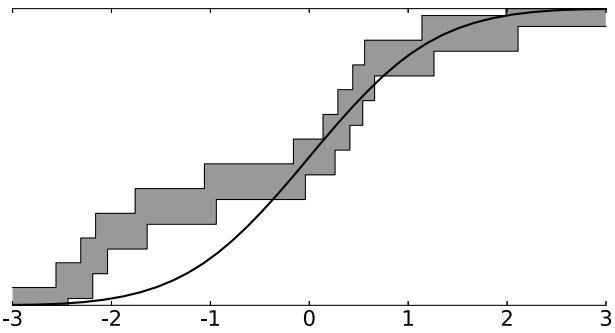

Fig. 2. The shaded region is $B_L(\hat{\mu}_n, 2\delta)$ and the solid line is P_0.

Remark 12: For a test Ω defined by the partition of the probability space (Ω_1, Ω_2), either Ω_1 or Ω_2, may consist of only empirical distributions (the optimal test can depend on the observations only through the empirical distributions). Suppose Ω_1 consists of only empirical distributions. Then $int\Omega_1$ is empty and $cl\Omega_2$ 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 decay rates. That is why in Theorem 8, for an arbitrary test Ω, we need to first perform $\delta-$smooth operation on it before comparing its error decay rates to those of the test Λ.

One of the difficulties in generalizing the discrete case to the continuous case, as first pointed out in [5], is that the superlevel set $\{\mu \in \mathcal{P} : D(\mu||P_0) \ge \eta\}$ is not closed in $\mathcal{P}$. This has been discussed in detail in Section III-A. In the following, instead of implementing the "$\delta-$smooth" operation on the detector as in (11), we generalize the original hypothesis test by considering a robust version of it. Specifically, $\mathcal{H}_0$, the single distribution P_0 is replaced by a small Lévy ball centered around P_0, i.e., $B_L(P_0, \delta_0)$. The introduction of the Lévy ball in place of the single distribution P_0 is itself an arguably more meaningful version for UHT: in practice, it is probably more important to tell if the sequence significantly departs from the nominal distribution P_0 and the radius δ_0 of the Lévy ball can be used to quantify the significance level. With this formulation of the UHT, we show in the following that the empirical likelihood ratio is asymptotically optimal under the minimax criterion.

B. Robust Universal Hypothesis Testing

The robust UHT problem is defined as followings. Givens samples drawn from a distribution μ, consider two hypotheses,

$$\mathcal{H}_0 : \mu \in \mathcal{P}_0, \quad \mathcal{H}_1 : \mu = Q, \tag{14}$$

where $\mathcal{P}_0 := B_L(P_0, \delta_0)$, P_0 is a known continuous distribution, $\delta_0 > 0$, and $Q \notin \mathcal{P}_0$ but is otherwise unknown.

The goal is to find the optimal detector under the asymptotic NP criterion:

$$\max_{\Omega} I^{Q}(\Omega) \quad \text{s.t. } J^{\mathcal{P}_0}(\Omega) \geq \eta, \tag{15}$$

where

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

Therefore, the goal is to maximize the decay rate of type II error probability when the worst case type I error probability has a decay rate that is bounded below by η.

The reason that the Lévy metric is used to define $\mathcal{P}_0$ is that the Lévy metric is the weakest hence also the most general one [13]. In another word, $B_L(P_0, \delta_0)$ 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, which might not be the case if $\mathcal{P}_0$ is constructed using other metrics. Theorem 13 below describes the optimal solution to the robust UHT, the proof of which can be found in Appendix B.

Theorem 13: For the robust UHT problem in (15). The detector $\Lambda = \{\Lambda_0, \Lambda_1\}$ *defined by,*

$$\Lambda_1 = \{\mu : D(\mu||\mathcal{P}_0)) > \eta\}, \quad \Lambda_0 = \mathcal{P} \setminus \Lambda_1, \tag{16}$$

satisfies the following properties:

1) $J^{\mathcal{P}_0}(\Lambda) = \eta$.
2) $I^{Q}(\Lambda) = D(\Lambda_0||Q)$.
3) *For any detector* Ω *with* $\Omega_1(n) = \Omega_1$ *with* Ω_1 *open, if*

$$J^{\mathcal{P}_0}(\Omega) > \eta, \tag{17}$$

then for any $Q \notin B_L(P_0, \delta_0)$,

$$I^{Q}(\Omega) \leq I^{Q}(\Lambda). \tag{18}$$

Remark 14: Theorem 13 states that the empirical likelihood ratio test,

$$D(\hat{\mu}_n||B_L(P_0, \delta_0) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{19}$$

achieves the optimal 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 Λ. In particular, the optimal type II error decay rate is precisely $D(\Lambda_0||Q)$ when Q is the true distribution under H_1.

Remark 15: 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 $\hat{\mu}_n$ (see Fig. 3). This can be shown to be a convex optimization problem with linear constraints in Section V-A, thus can be readily solved via standard convex programs. This contrasts with that of Theorem 8 where the detector involves an infinite dimension optimization problem (see Remark 2).

Remark 16: From Theorem 8, the detector developed in [5] can not be compared directly to an arbitrary detector Ω; instead, Ω^{δ} is used in establishing the optimality of the proposed detector. This ensures that Ω_1^{δ} is open and Ω_0^{δ} is closed yet this leads to a weaker sense of optimality, i.e., δ-optimality.

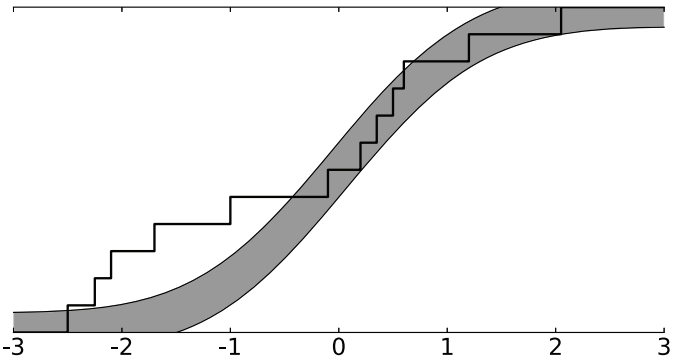

Fig. 3. The shaded region is a Lévy ball of the normal distribution and the step function is an example of the empirical distribution $\hat{\mu}_n$.

In Theorem 13, by restricting Ω_1 to be independent of n and assuming Ω_1 is open, asymptotic NP optimality is established which is stronger than $\delta-$optimality.

Remark 17: Theorem 8 only provides lower bounds for error decay rates, while Theorem 13 characterizes the exact values of error decay rates. Furthermore, while the error decay rates 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 decay rates. Therefore, Theorem 13 gives an exact characterization of error decay rates.

Summarizing, by considering the UHT 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. In the following, we address the computation of the developed test statistic and compares its performance to a recently proposed kernel based goodness of fit test.

V. Numerical Experiments and Performance Comparison

Theorem 13 established that the empirical robust KL divergence $D(\hat{\mu}_n||B_L(P_0, \delta_0))$ is the optimal statistic for the robust UHT under the asymptotic minimax NP criterion. The robust KL divergence between the empirical distribution and the Lévy ball around the nominal distribution is a finite dimension optimization problem (see Remark 6). We further establish in this section that this can be reformulated as a convex optimization problem hence can be readily solved using a convex program. Subsequently, we conduct numerical experiments to compare the proposed test with a recently proposed kernel based test in the context of the goodness of fit test.

A. Computing the Empirical Robust KL Divergence

Denote by $(x_0, x_1, \cdots, x_{n-1})$ the length-n sequence of samples, and without loss of generality, these samples are arranged in ascending order. As such, computing $D(\hat{\mu}_n||B_L(P_0, \delta_0))$ is equivalent to the following optimization problem,

$$\underset{P}{\text{minimize}} \sum_{i=0}^{n-1} \frac{1}{n} \log \frac{1/n}{p_i} \tag{20a}$$

$$\text{s.t. } P \in B_L(P_0, \delta_0), \tag{20b}$$

where $p_i = P_r(X = x_i) = P(x_i) - P(x_i-)$, i.e., the point mass at x_i for the discrete distribution P that is within the Lévy ball $B_L(P_0, \delta_0)$. Denote by $y_i = P(x_i)$ and $y_{-1} = 0$. Since P is non-decreasing and bounded below by the lower bound of $B_L(P_0, \delta_0)$ at all value of x_i, $P(x_i-)$ must be greater than or equal to $\max(y_{i-1}, l_i)$, where $l_i = \max(P_0(x_i - \delta_0) - \delta_0, 0)$, i.e., it is the lower boundary point of the Lévy ball at x_i. Thus we can rewrite the above optimization problem (20) as the following,

$$\underset{\mathbf{y},\mathbf{p}}{\text{minimize}} \sum_{i=0}^{n-1} \frac{1}{n} \log \frac{1/n}{p_i} \tag{21a}$$

$$\text{s.t. for } 0 \le i \le n-1,$$

$$l_i \le y_i \le u_i, \tag{21b}$$

$$0 \le p_i \le y_i - \max(y_{i-1}, l_i), \tag{21c}$$

where $u_i = \min(P_0(x_i + \delta_0) + \delta_0, 1)$, i.e., the upper boundary point of the Lévy ball at x_i. Unfolding condition (21c) to linear constraints, we have the following convex optimization problem,

$$\underset{\mathbf{y},\mathbf{p}}{\text{minimize}} \sum_{i=0}^{n-1} \frac{1}{n} \log \frac{1/n}{p_i} \tag{22a}$$

$$\text{s.t. for } 0 \le i \le n-1,$$

$$l_i \le y_i \le u_i, \tag{22b}$$

$$y_i - p_i - y_{i-1} \ge 0, \tag{22c}$$

$$y_i - p_i - l_i \ge 0, \tag{22d}$$

$$p_i \ge 0. \tag{22e}$$

Therefore, problem (20) of searching distribution P within $B_L(P_0, \delta_0)$, is reduced to a $2n$ dimensional convex optimization problem with separable convex objective functions and linear constraints, of which numerical solutions can be readily obtained via standard convex programs.

B. Goodness of Fit Test

The UHT described in (5) is also known to be the goodness of fit test which aims to determine whether the observed data samples are consistent with a nominal distribution P_0. The goodness of fit test has broad applications in various statistical data analysis, e.g., in regression analysis.

Notice that the our proposed test is derived for the robust version of the UHT where the null hypothesis is replaced by a Lévy ball centered at the nominal distribution. Nevertheless, the test statistic can be directly applied to the original UHT where δ_0 becomes a tuning parameter. The fact that this test statistic is still applicable to the original UHT problem is made clear by the following Lemma, whose proof is given in Appendix C.

Corollary 18: Given $P_0 \in \mathcal{P}$, if $P_0(t)$ is continuous in t, $D(\hat{\mu}_n || B_L(P_0, \delta_0)) \xrightarrow{a.s.} D(\mu || B_L(P_0, \delta_0))$ as $n \to \infty$.

Therefore, if $\mu = P_0$, $D(\hat{\mu}_n || B_L(P_0, \delta_0)) \xrightarrow{a.s.} 0$. On the other hand, if $\mu = Q \notin B_L(P_0, \delta_0)$, $D(\hat{\mu}_n || B_L(P_0, \delta_0))$ converges almost surely to $D(Q || B_L(P_0, \delta_0))$. Note that for the goodness of fit test (or the original UHT), we do not have prior knowledge on Q. Thus δ_0 needs to be chosen to be sufficiently small such that $Q \notin B_L(P_0, \delta_0)$.

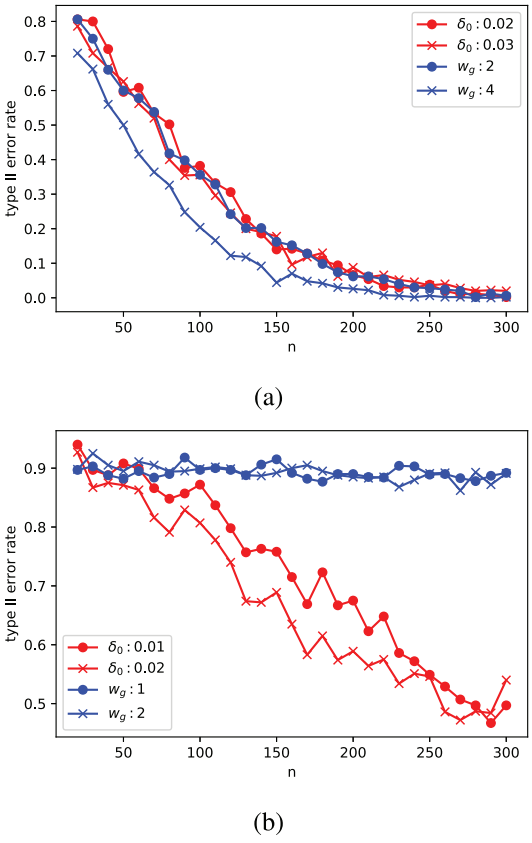

Fig. 4. Type II error rates. (a) Gaussian vs. Laplace. (b) t-distribution vs. t-distribution.

While there exists many classical tests for goodness of fit, there has been much recent effort in constructing kernel based test from the machine learning community. In the following, we compare the performance of the proposed test with one of the most recently proposed test, the kernel Stein discrepancy (KSD) test [7] that has been shown to exhibit superior performance.

KSD test is based on Stein transformed Reproducing Kernel Hilbert Space (RKHS) functions. The squared KSD is the norm of the smoothness-constrained function with largest expectation under the distribution of samples, which is defined as $E_{x\sim\mu} E_{x'\sim\mu} h_p(x, x')$, where h_p is a function dependent on the kernel of RKHS and probability density function of P_0. The squared KSD has zero expectation *if and only if* $\mu = P_0$, under some conditions.

For the following two experiments, we control the type I error probability to be 0.1 for all tests and evaluate type II error probabilities using 1000 trials. Notice that we are not examining the error decay rate but rather demonstrate the practicality of the developed test with only finite sample test. For the proposed robust KL divergence method, the threshold is estimated by drawing samples under the model P_0, i.e., using the Monte Carlo method. For the KSD method, the threshold is chosen using the wild bootstrap procedure as in the original paper [7]. The corresponding type II errors are shown in Fig. 4, where red lines represent robust KL

divergence method and blue lines represent KSD method with Gaussian kernels.

1) **Gaussian vs. Laplace.** We consider a problem in which $P_0 \sim \mathcal{N}(0, 2\sqrt{2})$ and $Q \sim Laplace(0, 2)$. P_0 and Q have the same mean and variance. We choose δ_0 to be 0.02 or 0.03 for the robust KL divergence method, and set Gaussian kernel width w_g to be 1 or 2 for the KSD method.
2) **t-distribution vs. t-distribution.** In this experiment P_0 is t-distribution with zero mean and degree of freedom 2, and Q is t-distribution with zero mean and degree of freedom 3. δ_0 is chosen to be 0.01 or 0.02, while Gaussian kernel width w_g is set to be 1 or 2.

As evident from the first experiment, the KSD method performs favorably compared to the proposed method when the selected kernel matches the distribution P_0. However, for the second example, the type II error probability of the KSD method decays in a much slower speed compared with the robust KLD based test (we also used Laplace kernels for KSD but the resulting type II errors are close to 1). One reason of the underperformance of the KSD method could be that there does not exist a natural choice of kernels for t-distributions. Indeed, extensive experiments showed that KSD is quite sensitive to the kernel choice whereas the proposed scheme exhibits rather robust performance.

VI. Conclusion

The KL divergence 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 the KL divergence 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 KL divergence 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 KL divergence, is known to be asymptotically NP optimal. 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 robust KL divergence, defined as the KL divergence between a distribution to the Lévy ball of a second distribution. In contrast to the classical KL divergence, this robust KL divergence 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 perturbation of the nominal distribution using a Lévy ball, it was established that the generalized empirical likelihood ratio test is optimal under the asymptotic minimax NP criterion whose error decay rates were characterized precisely. The developed test is also much easier to evaluate and exhibits robust performance when compared with some recently proposed kernel based tests.

Appendix A
Proof of Theorem 6

It is straightforward to prove $D(\mu||B_L(P_0, \delta_0))$ is lower semicontinuous in μ (Lemma 26); proving that it is also upper semicontinuous is more involved (Lemma 25). Before we can prove the upper semicontinuity, we will need several properties of the robust KL divergence (Lemmas 20 to 24).

We first introduce another widely used definition of KL divergence through partitions. This alternative definition is equivalent to the classical definition using the Radon-Nikodym derivative (see, e.g., [15 and 16, Sec. 2.4]).

A partition $\mathcal{A} = (A_1, \cdots, A_{|\mathcal{A}|})$ of $\mathcal{R}$ divides the real line into a finite number of sets A_i. The set of all finite partitions of $\mathcal{R}$ is denoted by Π. 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 $(P(A_1), P(A_2), \cdots, P(A_{|\mathcal{A}|})) \in \mathcal{R}^{|\mathcal{A}|}$.

Definition 19: The KL divergence between $P \in \mathcal{P}$ and $Q \in \mathcal{P}$ is defined as,

$$D(P||Q) = \sup_{\mathcal{A} \in \Pi} D(P^{\mathcal{A}}||Q^{\mathcal{A}}), \tag{23}$$

where

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

This definition will be used together with the classical Radon-Nikodym derivative throughout the appendix.

Eq. (23) states that the quantized KL divergence converges to the true KL divergence as the quantization becomes finer. The following lemma generalizes (23) from the classical KL divergence to the robust KL divergence. The proof is in essence proving that a max-min inequality is in fact an equality. For set $\Gamma \subseteq \mathcal{P}$, we define $\Gamma^{\mathcal{A}} := \{P^{\mathcal{A}} : P \in \Gamma\}$.

Lemma 20: 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: Let $\mathcal{M}$ denote the space of *finitely* additive[1] and non-negative set functions on $(\mathcal{R}, \mathcal{F})$ with $M(\mathcal{R}) = 1$ for $M \in \mathcal{M}$. Define $M(t) := M((-\infty, t])$, then 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. Clearly, $\mathcal{P} \subset \mathcal{M}$. The difference between $\mathcal{P}$ and $\mathcal{M}$ is that, for $P \in \mathcal{P}$ we have $P(-\infty) = 0$ and $P(+\infty) = 1$, while for $M \in \mathcal{M}$ we have $M(-\infty) \geq 0$ and $M(+\infty) \leq 1$. Another important difference is that $\mathcal{M}$ is compact with respect to the topology of weak convergence [10] (pg. 179), while $\mathcal{P}$ is not (a sequence of normal distributions with mean n and variance 1 does not converge weakly to any $P \in \mathcal{P}$).

For $F \in \mathcal{M}$ and $G \in \mathcal{M}$, Lévy metric d_L and the KL divergence are defined in exactly the same way as in Definition 2 and Definition 19.

[1]This contrasts with the probability function which requires countable additivity.

The following three steps constitute the proof of the lemma,

$$D(\mu||B_L(P_0,\delta_0)) = D(\mu||\bar{B}_L(P_0,\delta_0)), \quad (24)$$

$$D(\mu||\bar{B}_L(P_0,\delta_0)) = \sup_{\mathcal{A}\in\Pi} D(\mu^{\mathcal{A}}||\bar{B}_L^{\mathcal{A}}(P_0,\delta_0)), \quad (25)$$

$$\sup_{\mathcal{A}\in\Pi} D(\mu^{\mathcal{A}}||\bar{B}_L^{\mathcal{A}}(P_0,\delta_0)) = \sup_{\mathcal{A}\in\Pi} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0,\delta_0)), \quad (26)$$

where $\bar{B}_L(P_0,\delta_0) := \{P \in \mathcal{M} : d_L(P,P_0) \le \delta_0\}$. We first prove (24). 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

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

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 $P_\mu(-\infty) \ge \delta$ or $P_\mu(+\infty) \le 1-\delta$. We can assume $P_\mu(-\infty) = \delta$ and $P_\mu(+\infty) = 1$, as other cases can be proved in a similar manner. Let s denote the minimum t such that $P_0(t-\delta_0) \ge \delta_0$, we construct Q_μ as follows.

- If $P_\mu(s) = \delta$, let

$$Q_\mu(t) = \begin{cases} 0 & \text{if } t < s, \\ P_\mu(t) & \text{if } t \ge s. \end{cases}$$

  Since $\inf_t Q_\mu(t) = 0$ and $\sup_t Q_\mu(t) = 1$, $Q_\mu \in \mathcal{P}$. In addition, it can be easily verified that $d_L(Q_\mu,P_0) \le \delta_0$. Therefore, $Q_\mu \in B_L(P_0,\delta_0)$ and $D(\mu||Q_\mu) = D(\mu||P_\mu)$.
- If $P_\mu(s) > \delta$, let

$$Q_\mu(t) = \begin{cases} \dfrac{(P_\mu(t)-\delta)P_\mu(s)}{P_\mu(s)-\delta} & \text{if } t < s, \\ P_\mu(t) & \text{if } t \ge s. \end{cases}$$

  Again, since $\inf_t Q_\mu(t) = 0$ and $\sup_t Q_\mu(t) = 1$, $Q_\mu \in \mathcal{P}$. For $t < s$,

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

  then $d_L(Q_\mu,P_0) \le \delta_0$ because

$$P_0(t-\delta_0)-\delta_0 < 0 \le Q_\mu(t) \le P_\mu(t) \le P_0(t+\delta_0)+\delta_0.$$

  Therefore, we have $Q_\mu \in B_L(P_0,\delta_0)$. Also Q_μ achieves the infimum since,

$$\begin{aligned} &D(\mu||Q_\mu) \\ &= \int_{-\infty}^{s-} d\mu(t)\log\frac{d\mu(t)}{dQ_\mu(t)} + \int_s^\infty d\mu(t)\log\frac{d\mu(t)}{dQ_\mu(t)} \\ &= \mu(s-)\log\frac{P_\mu(s)-\delta}{P_\mu(s)} + \int_{-\infty}^{s-} d\mu(t)\log\frac{d\mu(t)}{dP_\mu(t)} \\ &\quad + \int_s^\infty d\mu(t)\log\frac{d\mu(t)}{dP_\mu(t)} \\ &= \mu(s-)\log\frac{P_\mu(s)-\delta}{P_\mu(s)} + D(\mu||P_\mu) \\ &\le D(\mu||P_\mu). \end{aligned}$$

Therefore in either case, there exists $Q_\mu \in B_L(P_0,\delta_0)$ such that

$$Q_\mu = \operatorname*{arg\,min}_{\{P\in\bar{B}_L(P_0,\delta_0)\}} D(\mu||P). \quad (27)$$

To prove (25), note that [17, eq. (2.5), Lemma 2.4] implies that

$$D(B_L(P_0,\delta_0)||\mu) = \sup_{\mathcal{A}\in\Pi} D(B_L^{\mathcal{A}}(P_0,\delta_0)||\mu^{\mathcal{A}}).$$

Using parallel proofs as in Lemma 2.3 and Lemma 2.4 in [17], one can establish that

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

i.e., (25) holds.

Finally, (26) holds since for any $\mathcal{A} \in \Pi$, $\bar{B}_L^{\mathcal{A}}(P_0,\delta_0) = B_L^{\mathcal{A}}(P_0,\delta_0)$. □

The robust KL divergence is continuous in the radius of the Lévy ball under a mild assumption. This property will be used towards proving Lemma 25.

Lemma 21: 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 continuous in δ_0 if $D(\mu||\{P \in \mathcal{P} : d_L(P,P_0) < \delta_0\}) = D(\mu||B_L(P_0,\delta_0))$. Clearly, $\{P \in \mathcal{P} : d_L(P,P_0) < \delta_0\} \subset B_L(P_0,\delta_0)$ implies

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

Thus we only need to show the other direction. Denote

$$P_\delta = \operatorname*{arg\,inf}_{\{P\in B_L(P_0,\delta)\}} D(\mu||P),$$
$$P_{\delta_0} = \operatorname*{arg\,inf}_{\{P\in B_L(P_0,\delta_0)\}} D(\mu||P).$$

The existence of P_δ and P_{δ_0} is guaranteed by (27). 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,

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

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:

$$\lim_{n\to\infty} D\left(\mu||B_L\left(P_0,\delta_0+\frac{1}{n}\right)\right) \ge D(\mu||B_L(P_0,\delta_0)). \quad (28)$$

From (27), 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,

$$\begin{aligned}\lim_{n\to\infty} D\left(\mu||B_L\left(P_0, \delta_0 + \frac{1}{n}\right)\right) &= \lim_{n\to\infty} D(\mu||P_n) \\ &\geq D(\mu||P^*) \\ &\geq D(\mu||\bar{B}_L(P_0, \delta_0)) \\ &= D(\mu||B_L(P_0, \delta_0)), \quad (29)\end{aligned}$$

where (29) comes from the fact that the KL divergence is lower semicontinuous and the last equality was proved in (24). Therefore $D(\mu||B_L(P_0, \delta_0))$ is right continuous in δ_0. □

The KL divergence is convex [9], so are the robust KL divergence and the quantized robust KL divergence. This is stated in the following lemma.

Lemma 22: 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. In addition, for any partition $\mathcal{A}$ of the real line, $D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0))$ is convex in $\mu^{\mathcal{A}}$.

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

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

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. □

To prove the upper semicontinuity of the robust KL divergence at μ_0, we need to characterize the supremum of $D(\mu||B_L(P_0, \delta_0))$ over $\mu \in B_L(\mu_0, \delta)$. The lemma below finds out that the supremum 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 23: 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

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

Proof: We have

$$\begin{aligned}&\sup_{\mu\in B_L(\mu_0,\delta)} D(\mu||B_L(P_0, \delta_0)) \\ &\quad= \sup_{\mu\in B_L(\mu_0,\delta)} \sup_{\mathcal{A}\in\Pi} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0)) \quad (30) \\ &\quad= \sup_{\mathcal{A}\in\Pi} \sup_{\mu\in B_L(\mu_0,\delta)} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0)) \\ &\quad= \sup_{\mathcal{A}\in\Pi} \sup_{\mu^{\mathcal{A}}\in B_L^{\mathcal{A}}(\mu_0,\delta)} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0)). \quad (31)\end{aligned}$$

Equality (30) comes from Lemma 20.

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

$$\mathcal{A} = \{(-\infty, a_1], (a_1, a_2], \cdots, (a_{n-1}, \infty)\}.$$

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

$$\mathcal{P}^{\mathcal{A}} = \{\mathbf{x} \in \mathcal{R}^n : \sum_i x_i = 1, 0 \leq x_i \leq 1 \text{ for } 1 \leq i \leq n\}.$$

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

$$B_L^{\mathcal{A}}(\mu_0, \delta) = \{\mathbf{x} \in \mathcal{P}^{\mathcal{A}} : L_j \leq \sum_{i=1}^{j} x_i \leq U_j \text{ for } 1 \leq j \leq n\text{-}1\},$$

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 $\mathbf{x}$ 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 $0 \leq x_i \leq 1$. If § is a vertex of $B_L^{\mathcal{A}}(\mu_0, \delta)$, then for some $1 \leq j \leq n-2$ we have $\sum_{i=1}^{j} x_i = U_j$. Since $U_j > L_{j+1}$ for any $1 \leq j \leq n-2$, if $\sum_{i=1}^{j} x_i = U_j$ for a vertex x and some $1 \leq j \leq n-2$, then for any $k > j$, $\sum_{i=1}^{k} x_i = U_k$.

Therefore there are n vertices $\mathbf{x}^1, \cdots, \mathbf{x}^n$ of $B_L^{\mathcal{A}}(\mu_0, \delta)$. And for each $\mathbf{x}^k$, $\sum_{i=1}^{j} x_i^k = L_j$ 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 22, $D(\cdot||B_L^{\mathcal{A}}(P_0, \delta_0))$ is a convex function, thus the supremum on the polytope $B_L^{\mathcal{A}}(\mu_0, \delta)$ is achieved at its vertices. Let

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

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

$$\begin{aligned}&\sup_{\mu^{\mathcal{A}}\in B_L^{\mathcal{A}}(\mu_0,\delta)} D(\mu^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0)) \\ &\quad= \max_k D(\mathbf{x}^k||B_L^{\mathcal{A}}(P_0, \delta_0)) \\ &\quad\leq \sup_x D((\mu_x^\delta)^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0)) \\ &\quad\leq \sup_{\mathcal{A}\in\Pi} \sup_{x\in\mathcal{R}} D((\mu_x^\delta)^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0)), \\ &\quad= \sup_{x\in\mathcal{R}} \sup_{\mathcal{A}\in\Pi} D((\mu_x^\delta)^{\mathcal{A}}||B_L^{\mathcal{A}}(P_0, \delta_0)), \\ &\quad= \sup_{x\in\mathcal{R}} D(\mu_x^\delta||B_L(P_0, \delta_0)). \quad (32)\end{aligned}$$

The last equality comes from Lemma 20. From (31) and (32), we have

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

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

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

Therefore,

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

□

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

Lemma 24: The robust KL divergence is bounded above, and its maximum is $\log\frac{1}{\delta_0}$*, i.e.,*

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

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 $t \ge 0$, then $\mathcal{P} = B_L(S_0, 1)$ since d_L is bounded by 1. According to Lemma 23,

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

where $\mu_x^1(t) = 0$ for $t < x$, and $\mu_x^1(t) = 1$ for $t \ge x$. Denote $P_0^u(x) = \min(1, P_0(x+\delta_0)+\delta_0)$ and $P_0^l(x) = \max(0, P_0(x-\delta_0)-\delta_0)$, we have

$$\begin{aligned}\sup_{x\in\mathcal{R}} D(\mu_x^\delta||B_L(P_0,\delta_0)) &= \sup_{x\in\mathcal{R}} \log\frac{1}{P_0^u(x)-P_0^l(x)}\\ &= \log\frac{1}{\delta_0},\end{aligned}$$

where the last equality comes from the fact that

$$P_0^u(x) - P_0^l(x) \ge \delta_0$$

and

$$\lim_{x\to\infty}(P_0^u(x) - P_0^l(x)) = \delta_0.$$

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

The intuition behind the proof of the following upper semicontinuity property is the following. 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. The Lévy ball $B_L(P_0,\delta_0)$ provides the necessary tolerance to the perturbation on μ, since the Lévy 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 on the edge of $B_L(\mu,\delta)$. Such shifts can be tolerated by $B_L(P_0,\delta_0)$, so that the level of perturbation on μ decreases to 0, and the corresponding variation in $D(\mu||B_L(P_0,\delta_0))$ diminishes.

Lemma 25: 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.

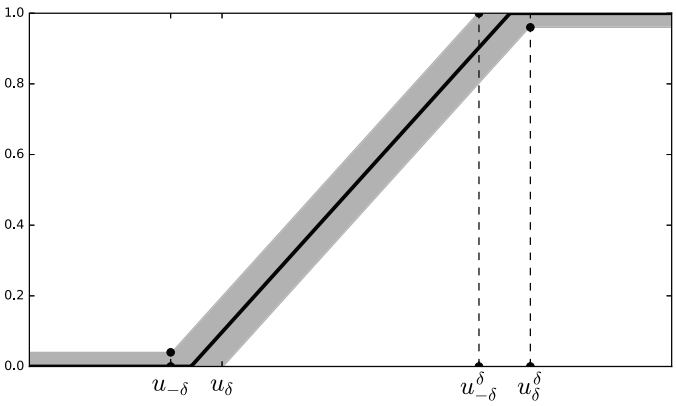

Fig. 5. Illustration of $u_{-\delta}, u_\delta, u_{-\delta}^\delta$ and u_δ^δ. The solid line represents μ_0 and shaded region represents $B_L(\mu_0,\delta)$.

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

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

From Lemma 23, it is equivalent to proving

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

Denote $u_{-\delta}$ as the left boundary of support set of distribution $\mu(t+\delta)$ and $u_{-\delta}^\delta$ as the infimum x such that $\mu(x+\delta) = 1-\delta$. Similarly, denote u_δ as the left boundary of distribution $\mu(t-\delta))$ and u_δ^δ as the infimum x such that $\mu(x-\delta) = 1$. Fig. 5 illustrates these locations. Note that these boundary points may not be finite. We will first prove that for any $\delta_1 \in (0,\delta_0)$,

$$\lim_{\delta\to 0}\sup_{x\in\mathcal{R}} D(\mu_x^\delta||B_L(P_0,\delta_0)) \le D(\mu_0||B_L(P_0,\delta_0-\delta_1)). \quad (33)$$

Now fix δ_1, we then establish that $D(\mu_x^\delta||B_L(P_0,\delta_0))$ can be uniformly bounded as x varies. Denote

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

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

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

For any x, construct $P_{\delta_0-\delta_1}^x$ in a similar manner as μ_x^δ,

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

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

$$\begin{aligned} d_L(P_{\delta_0-\delta_1}^x, P_0) &\le d_L(P_{\delta_0-\delta_1}^x, P_{\delta_0-\delta_1}) + d_L(P_{\delta_0-\delta_1}, P_0)\\ &\le \delta_1 + (\delta_0-\delta_1) = \delta_0,\end{aligned}$$

where the first inequality holds because $(\mathcal{P}, d_L)$ is a metric space (i.e., d_L satisfies the triangle inequality); the second inequality comes from (34) and the definition of $P_{\delta_0-\delta_1}$.

From Lemma 24, $D(\mu_0||P_{\delta_0-\delta_1}) = D(\mu_0||B_L(P_0, \delta_0 - \delta_1)) < \infty$. Therefore μ_0 is absolutely continuous with respect to $P_{\delta_0-\delta_1}$. From the construction of μ_x^δ and $P_{\delta_0-\delta_1}^x$, we can see that μ_x^δ is absolutely continuous with respect to $P_{\delta_0-\delta_1}^x$ as well. Therefore, we have

$$\begin{aligned}\lim_{\delta\to 0}\sup_{x\in\mathcal{R}} D(\mu_x^\delta||B_L(P_0,\delta_0)) &= \lim_{\delta\to 0}\sup_{x\in\mathcal{R}}\inf_{\{P\in B_L(P_0,\delta_0)\}} D(\mu_x^\delta||P)\\ &\le \lim_{\delta\to 0}\sup_{x\in\mathcal{R}} D(\mu_x^\delta||P_{\delta_0-\delta_1}^x),\end{aligned}$$

establishing (33). We now prove $D(\mu_x^\delta||P_{\delta_0-\delta_1}^x)$ can be uniformly bounded as x varies. Inequalities appear in cases 1)-3) are due to the log sum inequality (see [2, Ch. 2.7]) unless otherwise stated.

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

$$\begin{aligned}&D(\mu_x^\delta||P_{\delta_0-\delta_1}^x)\\ &\le \delta\log\frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u_{-\delta}^\delta} d(\mu_0(t+\delta)+\delta)\\ &\quad\times\log\frac{d(\mu_0(t+\delta)+\delta)}{d((1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1)}\\ &= \delta\log\frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u_{-\delta}^\delta} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{(1-\delta_1)d(P_{\delta_0-\delta_1}(t+\delta))}\\ &= \delta\log\frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u_{-\delta}^\delta} d(\mu_0(t+\delta))\log\frac{1}{(1-\delta_1)}\\ &\quad+\int_{u_{-\delta}}^{u_{-\delta}^\delta} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{d(P_{\delta_0-\delta_1}(t+\delta))}\\ &= \delta\log\frac{\delta}{\delta_1} + (1-\delta)\log\frac{1}{(1-\delta_1)}\\ &\quad+\int_{u_{-\delta}+\delta}^{u_{-\delta}^\delta+\delta} d(\mu_0(t))\log\frac{d(\mu_0(t))}{d(P_{\delta_0-\delta_1}(t))},\end{aligned}\tag{34}$$

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

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

2.) For $u_{-\delta}\le x\le u_\delta$,

$$\begin{aligned}&D(\mu_x^\delta||P_{\delta_0-\delta_1}^x)\\ &= (u_0(x+\delta)+\delta)\log\frac{(u_0(x+\delta)+\delta)}{(1-\delta_1)P_{\delta_0-\delta_1}(x+\delta)+\delta_1}\\ &\quad+\int_{x+}^{u_{-\delta}^\delta} d(\mu_0(t+\delta)+\delta)\log\frac{d(\mu_0(t+\delta)+\delta)}{d((1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1)}\\ &\le \delta\log\frac{\delta}{\delta_1} + (u_0(x+\delta))\log\frac{(u_0(x+\delta))}{(1-\delta_1)P_{\delta_0-\delta_1}(x+\delta)}\\ &\quad+\int_{x+}^{u_{-\delta}^\delta} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{(1-\delta_1)d(P_{\delta_0-\delta_1}(t+\delta))}\\ &\le \delta\log\frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{x} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{(1-\delta_1)d(P_{\delta_0-\delta_1}(t+\delta))}\\ &\quad+\int_{x+}^{u_{-\delta}^\delta} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{(1-\delta_1)d(P_{\delta_0-\delta_1}(t+\delta))}\\ &= \delta\log\frac{\delta}{\delta_1} + \int_{u_{-\delta}}^{u_{-\delta}^\delta} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t+\delta)}\\ &\quad+\int_{u_\delta}^{x} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t+\delta)}\end{aligned}\tag{35}$$

which degenerates to the case of $x < u_{-\delta}$ since (35) is the same as (34).

3.) For $u_\delta < x \le u_{-\delta}^\delta$,

$$\begin{aligned}&D(\mu_x^\delta||P_{\delta_0-\delta_1}^x)\\ &= \int_{u_\delta}^{x-} d(\mu_0(t-\delta)-\delta)\log\frac{d(\mu_0(t-\delta)-\delta)}{d((1-\delta_1)P_{\delta_0-\delta_1}(t-\delta))}\\ &\quad+(\mu_0(x+\delta)+\delta-(\mu_0(x-\delta)-\delta))\\ &\quad*\log\frac{(\mu_0(x+\delta)+\delta-(\mu_0(x-\delta)-\delta))}{(1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1-(1-\delta_1)P_{\delta_0-\delta_1}(t-\delta)}\\ &\quad+\int_{x+}^{u_{-\delta}^\delta} d(\mu_0(t+\delta)+\delta)\log\frac{d(\mu_0(t+\delta)+\delta)}{d((1-\delta_1)P_{\delta_0-\delta_1}(t+\delta)+\delta_1)}\\ &= \int_{u_\delta}^{x-} d(\mu_0(t-\delta))\log\frac{d(\mu_0(t-\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t-\delta)}\\ &\quad+(2\delta+\mu_0(x+\delta)-\mu_0(x-\delta))\\ &\quad*\log\frac{2\delta+\mu_0(x+\delta)-\mu_0(x-\delta)}{\delta_1+(1-\delta_1)(P_{\delta_0-\delta_1}(t+\delta)-P_{\delta_0-\delta_1}(t-\delta))}\\ &\quad+\int_{x+}^{u_{-\delta}^\delta} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t+\delta)}\\ &\le \int_{u_\delta}^{x-} d(\mu_0(t-\delta))\log\frac{d(\mu_0(t-\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t-\delta)}\\ &\quad+2\delta\log\frac{2\delta}{\delta_1}+\int_{x-\delta}^{x+\delta} d\mu_0(t)\log\frac{d\mu_0(t)}{(1-\delta_1)dP_{\delta_0-\delta_1}(t)}\\ &\quad+\int_{x+}^{u_{-\delta}^\delta} d(\mu_0(t+\delta))\log\frac{d(\mu_0(t+\delta))}{(1-\delta_1)dP_{\delta_0-\delta_1}(t+\delta)}\\ &= 2\delta\log\frac{2\delta}{\delta_1}+\int_{u_\delta-\delta}^{u_{-\delta}^\delta+\delta} d\mu_0(t)\log\frac{d\mu_0(t)}{(1-\delta_1)dP_{\delta_0-\delta_1}(t)}\\ &= 2\delta\log\frac{2\delta}{\delta_1}+(1-2\delta)\log\frac{1}{1-\delta_1}\\ &\quad+\int_{u_\delta-\delta}^{u_{-\delta}^\delta+\delta} d\mu_0(t)\log\frac{d\mu_0(t)}{dP_{\delta_0-\delta_1}(t)},\end{aligned}$$

which, as $\delta\to 0$, 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

$$\begin{aligned}&\lim_{\delta\to 0}\sup_{x\in\mathcal{R}} D(\mu_x^\delta||B_L(P_0,\delta_0))\\ &\le \log\frac{1}{1-\delta_1} + D(\mu_0||B_L(P_0,\delta_0-\delta_1)).\end{aligned}$$

Notice that this is true for any δ_1. Letting $\delta_1\to 0$, we have

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

the last equality comes from Lemma 21: $D(\mu_0||B_L(P_0,\delta_0))$ is left continuous in δ_0 if $P_0(t)$ is continuous in t. □

Lemma 26: 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 (27), 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 P_n) that converge to $P_{\mu_0} \in \bar{B}_L(P_0,\delta_0)$. $D(\mu||P_{\mu_0}) \le \liminf_{n\to\infty} D(\mu_n||P_n)$ because $(\mu_n, P_n) \to (\mu_0, P_{\mu_0})$ and the KL divergence is lower semi-continuous. Therefore we have

$$\begin{aligned} D(\mu_0||B_L(P_0,\delta_0)) &= D(\mu_0||\bar{B}_L(P_0,\delta_0)) \\ &\le D(\mu_0||P_{\mu_0}) \\ &\le \liminf_{n\to\infty} D(\mu_n||P_n) \\ &= \liminf_{n\to\infty} D(\mu_n||B_L(P_0,\delta_0)) \quad (36) \end{aligned}$$

where (36) comes from (24). □

As $D(\mu||B_L(P_0,\delta_0))$ is both upper semicontinuous (Lemma 25) and lower semicontinuous (Lemma 26), it is continuous in μ with respect to the weak convergence.

Appendix B
Proof of Theorem 13

Proof: The three parts of the Theorem 13 are proved below.

1) From the general Sanov's theorem, we have

$$\begin{aligned} \inf_{P\in\mathcal{P}_0} J^P(\Lambda) &= \inf_{P\in\mathcal{P}_0} \liminf_{n\to\infty} -\frac{1}{n}\log P(\{x^n : \hat{\mu}_n \in \Lambda_1\}) \\ &\ge \inf_{P\in\mathcal{P}_0} \inf_{\mu\in cl\Lambda_1} D(\mu||P) \\ &= \inf_{\mu\in cl\Lambda_1} D(\mu||\mathcal{P}_0) \\ &= \eta. \end{aligned}$$

The last equality holds since $D(\mu||\mathcal{P}_0)$ is continuous in μ thus $cl\Lambda_1 \subseteq \{\mu : D(\mu||\mathcal{P}_0) \ge \eta\}$. On the other hand,

$$\begin{aligned} &\inf_{P\in\mathcal{P}_0} J^P(\Lambda) \quad (37) \\ &\le \inf_{P\in\mathcal{P}_0} \limsup_{n\to\infty} -\frac{1}{n}\log P(\{x^n : \hat{\mu}_n \in \Lambda_1\}) \\ &\le \inf_{P\in\mathcal{P}_0} \inf_{\mu\in int\Lambda_1} D(\mu||P) \\ &= \eta. \quad (38) \end{aligned}$$

The last equality holds since $int\Lambda_1 = \Lambda_1$.

2) Again from the general Sanov's theorem, we have

$$\begin{aligned} I^Q(\Lambda) &= \liminf_{n\to\infty} -\frac{1}{n}\log Q(\{x^n : \hat{\mu}_n \in \Lambda_0\}) \\ &\ge \inf_{\mu\in cl\Lambda_0} D(\mu||Q) \\ &= D(\Lambda_0||Q). \end{aligned}$$

The last equality holds since $cl\Lambda_0 = \Lambda_0$. On the other hand, $\{\mu : D(\mu||\mathcal{P}_0) < \eta\} \subseteq int\Lambda_0$, thus,

$$\begin{aligned} I^Q(\Lambda) &\le \limsup_{n\to\infty} -\frac{1}{n}\log Q(\{x^n : \hat{\mu}_n \in \Lambda_0\}) \\ &\le \inf_{\mu\in int\Lambda_0} D(\mu||Q) \\ &\le \inf_{\mu\in\{\mu: D(\mu||\mathcal{P}_1)<\eta\}} D(\mu||Q) \\ &\le D(\Lambda_0||Q). \quad (39) \end{aligned}$$

Inequality (39) holds because of the following. There exists a distribution $P \in \mathcal{P}_0$ such that $D(P||Q) < \infty$. 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

$$\begin{aligned} &D((1-\lambda)P_c + \lambda P||\mathcal{P}_0) \\ &\le (1-\lambda)D(P_c||\mathcal{P}_0) + \lambda D(P||\mathcal{P}_0)) \quad (40) \\ &< (1-\lambda)\eta + 0 \quad (41) \\ &< \eta, \end{aligned}$$

where (40) comes from the fact that $D(\mu||\mathcal{P}_0)$ is convex in μ while inequality (41) is due to the fact $P \in \mathcal{P}_0$. Thus,

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

the last inequality holds since $D(P||Q) < \infty$. The above inequalities hold for any $P_c \in \Lambda_0$, thus we have

$$\inf_{\mu\in\{\mu: D(\mu||\mathcal{P}_0)<\eta\}} D(\mu||Q) \le D(\Lambda_1||Q).$$

3) We have

$$\begin{aligned} &\inf_{P\in\mathcal{P}_0} D(\Omega_1||P) \\ &= \inf_{P\in\mathcal{P}_0} D(int\Omega_1||P) \\ &\ge \inf_{P\in\mathcal{P}_0} \liminf_{n\to\infty} -\frac{1}{n}\log P(\{x^n : \hat{\mu}_n \in \Omega_1\}) \\ &> \eta. \quad (42) \end{aligned}$$

Therefore, $\Omega_1 \subseteq \Lambda_1$, or equivalently, $\Lambda_0 \subseteq \Omega_0$. Next,

$$\begin{aligned} I^Q(\Omega) &= \liminf_{n\to\infty} -\frac{1}{n}\log Q(\{x^n : \hat{\mu}_n \in \Omega_0\}) \\ &\le \liminf_{n\to\infty} -\frac{1}{n}\log Q(\{x^n : \hat{\mu}_n \in \Lambda_0\}) \\ &= I^Q(\Lambda). \end{aligned}$$

□

Appendix C
Proof of Corollary 18

Proof: Denote $C_\mu \subseteq \mathcal{R}$ as the continuity set of $\mu(t)$. From Theorem 6, $D(\mu||B_L(P_0,\delta_0))$ 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,

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

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

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. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1991.

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY, USA: Springer, 1998.

[4] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, no. 2, pp. 369–401, 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. IT-21, no. 4, pp. 388–391, Jul. 1975.

[7] K. Chwialkowski, H. Strathmann, and A. Gretton, "A kernel test of goodness of fit," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2606–2615.

[8] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[9] T. V. Erven and P. Harremoës, "Rényi divergence and kullback-leibler divergence," *IEEE Trans. Inf. Theory.*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

[10] M. Loève, *Probability Theory*, 2nd ed. Princeton, NJ, USA: Van Nostrand, 1960.

[11] P. Billingsley, *Convergence of Probability Measures*. Hoboken, NJ, USA: Wiley, 1999.

[12] P. F. Yang and B. Chen, "Deviation detection with continuous observations," in *Proc. IEEE GlobalSIP*, Orlando, FL, USA, Dec. 2015, pp. 537–541.

[13] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Statist. Rev.*, vol. 70, no. 3, pp. 419–435, Dec. 2002.

[14] P. Yang, "Robust Kullback-Leibler divergence and its applications in universal hypothesis testing and deviation detection," Ph.D. dissertation, Dept. Elect. Comp. Eng., Syracuse Univ., Syracuse, NY, USA, 2016.

[15] I. Csiszár, "A simple proof of Sanov's theorem," *Bull. Brazilian Math. Soc.*, vol. 37, no. 4, pp. 453–459, 2006.

[16] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, San Francisco, CA, USA: Holden Day, 1964.

[17] 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.

Pengfei Yang received his the B.S. degree in statistics from the University of Science and Technology of China, Hefei, China, and the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY, in 2010 and 2016, respectively. Since 2016, he has been a quantitative researcher working on statistical arbitrage trading strategies at Point72 Asset Management in New York, NY.

Biao Chen received the B.E. and M.E. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1992 and 1994, respectively. He joined the University of Connecticut, Storrs, CT, in 1995, where he received the M.S. degree in statistics and the Ph.D. degree in electrical engineering, in 1998 and 1999, respectively.

From 1994 to 1995, he worked at AT&T (China) Inc., Beijing, China. From 1999 to 2000, he was with Cornell University, Ithaca, NY, as a Postdoctoral Associate. Since 2000, he has been with Syracuse University, Syracuse, NY, where he is currently the John E. and Patricia A. Breyer Professor in the Department of Electrical Engineering and Computer Science. His area of interest mainly focuses on signal processing and information theory for wireless communications and sensor networks.

Prof. Chen has served as Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, Associate Editor for the IEEE COMMUNICATIONS LETTERS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the *EURASIP Journal on Wireless Communications and Networking* (JWCN). He is the recipient of an NSF CAREER Award in 2006.