Limit Distribution Theory for KL divergence and Applications to Auditing Differential Privacy

Sreejith Sreekumar
Cornell University
sreejithsreekumar@cornell.edu

Ziv Goldfeld
Cornell University
goldfeld@cornell.edu

Kengo Kato
Cornell University
kk976@cornell.edu

***Abstract*—The Kullback-Leibler (KL) divergence is a discrepancy measure between probability distribution that plays a central role in information theory, statistics and machine learning. While there are numerous methods for estimating this quantity from data, a limit distribution theory which quantifies fluctuations of the estimation error is largely obscure. In this paper, we close this gap by identifying sufficient conditions on the population distributions for the existence of distributional limits and characterizing the limiting variables. These results are used to derive one- and two-sample limit theorems for Gaussian-smoothed KL divergence, both under the null and the alternative. Finally, an application of the limit distribution result to auditing differential privacy is proposed and analyzed for significance level and power against local alternatives.**

I. Introduction

Statistical inference often boils down to estimation of a certain functional of the underlying probability measures. Discrepancy measures between probability distributions, also known as statistical divergences, such as f-divergences [1], [2], Rényi divergences [3], [4], integral probability metrics [5], [6], Wasserstein distances [7], [8], etc., form an important class of such functionals. Among the f-divergences, the Kullback-Leibler (KL) divergence is ubiquitous in information theory and statistics, naturally emerging as a quantifier of operational channel capacity and hypothesis testing problems [9], [10]. The KL divergence also has applications to modeling, analysis, and design of machine learning algorithms, including generative modeling [11], goodness-of-fit [12], model fusion [13] and anomaly detection [14].

In data-driven applications, one only has samples from the population distributions, which necessitates estimating the KL divergence. While there is an abundance of consistent estimators with known convergence rates, a limit distribution theory for the empirical estimation error has remained partial and premature—we close this gap in this paper.

A. Contributions

Denote the KL divergence between μ, ν by $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu)$, and let μ_n, ν_n be empirical estimates of μ, ν. Limit theorems seek to identify a scaling rate $r_n \to \infty$ and a limiting variable G such that the following convergence in distribution holds[1]

$$r_n\big(\mathsf{D}_{\mathsf{KL}}(\mu_n\|\nu_n) - \mathsf{D}_{\mathsf{KL}}(\mu\|\nu)\big) \xrightarrow{d} G.$$

As such, these results characterize the probability laws governing the random fluctuations of the error and serve as a central constituent for valid statistical inference. Indeed, distributional limits enable constructing confidence intervals, devising consistent resampling methods, proving guarantees for applications of hypothesis testing, and more.

To derive our main limit distribution result, we identify regularity conditions on the population distributions that allow bounding the second order term in a Taylor's expansion of KL divergence. This enables lifting weak convergence of the estimates of the distributions to convergence of the KL divergence between them, with the limiting variable identified in terms of the first or second order term in the expansion. We obtain the one- and two-sample distributional limits, under both the null ($\mu = \nu$) and the alternative ($\mu \neq \nu$) via this approach. The results hold under the high-level weak convergence assumptions on the empirical estimates of μ, ν with a given scaling law r_n.

To obtain limit theorems under basic conditions on the population distributions with explicit rates, we then consider the Gaussian-smoothed KL-divergence $\mathsf{D}_{\mathsf{KL}}(\mu * \gamma_\sigma\|\nu * \gamma_\sigma)$, where $\gamma_\sigma = \mathcal{N}(0, \sigma^2 I_d)$, and estimate μ, ν by the empirical measures $\hat{\mu}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_n = n^{-1}\sum_{i=1}^n \delta_{Y_i}$, respectively. Under this setup, we derive primitive conditions[2] on μ, ν that guarantees weak convergence of the smooth empirical measures $\hat{\mu}_n * \gamma_\sigma, \hat{\nu}_n * \gamma_\sigma$, utilizing the central limit theorem (CLT) in L^2 spaces [15, Proposition 2.1.11]. Under the null, we identify the scaling law as $r_n = n$ and the limiting variable as a weighted sum of independent and identically distributed (i.i.d.) χ^2 random variables. Under the alternative, we show that $r_n = \sqrt{n}$ and the limit is a centered Gaussian.

As an application of our limit distribution theory, we consider auditing ϵ-differential privacy (DP). An audit of a black-box privacy mechanism seeks to certify whether it satisfies a promised DP guarantee. While existing auditing methods are heuristic [16], [17] or lack in formal guarantees [18],

S. Sreekumar is partially supported by the NSF grant CCF-1740822. Z. Goldfeld is supported by NSF grants CCF-1947801, CCF-2046018, DMS-2210368, and the 2020 IBM Academic Award. K. Kato is supported by the NSF grants DMS-1952306, DMS-2014636, and DMS-2210368.

[1] In the one-sample case, ν is not estimated and the relevant convergence of interest is $r_n\big(\mathsf{D}_{\mathsf{KL}}(\mu_n\|\nu) - \mathsf{D}_{\mathsf{KL}}(\mu\|\nu)\big) \xrightarrow{d} G$.

[2] The condition is sharp in the one-sample null case; cf. Proposition $1(i)$.

we propose a principled hypothesis testing pipeline for DP auditing with a full asymptotic analysis of significance level and power against local alternatives. The key idea is to relax the ϵ-DP constraint to a KL divergence bound, which is further relaxed to the Gaussian-smoothed KL divergence via the data-processing inequality. We then test for the smooth KL divergence value and leverage our limit theorem for the significance and power analysis. Lastly, we establish a stability lemma that bounds the gap due to smoothing, namely $\big|\mathsf{D}_{\mathsf{KL}}(\mu * \gamma_\sigma \| \nu * \gamma_\sigma) - \mathsf{D}_{\mathsf{KL}}(\mu \| \nu)\big|$. This enables translating the audit to test for the KL divergence value itself, for which we show that any non-zero significance level with power 1 is asymptotically achievable.

B. Related Work

Statistical analysis of divergence estimators has been an active area of research in recent years. Convergence rates for various estimators, which subsumes entropy and mutual information as special cases, have been studied in [19]–[31] (see also references therein). Literature on limit distributions for f-divergences mainly focused on analyzing specific estimators on a case-by-case basis. In [32], limit distributions for f-divergences between maximum likelihood estimates of probability distributions over a certain parametric class is established, with the limit variable shown to be either normal or χ^2. The authors of [21] study plug-in methods of kernel density estimators and show asymptotic normality subject to high Hölder smoothness and compact support of the densities. The case when the density estimates are constructed via k-nearest neighbour techniques is treated in [33]. One-sample null distributional limits of Gaussian-smoothed TV distance and χ^2 divergence have been derived in [34] by invoking the CLT in $L^1(\mathbb{R}^d)$ and $L^2(\mathbb{R}^d)$, respectively. Limit distributions for plug-in estimators of entropy and mutual information in the discrete setting have been considered in [35], [36].

II. Preliminaries and Problem Setup

A. Notation

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a sufficiently rich probability space on which all random variables are defined. Let $(\mathfrak{S}, \mathcal{S})$ be a separable measurable space equipped with a σ-finite measure ρ. When $\mathfrak{S}$ is a topological space, we use $\mathcal{B}(\mathfrak{S})$ to denote the Borel σ-field on $\mathfrak{S}$. In the sequel, we adapt ρ on a case-by-case basis, but given ρ, all considered measures are assumed to be absolutely continuous with respect to (w.r.t.) it. For $\eta \ll \rho$, we write $p_\eta = d\eta/d\rho$ for the Radon-Nikodym derivative of η w.r.t. ρ. $\eta^{\otimes n}$ stands for the n-fold product measure, and δ_x represents the Dirac measure at x. We use $\mathcal{P}(\mathfrak{S})$ to denote the space of probability measures on $(\mathfrak{S}, \mathcal{S})$, leaving the σ-field implicit. When $\mathfrak{S} = \mathbb{R}^d$, we always take $\mathcal{S} = \mathcal{B}(\mathbb{R}^d)$ and $\mathcal{P}(\mathbb{R}^d)$ as the set of Borel probability measures. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\mu * \nu$ denotes the convolution of μ and ν; likewise, $f * g$ represents convolution of two measurable functions $f, g : \mathbb{R}^d \to \mathbb{R}$. We write $\gamma_\sigma = N(0, \sigma^2 I_d)$ for the centered Gaussian distribution on $\mathbb{R}^d$ with covariance matrix $\sigma^2 I_d$, and use $\varphi_\sigma = (2\pi\sigma^2)^{-d/2} e^{-\|\cdot\|^2/(2\sigma^2)}$ for the corresponding density. We say that $\mu \in \mathcal{P}(\mathbb{R}^d)$ is β-sub-Gaussian for $\beta \geq 0$, if $X \sim \mu$ satisfies $\mathbb{E}\big[e^{\alpha \cdot (X - \mathbb{E}[X])}\big] \leq \exp\big(\beta^2 \|\alpha\|^2/2\big)$, for all $\alpha \in \mathbb{R}^d$. Let $\xrightarrow{w}$ and $\xrightarrow{d}$ denote weak convergence of probability measures and convergence in distribution of random variables, respectively. For $1 \leq r \leq \infty$, let $L^r(\rho) = L^r(\mathfrak{S}, \mathcal{S}, \rho)$ be the space of all real-valued measurable functions f on $\mathfrak{S}$ such that $\|f\|_{r,\rho} := \big(\int_{\mathfrak{S}} |f|^r d\rho\big)^{1/r} < \infty$, with the usual identification of functions that are equal ρ-almost everywhere (a.e.). When ρ is the Lebesgue measure λ on $\mathbb{R}^d$, we use $\|\cdot\|_r$ to denote the corresponding L^r norm. $\|\cdot\|$ designates Euclidean norm.

B. Problem Setup

Let $\mu, \nu \in \mathcal{P}(\mathfrak{S})$ and consider a sequence $(\mu_n, \nu_n)_{n \in \mathbb{N}}$ of random probability measures[3] on $\mathfrak{S}$ such that $\mathbb{E}[\mu_n] = \mu$, $\mathbb{E}[\nu_n] = \nu$, and μ_n and ν_n converges weakly to μ and ν, respectively. Accordingly, (μ_n, ν_n) can be viewed as weakly convergent and unbiased estimators of the population distribution (μ, ν). Below, the one- and two-sample settings refer to when only μ or both (μ, ν) are approximated by μ_n or (μ_n, ν_n), respectively. Also recall that the terms 'null' and 'alternative' refer to when $\mu = \nu$ or $\mu \neq \nu$, respectively. Further, for two positive measures η_1, η_2, $\mathfrak{D}$ denotes the normed space defined as

$$\mathfrak{D} := \begin{Bmatrix} (g_1 - p_\mu, g_2 - p_\nu) : \, g_1, g_2 \in L^1(\rho), \\ \|(g_1 - p_\mu, g_2 - p_\nu)\|_{\mathfrak{D}} < \infty \end{Bmatrix},$$

where $\|(h_1, h_2)\|_{\mathfrak{D}} := \|h_1\|_{2,\eta_1} + \|h_2\|_{2,\eta_2}$.

III. Main Results

We present our limit distribution results for the KL divergence, first under a general setting and then specialized to the Gaussian-smoothed case. The proofs are omitted due to space constraints and can be found in the extended version [37].

A. General Setting

In the following, inequalities involving relative densities (e.g., $p_\mu > 0$) are understood as holding ρ-a.e. Let $(r_n)_{n \in \mathbb{N}}$ denote a diverging sequence, and q be a measurable function.

Theorem 1 (Limit distribution for KL divergence) *The following hold:*

(i) (One-sample null) Let $\mu_n \ll \mu = \rho$ be such that $\mathsf{D}_{\mathsf{KL}}(\mu_n \| \mu) < \infty$ almost surely (a.s.). If $r_n(p_{\mu_n} - 1) \xrightarrow{w} B$ in $L^2(\mu)$, then

$$r_n^2 \mathsf{D}_{\mathsf{KL}}(\mu_n \| \mu) \xrightarrow{d} \frac{1}{2} \int_{\mathfrak{S}} B^2 d\mu. \tag{1}$$

(ii) (One-sample alternative) Let $\mu \ll \nu = \rho$ and $\mu_n \ll \nu$ satisfy $p_\mu > 0$, $\log p_\mu \in L^2(\nu)$, $\mathsf{D}_{\mathsf{KL}}(\mu \| \nu) < \infty$, and $\mathsf{D}_{\mathsf{KL}}(\mu_n \| \nu) < \infty$ a.s. If $r_n(p_{\mu_n} - p_\mu) \xrightarrow{w} B$ in $L^2(\eta)$, where η has relative density $p_\eta = 1 + (1/p_\mu)$, then

$$r_n\big(\mathsf{D}_{\mathsf{KL}}(\mu_n \| \nu) - \mathsf{D}_{\mathsf{KL}}(\mu \| \nu)\big) \xrightarrow{d} \int_{\mathfrak{S}} B \log p_\mu d\nu. \tag{2}$$

[3] A random probability measure on $\mathfrak{S}$ is a map $\zeta : \Omega \times \mathcal{S} \to [0, 1]$ satisfying (i) for every $\mathcal{C} \in \mathcal{S}$, $\omega \to \zeta(\omega, \mathcal{C})$ is measurable from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$; and (ii) for every $\omega \in \Omega$, $\zeta(\omega, \cdot) \in \mathcal{P}(\mathfrak{S})$.

(iii) (Two-sample null) Let $\mu_n \ll \nu_n \ll \mu = \rho$ be such that $\mathsf{D}_{\mathsf{KL}}(\mu_n\|\nu_n) < \infty$, $p_{\nu_n} > 0$, and $p_{\mu_n}/p_{\nu_n} \le q$ a.s. Let $\eta_1 = \mu$ and η_2 be the measure with relative density $p_{\eta_2} = 1+q$. If $\big(r_n(p_{\mu_n}-1), r_n(p_{\nu_n}-1)\big) \xrightarrow{w} (B_1, B_2)$ in $\mathfrak{D}$, then

$$r_n^2 \mathsf{D}_{\mathsf{KL}}(\mu_n\|\nu_n) \xrightarrow{d} \frac{1}{2}\int_{\mathfrak{S}} (B_1 - B_2)^2 d\mu. \tag{3}$$

(iv) (Two-sample alternative) Let $\mu \ll \nu = \rho$ and $\mu_n \ll \nu_n \ll \nu$ satisfy $p_\mu > 0$, $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) < \infty$, $p_\mu, \log p_\mu \in L^2(\nu)$, $\mathsf{D}_{\mathsf{KL}}(\mu_n\|\nu_n) < \infty$, $p_{\nu_n} > 0$, and $p_{\mu_n}/p_{\nu_n} \le q$ a.s. Let η_1 and η_2 be measures with relative densities $p_{\eta_1} = 1 + (1/p_\mu)$ and $p_{\eta_2} = 1 + p_\mu + q$, respectively. If $\big(r_n(p_{\mu_n} - p_\mu), r_n(p_{\nu_n} - 1)\big) \xrightarrow{w} (B_1, B_2)$ in $\mathfrak{D}$, then

$$r_n\big(\mathsf{D}_{\mathsf{KL}}(\mu_n\|\nu_n) - \mathsf{D}_{\mathsf{KL}}(\mu\|\nu)\big) \xrightarrow{d} \int_{\mathfrak{S}} B_1 \log p_\mu d\nu - \int_{\mathfrak{S}} B_2 d\mu. \tag{4}$$

The proof of Theorem 1 identifies regularity conditions on the population distributions that allow bounding the second order term in a Taylor expansion of KL divergence. This enables translating weak convergence of the estimates of the distributions to that of the KL divergence between them. Note that the regularity assumptions in Theorem 1 are automatically satisfied in the one-sample case when $\mathfrak{S}$ is discrete (of finite cardinality) and $\mu \lll \nu$. Then, the multivariate CLT implies that (1)-(2) hold with $r_n = n^{1/2}$ and B as a Gaussian vector.

B. Gaussian-Smoothed KL Divergence

To obtain explicit scaling rates and distributional limits, we consider the Gaussian-smoothed KL divergence, i.e., the population objective is now $\mathsf{D}_{\mathsf{KL}}(\mu * \gamma_\sigma \| \nu * \gamma_\sigma)$ [38]. We estimate μ (or both μ and ν) from samples, while assuming that the Gaussian kernel is known. The Gaussian smoothing alleviates mismatched support issues that f-divergences often suffer from and gives rise to a well-posed empirical approximation setting. Henceforth, we assume $\mathfrak{S} = \mathbb{R}^d$ and $\mathcal{S} = \mathcal{B}(\mathbb{R}^d)$. Some preliminaries are due before stating the results.

In defining the empirical measures of μ and ν we allow arbitrary correlation between their samples, which is necessary for the application to auditing DP considered below. Let $(X, Y) \sim \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with X, Y marginals μ, ν, respectively. Set $\hat{\mu}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ as the empirical distribution of $(X_1, \ldots, X_n)$ and $\hat{\nu}_n = n^{-1}\sum_{i=1}^n \delta_{Y_i}$ as that of $(Y_1, \ldots, Y_n)$, where $(X_i, Y_i) \sim \pi$, $1 \le i \le n$, are pairwise i.i.d. Recalling that φ_σ is the density of γ_σ, the Lebesgue densities of $\hat{\mu}_n * \gamma_\sigma$ and $\hat{\nu}_n * \gamma_\sigma$ are $\hat{\mu}_n * \varphi_\sigma$ and $\hat{\nu}_n * \varphi_\sigma$, respectively. We study distributional limits of $\mathsf{D}_{\mathsf{KL}}(\hat{\mu}_n * \gamma_\sigma \| \nu * \gamma_\sigma)$ as well as its two-sample analogues, under the null and the alternative.

Our limit variables are characterized as integral forms of a certain Gaussian process, which is introduced next. Consider the 2-dimensional centered Gaussian process $(G_{\mu,\sigma}, G_{\nu,\sigma}) := \big(G_{\mu,\sigma}(x), G_{\nu,\sigma}(y)\big)_{(x,y)\in\mathbb{R}^d\times\mathbb{R}^d}$ with covariance function

$$\Sigma_{\mu,\nu,\sigma}\big((x,y),(\tilde{x},\tilde{y})\big) := \begin{bmatrix} \mathbb{E}\big[G_{\mu,\sigma}(x)G_{\mu,\sigma}(\tilde{x})\big] & \mathbb{E}\big[G_{\mu,\sigma}(x)G_{\nu,\sigma}(\tilde{y})\big] \\ \mathbb{E}\big[G_{\nu,\sigma}(y)G_{\mu,\sigma}(\tilde{x})\big] & \mathbb{E}\big[G_{\nu,\sigma}(y)G_{\nu,\sigma}(\tilde{y})\big] \end{bmatrix}, \tag{5}$$

where $\mathbb{E}\big[G_{\mu,\sigma}(x)G_{\mu,\sigma}(\tilde{x})\big] = \mathsf{cov}\big(\varphi_\sigma(x-X), \varphi_\sigma(\tilde{x}-X)\big)$, $\mathbb{E}\big[G_{\mu,\sigma}(x)G_{\nu,\sigma}(\tilde{y})\big] = \mathsf{cov}\big(\varphi_\sigma(x-X), \varphi_\sigma(\tilde{y}-Y)\big)$, and $\mathbb{E}\big[G_{\nu,\sigma}(y)G_{\nu,\sigma}(\tilde{y})\big] = \mathsf{cov}\big(\varphi_\sigma(y-Y), \varphi_\sigma(\tilde{y}-Y)\big)$. For $i, j \in \{1,2\}$, denote the (i,j)-th entry of $\Sigma_{\mu,\nu,\sigma}$ by $\Sigma^{(i,j)}_{\mu,\nu,\sigma}$. Note that each such entry depends only on two coordinates among $\big((x,y),(\tilde{x},\tilde{y})\big)$. Hence, by some abuse of notation, we omit the redundant coordinates and use the remaining coordinates in the same order they appear, e.g., $\Sigma^{(2,1)}_{\mu,\nu,\sigma}(y,\tilde{x})$ for $\Sigma^{(2,1)}_{\mu,\nu,\sigma}\big((x,y),(\tilde{x},\tilde{y})\big)$. Further, when $\nu = \mu$ (viz. $X \overset{d}{=} Y$), we denote $G_{\nu,\sigma}$ by $\tilde{G}_{\mu,\sigma}$ to avoid confusion with $G_{\mu,\sigma}$.

Proposition 1 (Limit distribution for Gaussian-smoothed KL divergence) *The following hold:*

(i) (One-sample null) If

$$\int_{\mathbb{R}^d} \frac{\mathsf{Var}_\mu\big(\varphi_\sigma(x-\cdot)\big)}{\mu * \varphi_\sigma(x)}\, dx < \infty, \tag{6}$$

then there exists a version of $G_{\mu,\sigma}$ such that $G_{\mu,\sigma}/\sqrt{\mu * \varphi_\sigma}$ is $L^2(\mathbb{R}^d)$-valued, and

$$n\mathsf{D}_{\mathsf{KL}}(\hat{\mu}_n * \gamma_\sigma \| \mu * \gamma_\sigma) \xrightarrow{d} \frac{1}{2}\int_{\mathbb{R}^d} \frac{G^2_{\mu,\sigma}(x)}{\mu * \varphi_\sigma(x)} dx, \tag{7}$$

where the limit can be represented as a weighted sum of i.i.d. χ^2 random variables with 1 degree of freedom. In particular, (6) *and* (7) *holds for β-sub-Gaussian μ with $\beta < \sigma$. Conversely, if* (6) *is violated, then we have $\liminf_{n\to\infty} n\mathbb{E}\big[\mathsf{D}_{\mathsf{KL}}(\hat{\mu}_n * \gamma_\sigma \| \mu * \gamma_\sigma)\big] = \infty$.*

(ii) (One-sample alternative) If (6) *holds, $\log\big(\mu * \varphi_\sigma / \nu * \varphi_\sigma\big) \in L^2(\nu * \varphi_\sigma)$, $\big\|(\nu * \varphi_\sigma)^2/\mu * \varphi_\sigma\big\|_\infty < \infty$, and*

$$\int_{\mathbb{R}^d} \frac{\mathsf{Var}_\mu\big(\varphi_\sigma(x-\cdot)\big)}{\nu * \varphi_\sigma(x)}\, dx < \infty,$$

then

$$n^{\frac{1}{2}}\big(\mathsf{D}_{\mathsf{KL}}(\hat{\mu}_n * \gamma_\sigma \| \nu * \gamma_\sigma) - \mathsf{D}_{\mathsf{KL}}(\mu * \gamma_\sigma \| \nu * \gamma_\sigma)\big) \xrightarrow{d} N\big(0, v_1^2(\mu,\nu,\sigma)\big), \tag{8}$$

where

$$v_1^2(\mu,\nu,\sigma) := \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Sigma^{(1,1)}_{\mu,\nu,\sigma}(x,y) \log\left(\frac{\mu * \varphi_\sigma(x)}{\nu * \varphi_\sigma(x)}\right) \times \log\left(\frac{\mu * \varphi_\sigma(y)}{\nu * \varphi_\sigma(y)}\right) dx\, dy.$$

In particular, (8) *holds for β-sub-Gaussian μ with $\beta < \sigma$ such that $\nu \ll \mu \ll \nu$ and $\|d\mu/d\nu\|_\infty \vee \|d\nu/d\mu\|_\infty < \infty$.*

(iii) (Two-sample null) If μ has compact support, then there exists a version of $G_{\mu,\sigma}$ and $\tilde{G}_{\mu,\sigma}$ such that $G_{\mu,\sigma}/\sqrt{\mu * \varphi_\sigma}$ and $\tilde{G}_{\mu,\sigma}/\sqrt{\mu * \varphi_\sigma}$ are $L^2(\mathbb{R}^d)$-valued, and

$$n\mathsf{D}_{\mathsf{KL}}(\hat{\mu}_n * \gamma_\sigma \| \hat{\nu}_n * \gamma_\sigma) \xrightarrow{d} \frac{1}{2}\int_{\mathbb{R}^d} \frac{\big(G_{\mu,\sigma}(x) - \tilde{G}_{\mu,\sigma}(x)\big)^2}{\mu * \varphi_\sigma(x)} dx, \tag{9}$$

where the limit can be represented as a weighted sum of i.i.d. χ^2 random variables with 1 degree of freedom.

(iv) (Two-sample alternative) If μ, ν have compact supports, then

$$n^{\frac{1}{2}}\big(\mathsf{D}_{\mathsf{KL}}(\hat{\mu}_n * \gamma_\sigma \| \hat{\nu}_n * \gamma_\sigma) - \mathsf{D}_{\mathsf{KL}}(\mu * \gamma_\sigma \| \nu * \gamma_\sigma)\big) \xrightarrow{d} N\big(0, v_2^2(\mu, \nu, \sigma)\big), \tag{10}$$

where

$$v_2^2(\mu,\nu,\sigma) := \sum_{1\le i,j\le 2} \int_{\mathbb{R}^d\times\mathbb{R}^d} \Sigma^{(i,j)}_{\mu,\nu,\sigma}(x,y)\tilde{L}_i(x)\tilde{L}_j(y)dx\,dy,$$

with $\tilde{L}_1 := \log(\mu * \varphi_\sigma / \nu * \varphi_\sigma)$ *and* $\tilde{L}_2 := -\mu * \varphi_\sigma / \nu * \varphi_\sigma$.

The proof of Proposition 1 hinges on Theorem 1 by identifying primitive conditions in terms of μ, ν, and σ that guarantee the regularity assumptions therein.

IV. Application to Auditing Differential Privacy

We consider the application of our limit distribution theory to auditing DP, which was introduced in [39] as an approach for quantifying privacy leakage of privatization mechanisms. We recall some DP notions that are relevant to our setting. Consider a set $\mathfrak{U}$ with a relation $\sim$ such that $u \sim v$, for $u, v \in \mathfrak{U}$, denotes that u and v are adjacent. In the DP context, $\mathfrak{U}$ is a set of databases, while $u \sim v$ denotes that u and v are adjacent databases, differing on a single entry. Let $\epsilon, \delta \ge 0$. A randomized (measurable) mechanism $f : \mathfrak{U} \to \mathbb{R}^d$ is

(i) ϵ-differentially private if $\mathbb{P}(f(u) \in \mathcal{T}) \le e^\epsilon \mathbb{P}(f(v) \in \mathcal{T})$ for every $u \sim v$ and $\mathcal{T} \in \mathcal{B}(\mathbb{R}^d)$;

(ii) ϵ-KL differentially private if $\mathsf{D}_{\mathsf{KL}}(\mu_u \| \mu_v) \le \epsilon$ for every $u \sim v$, where $\mu_u \in \mathcal{P}(\mathbb{R}^d)$ is the distribution of $f(u)$.

In addition, we say that a privacy mechanism is *ϵ-smoothed KL differentially private* if $\mathsf{D}_{\mathsf{KL}}(\mu_u * \gamma_\sigma \| \mu_v * \gamma_\sigma) \le \epsilon$ for every $u \sim v$, where $\sigma > 0$ is a pre-specified parameter. As ϵ-DP is equivalent to $\sup_{u\sim v} \mathsf{D}_\infty(\mu_u \| \mu_v) \le \epsilon$, where D_∞ is the ∞-order Renyi divergence, it is clear that KL DP is a relaxation of DP (see e.g. [40, Lemma 3.18]). By the data processing inequality, we further have that smoothed KL DP is a relaxation of KL DP.

In practice, given output samples from a privacy mechanism, one encounters the problem of ascertaining whether the mechanism is differentially private or not, referred to as auditing DP. In [18], a hypothesis test for auditing DP using a regularized kernel Rényi divergence was proposed, where the null hypothesis is that the mechanism satisfies (ϵ, δ)-DP. The authors propose a decision rule achieving any non-zero significance level (type I error probability), leaving the characterization of the power (equivalently, type II error probability) open. Here, utilizing Proposition 1, we put forth a principled hypothesis testing pipeline for auditing DP using the Gaussian-smoothed and the classical KL divergence. Our analysis accounts for both significance and power of the test. We start from the smoothed KL DP test.

A. Smoothed KL DP Test

The main objective of a privacy audit is to identify violations. For that reason, we set up an hypothesis test where the null H_0 corresponds to when privacy holds, and consider a sequence of local alternatives $H_{1,n}$ that become harder to distinguish from H_0 as n grows. This models a situation where the alternative hypothesis is arbitrarily close to the null, and we seek a powerful test that successfully rejects the null, even under these local alternatives. To define the local alternatives, we consider a sequence of privacy mechanisms that violate ϵ-smoothed KL DP by an $O(n^{-1/2})$ amount.

Fix $\sigma, \epsilon, b, C > 0$ and, for $n \in \mathbb{N}_0$, let $f_n : \mathfrak{U} \to \mathcal{I}_b := [-b, b]^d$ be a sequence of privacy mechanisms. Denote a pair of adjacent databases by $(U, V) \sim \tilde{\pi} \in \mathcal{P}(\mathfrak{U} \times \mathfrak{U})$. Let $\pi_n := (f_n, f_n)_\# \tilde{\pi}$ be the joint distribution of $\big(f_n(U), f_n(V)\big)$, where $\#$ is the pushforward operation. The first and second marginals of π_n are denoted by μ_n and ν_n, respectively. We impose the following assumption on the sequence $(\pi_n)_{n\in\mathbb{N}_0}$.

Assumption 1 *The sequence $(\pi_n)_{n\in\mathbb{N}_0}$ is such that*

(i) there exists $0 \ne h \in L^2(\pi_0)$ with $n\mathsf{H}^2(\pi_n, \pi_0) \to \|h/2\|^2_{2,\pi_0}$, $\int_{\mathbb{R}^d\times\mathbb{R}^d} h\, d\pi_0 = 0$, and

$$\left(n^{1/2}\left(\frac{\mu_n * \varphi_\sigma - \mu_0 * \varphi_\sigma}{\nu_0 * \varphi_\sigma}\right), n^{1/2}\left(\frac{\nu_n * \varphi_\sigma}{\nu_0 * \varphi_\sigma} - 1\right)\right) \to \left(\frac{\mathbb{E}_{\pi_0}[h(X,Y)\varphi_\sigma(\cdot - X)]}{\nu_0 * \varphi_\sigma}, \frac{\mathbb{E}_{\pi_0}[h(X,Y)\varphi_\sigma(\cdot - Y)]}{\nu_0 * \varphi_\sigma}\right)$$

in $L^\infty(\lambda) \times L^\infty(\lambda)$.

(ii) $\mathsf{D}_{\mathsf{KL}}(\mu_0 * \gamma_\sigma \| \nu_0 * \gamma_\sigma) \le \epsilon$ and $\mathsf{D}_{\mathsf{KL}}(\mu_n * \gamma_\sigma \| \nu_n * \gamma_\sigma) \ge \epsilon_{n,C} := \epsilon + Cn^{-1/2}$ for all n sufficiently large.

Observe that Assumption $1(ii)$ implies that f_0 satisfies ϵ-smoothed KL DP while f_n violates it for all n sufficiently large. On the other hand, Assumption $1(i)$ is a technical requirement that guarantees that the Gaussian-smoothed KL divergence limit theorems needed for the analysis continue to hold under the local alternatives setting, where π_n changes with n. Proposition 3 below presents an explicit construction of $(\pi_n)_{n\in\mathbb{N}_0}$ that satisfies Assumption 1 for any $\sigma, \epsilon, b, C > 0$.

For now, under this assumption, consider the following binary hypothesis test with a sequence of alternatives:

$$\begin{aligned} H_0 &: \mathsf{D}_{\mathsf{KL}}(\mu_0 * \gamma_\sigma \| \nu_0 * \gamma_\sigma) \le \epsilon, \\ H_{1,n} &: \mathsf{D}_{\mathsf{KL}}(\mu_n * \gamma_\sigma \| \nu_n * \gamma_\sigma) \ge \epsilon_{n,C}. \end{aligned} \tag{11}$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n) \sim \pi$ be pairwise i.i.d. samples of the privacy mechanism's output when acting on i.i.d. pairs of adjacent databases, where $\pi = \pi_0$ under H_0 and $\pi = \pi_n$ under $H_{1,n}$. Denote the empirical measures of $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ by $\hat{\mu}_n$ and $\hat{\nu}_n$, respectively. For a test statistic $T_n = T_n(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$, a standard class of tests rejects H_0 if $T_n > t_n$, where t_n is a critical value chosen according to the desired level $\tau \in (0, 1)$. The operational meaning of rejecting H_0 is declaring that ϵ-smoothed KL DP is violated, and hence, also ϵ-DP itself. We say that such a sequence has *asymptotic level* τ if $\limsup_{n\to\infty} \mathbb{P}(T_n > t_n | H_0) \le \tau$. The *power* of a test is the probability that it correctly rejects H_0, i.e., $\mathbb{P}(T_n > t_n | H_{1,n})$, and the *asymptotic power* is $\liminf_{n\to\infty} \mathbb{P}(T_n > t_n | H_{1,n})$. Lastly, the sequence of tests is called *asymptotically consistent* if its asymptotic

power is 1. These definitions specialize to the case of a fixed alternative H_1 by $H_{1,n} = H_1$ and $\pi_n = \pi_1$, for all $n \in \mathbb{N}$.

For $\tau \in [0,1]$, let $Q^{-1}(\tau) = \inf\{z \in \mathbb{R} : (2\pi)^{-1/2}\int_z^\infty e^{-u^2/2}du \leq \tau\}$. The following proposition provides a test statistic for the above hypothesis test and characterizes its asymptotic level and asymptotic power against local alternatives.

Proposition 2 (Smoothed KL DP audit) *Suppose Assumption 1 holds. For $0 < \tau, \tau' \leq 1$, there exists a constant $c_{b,d,\sigma}$ (see [37, Equation (108)]) such that the test statistic $T_n = \mathsf{D}_{\mathsf{KL}}(\hat{\mu}_n * \gamma_\sigma \| \hat{\nu}_n * \gamma_\sigma)$ with critical value $t_n = \epsilon + c_{b,d,\sigma} Q^{-1}(\tau) n^{-1/2}$ achieves an asymptotic level τ and asymptotic power at least $1-\tau'$ for the test in* (11)*, whenever $C > C_{b,d,\sigma,\tau,\tau'} \vee 0$ and $C_{b,d,\sigma,\tau,\tau'} = c_{b,d,\sigma}\big(Q^{-1}(\tau) - Q^{-1}(1-\tau')\big)$.*

The proof of the above claim relies on the limit distribution result for smoothed KL divergence given in (10) along with its refinement to account for the local alternatives scenario, i.e., treating the sequence of distribution pairs $(\mu_n * \gamma_\sigma, \nu_n * \gamma_\sigma)_{n\in\mathbb{N}}$, instead of a fixed one. This refinement is derived under Assumption 1 by invoking Le Cam's third lemma [15, Theorem 3.10.7]. Given these results and the fact that the relevant limit distributions are Gaussian, Proposition 2 follows by an analysis of the asymptotic level and power via the Portmanteau theorem [15, Theorem 1.3.4]. Note that the constant $C_{b,d,\sigma,\tau,\tau'}$ is positive whenever $\tau + \tau' < 1$, which is when the requirement $C > C_{b,d,\sigma,\tau,\tau'}$ is active. Operationally, $\tau + \tau' < 1$ means that the sum of type I and type II error probabilities is less than 1, which is the interesting regime for hypothesis testing; otherwise, a test based on a random coin flip is preferable.

We next provide an explicit construction of a sequence of couplings $(\pi_n)_{n\in\mathbb{N}_0}$ satisfying Assumption 1.

Proposition 3 (Construction for Assumption 1) *We have:*

(i) Let $\pi_0 \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ be such that $\mu_0 \neq \nu_0$, $\nu_0 \otimes \mu_0 \ll \pi_0$, $\|d(\nu_0 \otimes \mu_0)/d\pi_0\|_{\infty,\pi_0} < \infty$ and $\|h_{\pi_0,\bar{c}}\|_{2,\pi_0} < \infty$, where $h_{\pi_0,\bar{c}} := \bar{c}\big(d(\mu_0 \otimes \nu_0)/d\pi_0) - (d(\nu_0 \otimes \mu_0)/d\pi_0)\big)$ and $\bar{c} > 0$ is an arbitrary constant. Let $\pi_n \ll \pi_0$ be the probability measure specified by the relative density $d\pi_n/d\pi_0 = 1 + n^{-\frac{1}{2}} h_{\pi_0,\bar{c}}$, whenever the right-hand side (RHS) is non-negative π_0-a.s.; otherwise, set $\pi_n = \pi_0$. Then, π_n satisfies Assumption 1(i) with $h = h_{\pi_0,\bar{c}}$.

(ii) Let $\pi_0 \in \mathcal{P}(\mathcal{I}_b \times \mathcal{I}_b)$ and σ be such that $\mu_0 \neq \nu_0$, $\nu_0 \otimes \mu_0 \ll \pi_0$, $\|d(\nu_0 \otimes \mu_0)/d\pi_0\|_{\infty,\pi_0} \vee \|d(\mu_0 \otimes \nu_0)/d\pi_0\|_{2,\pi_0} < \infty$ and $\mathsf{D}_{\mathsf{KL}}(\mu_0 * \gamma_\sigma \| \nu_0 * \gamma_\sigma) = \epsilon$. Then, there exists a sufficiently large $\bar{c}$, such that π_n as defined in Part (i) satisfies Assumption 1 with $h = h_{\pi_0,\bar{c}}$ for any $C > 0$.

Proposition $3(ii)$ provides a method of constructing π_n for the hypothesis test in (11), given π_0 that satisfies the aforementioned regularity assumptions. This can be achieved, for instance, by choosing π_0 with $\mu_0 \lll \nu_0 \ll \lambda$, $\|d\nu_0/d\mu_0\|_\infty \vee \|d\mu_0/d\nu_0\|_\infty < \infty$, and $\mathsf{D}_{\mathsf{KL}}(\mu_0 * \gamma_\sigma \| \nu_0 * \gamma_\sigma) = \epsilon$.

B. KL DP test

A more stringent DP audit is realized by a hypothesis test for detecting ϵ-KL DP violations, instead of its smoothed version. We provide such a test against a fixed alternative:

$$H_0 : \mathsf{D}_{\mathsf{KL}}(\mu_0 \| \nu_0) \leq \epsilon, \quad H_1 : \mathsf{D}_{\mathsf{KL}}(\mu_1 \| \nu_1) \geq \tilde{\epsilon}, \tag{12}$$

where $\tilde{\epsilon} > \epsilon > 0$. We again employ the test statistic T_n from Proposition 2 with appropriately chosen σ. Doing so requires additional assumption on the output distributions of the DP mechanism, namely, that μ_i, ν_i, $i = 0, 1$, have smooth Lebesgue densities belonging to the Lipschitz class $\mathrm{Lip}_{s,1}(M, \mathcal{X})$, with smoothness parameter s, norm parameter M, and domain $\mathcal{X} \subseteq \mathbb{R}^d$; see [41].

Assumption 2 *For $i = 0, 1$, the Lebesgue densities $p_{\mu_i}, p_{\nu_i} \in \mathrm{Lip}_{s,1}(M, \mathcal{I}_b)$ and $\|p_{\mu_i}/p_{\nu_i}\|_\infty \vee \|p_{\nu_i}/p_{\mu_i}\|_\infty \leq M$ for some $0 < s \leq 1$ and $M > 0$. Further, $\mathsf{D}_{\mathsf{KL}}(\mu_0 \| \nu_0) \leq \epsilon$ and $\mathsf{D}_{\mathsf{KL}}(\mu_1 \| \nu_1) \geq \tilde{\epsilon}$ for some $\tilde{\epsilon} > \epsilon > 0$.*

Assumption 2 is not restrictive in practice. Indeed, the definition of DP itself necessitates that $\|p_{\mu_u}/p_{\mu_v}\|_\infty$ is bounded uniformly over all $u, v \in \mathfrak{U}$ with $u \sim v$. Moreover, the Lipschitz class grows as we shrink the smoothness parameter s, whereby $\cup_{M\geq 0}\mathrm{Lip}_{1,1}(M, \mathcal{I}_b) \subseteq \cup_{M\geq 0}\mathrm{Lip}_{s,1}(M, \mathcal{I}_b)$ (since we assume $s \in (0, 1]$). As functions with bounded variation (for $d = 1$) over $\mathcal{I}_b$ are elements of $\cup_{M\geq 0}\mathrm{Lip}_{1,1}(M, \mathcal{I}_b)$, Assumption 2 allows for most densities of practical interest.

We are now ready to state the ϵ-KL DP audit result. As it may be unrealistic to assume that the exact values of M, s, and $\tilde{\epsilon}$ are known when constructing T_n and choosing critical values, the following proposition only requires the existence of known constants $\bar{M}$,$\bar{\epsilon}$, $\underline{s}$, and $\bar{s}$ such that $M \leq \bar{M} < \infty$, $\epsilon < \bar{\epsilon} \leq \tilde{\epsilon}$, and $0 < \underline{s} \leq s \leq \bar{s} \leq 1$.

Proposition 4 (KL DP audit) *Suppose Assumption 2 holds. There exists constants $c_{b,d,\sigma}$ and $\sigma_{\epsilon,\bar{\epsilon},\underline{s},\bar{s},d,\bar{M}}$ such that for all $0 < \tau \leq 1$ and $0 < \sigma < \sigma_{\epsilon,\bar{\epsilon},\underline{s},\bar{s},d,\bar{M}}$, the test statistic $T_n = \mathsf{D}_{\mathsf{KL}}(\hat{\mu}_n * \gamma_\sigma \| \hat{\nu}_n * \gamma_\sigma)$ with critical value $t_n = \epsilon + c_{b,d,\sigma} Q^{-1}(\tau) n^{-1/2}$ is asymptotically consistent and achieves asymptotic level τ for the test in* (12)*.*

Explicit expressions for $c_{b,d,\sigma}$ and $\sigma_{\epsilon,\bar{\epsilon},\underline{s},\bar{s},d,\bar{M}}$ are given in [37, Equation (108)] and [37, Equation (110)], respectively. The key difference between the proof of this claim and that of Proposition 2 is that given $\bar{M}$, $\bar{\epsilon}$, $\underline{s}$, and $\bar{s}$, one may choose $\sigma > 0$ small enough so that $\mathsf{D}_{\mathsf{KL}}(\mu_1 * \gamma_\sigma \| \nu_1 * \gamma_\sigma) > \epsilon$ while $\mathsf{D}_{\mathsf{KL}}(\mu_0 * \gamma_\sigma \| \nu_0 * \gamma_\sigma) \leq \epsilon$. Choosing such a σ, the claim then follows by utilizing (10) along with the Portmanteau theorem to bound the type I and type II error probabilities associated with T_n. The aforementioned choice of σ relies on a stability lemma for smoothed KL divergence given next, which may be of independent interest.

Lemma 1 (Stability of smoothed KL divergence) *Let $\mathcal{X} \subseteq \mathbb{R}^d$, and $\mu, \nu \in \mathcal{P}(\mathcal{X})$ have Lebesgue densities p_μ and p_ν, respectively. Further, assume that $p_\mu, p_\nu \in \mathrm{Lip}_{s,1}(M, \mathcal{X})$ and $\|p_\mu/p_\nu\|_\infty \vee \|p_\nu/p_\mu\|_\infty \leq M$ for some $M \geq 1$. Then,*

$$\big|\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) - \mathsf{D}_{\mathsf{KL}}(\mu * \gamma_\sigma \| \nu * \gamma_\sigma)\big| \leq cM(1 + M + \log M)\sigma^s,$$

where $c = c_{d,s} := \int_{\mathbb{R}^d} \|z\|^s \varphi_1(z) dz$ only depends on s and d.

REFERENCES

[1] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, Jan. 1967.

[2] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[3] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. Berkely: University of California Press, 1961, pp. 547–561.

[4] T. van Erven and P. Harremoës, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

[5] V. M. Zolotarev, "Probability metrics," *Teoriya Veroyatnostei i ee Primeneniya*, vol. 28, no. 2, pp. 264–287, 1983.

[6] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, vol. 29, no. 2, pp. 429–443, Jun. 1997.

[7] C. Villani, *Optimal Transport: Old and New*. Springer Berlin, Heidelberg, 2008.

[8] F. Santambrogio, *Optimal Transport for Applied Mathematicians*. Birkhäuser Cham, 2015.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. NewYork: Wiley, 1991.

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

[11] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 29, Barcelona, Spain, Dec. 2016, pp. 271–279.

[12] K.-S. Song, "Goodness-of-fit tests based on Kullback-Leibler discrimination information," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1103–1117, 2002.

[13] S. Claici, M. Yurochkin, S. Ghosh, and J. Solomon, "Model fusion with Kullback-Leibler divergence," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 13–18 Jul 2020, pp. 2038–2047.

[14] M. Afgani, S. Sinanovic, and H. Haas, "Anomaly detection using the Kullback-Leibler divergence metric," in *Proceedings of the 2008 First International Symposium on Applied Sciences on Biomedical and Communication Technologies*, Oct. 2008, pp. 1–5.

[15] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

[16] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, Oct. 2018, pp. 475–489.

[17] M. Jagielski, J. Ullman, and A. Oprea, "Auditing differentially private machine learning: How private is private SGD?" in *Proceedings of Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 22 205–22 216.

[18] C. Domingo-Enrich and Y. Mroueh, "Auditing differential privacy in high dimensions with the kernel quantum Rényi divergence," *arXiv:2205.13941*, 2022.

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

[20] F. Perez-Cruz, "Kullback-Leibler divergence estimation of continuous distributions," in *Proceedings of the 2008 IEEE International Symposium on Information Theory*, Toronto, ON, Canada, Jul. 2008, pp. 1666–1670.

[21] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and J. M. Robins, "Nonparametric von Mises estimators for entropies, divergences and mutual informations," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 28, Montréal, Canada, Dec. 2015, pp. 397–405.

[22] S. Singh and B. Póczos, "Finite-sample analysis of fixed-k nearest neighbor density functional estimators," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 29, Barcelona, Spain, Dec. 2016, pp. 1225–1233.

[23] M. Noshad, K. R. Moon, S. Y. Sekeh, and A. O. Hero, "Direct estimation of information divergence using nearest neighbor ratios," in *Proceedings of the 2017 IEEE International Symposium on Information Theory*, Jun. 2017, pp. 903–907.

[24] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, Stockholm Sweden, Jul. 2018, pp. 531–540.

[25] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Ensemble estimation of information divergence," *Entropy*, vol. 20, no. 8, Aug. 2018.

[26] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, Oct. 2010.

[27] T. B. Berrett, R. J. Samworth, and M. Yuan, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *The Annals of Statistics*, vol. 47, no. 1, pp. 288–318, Feb. 2019.

[28] T. B. Berrett and R. J. Samworth, "Efficient two-sample functional estimation and the super-oracle phenomenon," *arXiv:1904.09347*, 2019.

[29] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over Lipschitz balls," *The Annals of Statistics*, vol. 48, no. 6, pp. 3228–3250, Dec. 2020.

[30] S. Sreekumar, Z. Zhang, and Z. Goldfeld, "Non-asymptotic performance guarantees for neural estimation of f-divergences," in *AISTATS*, 2021.

[31] S. Sreekumar and Z. Goldfeld, "Neural estimation of statistical divergences," *Journal of Machine Learning Research*, vol. 23, no. 126, pp. 1–75, 2022.

[32] M. Salicru, D. Morales, M. L. Menendez, and L. Pardo, "On the applications of divergence type measures in testing statistical hypotheses," *Journal of Multivariate Analysis*, vol. 51, no. 2, pp. 372–391, 1994.

[33] K. Moon and A. Hero, "Multivariate f-divergence estimation with confidence," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 27, Montréal, Canada, 2014, pp. 2420–2428.

[34] Z. Goldfeld and K. Kato, "Limit distributions for smooth total variation and χ^2-divergence in high dimensions," in *Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2640–2645.

[35] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[36] I. Kontoyiannis and M. Skoularidou, "Estimating the directed information and testing for causality," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6053–6067, 2016.

[37] S. Sreekumar, Z. Goldfeld, and K. Kato, "Limit distribution theory for f-divergences," *arXiv preprint arxiv: 2211.11184*, 2022.

[38] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy, "Convergence of smoothed empirical measures with applications to entropy estimation," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4368–4391, Jul. 2020.

[39] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Third Theory of Cryptography Conference*. Springer Berlin Heidelberg, 2006, pp. 265–284.

[40] C. Dwork and A. Roth, *The Algorithmic Foundations of Differential Privacy*. Hanover, MA, USA: Now Publishers Inc., Aug. 2014, vol. 9.

[41] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. Springer Berlin, Heidelberg, 1993.