Convergence of Smoothed Empirical Measures with Applications to Entropy Estimation

Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed and Yury Polyanskiy

***Abstract*—This paper studies convergence of empirical measures smoothed by a Gaussian kernel. Specifically, consider approximating $P * \mathcal{N}_\sigma$, for $\mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$, by $\hat{P}_n * \mathcal{N}_\sigma$ under different statistical distances, where $\hat{P}_n$ is the empirical measure. We examine the convergence in terms of the Wasserstein distance, total variation (TV), Kullback-Leibler (KL) divergence, and χ^2-divergence. We show that the approximation error under the TV distance and 1-Wasserstein distance (W_1) converges at the rate $e^{O(d)}n^{-1/2}$ in remarkable contrast to a (typical) $n^{-\frac{1}{d}}$ rate for unsmoothed W_1 (and $d \geq 3$). Similarly, for the KL divergence, squared 2-Wasserstein distance (W_2^2), and χ^2-divergence, the convergence rate is $e^{O(d)}n^{-1}$, but only if P achieves finite input-output χ^2 mutual information across the additive white Gaussian noise (AWGN) channel. If the latter condition is not met, the rate changes to $\omega\left(n^{-1}\right)$ for the KL divergence and W_2^2, while the χ^2-divergence becomes infinite – a curious dichotomy.**

As an application we consider estimating the differential entropy $h(S+Z)$, where $S \sim P$ and $Z \sim \mathcal{N}_\sigma$ are independent d-dimensional random variables. The distribution P is unknown and belongs to some nonparametric class, but n independently and identically distributed (i.i.d) samples from it are available. Despite the regularizing effect of noise, we first show that any good estimator (within an additive gap) for this problem must have a sample complexity that is exponential in d. We then leverage the above empirical approximation results to show that the absolute-error risk of the plug-in estimator converges as $e^{O(d)}n^{-1/2}$, thus attaining the parametric rate in n. This establishes the plug-in estimator as minimax rate-optimal for the considered problem, with sharp dependence of the convergence rate both in n and d. We provide numerical results comparing the performance of the plug-in estimator to that of general-purpose (unstructured) differential entropy estimators (based on kernel density estimation (KDE) or k nearest neighbors (kNN) techniques) applied to samples of $S+Z$. These results reveal a significant empirical superiority of the plug-in to state-of-the-art KDE and kNN methods. As a motivating utilization of the plug-in approach, we estimate information flows in deep neural networks and discuss Tishby's Information Bottleneck and the compression conjecture, among others.

***Index Terms*—Deep neural networks, differential entropy, estimation, empirical approximation, Gaussian kernel, minimax rates, mutual information.**

I. Introduction

This work is motivated by a new nonparametric and high-dimensional functional estimation problem, which we refer to as 'differential entropy estimation under Gaussian convolutions.' The goal of this problem is to estimate the differential entropy $h(S+Z)$ based on samples of S, while knowing the distribution of $Z \sim \mathcal{N}_\sigma \triangleq \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ which is independent of S. The analysis of the estimation risk reduces to evaluating the expected 1-Wasserstein distance or χ^2-divergence between $P * \mathcal{N}_\sigma$ and $\hat{P}_{S^n} * \mathcal{N}_\sigma$, where $S^n \triangleq (S_1, \ldots, S_n)$ are i.i.d. samples from P and $\hat{P}_{S^n} = \frac{1}{n}\sum_{i=1}^n \delta_{S_i}$ is the empirical measure.[1] Due to the popularity of the additive Gaussian noise model, we start by exploring this smoothed empirical approximation problem in detail, under several additional statistical distances.

A. Convergence of Smooth Empirical Measures

Consider the empirical approximation error $\mathbb{E}\delta(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)$ under some statistical distance δ. We consider various choices of δ, including the 1-Wasserstein and (squared) 2-Wasserstein distances, total variation (TV), Kullback-Leibler (KL) divergence, and χ^2-divergence. We show that, when P is subgaussian, the approximation error under the 1-Wasserstein and TV distances drops at the *parametric* rate for *any* dimension d. The exact rate is $c^d n^{-1/2}$, for a constant c. The parametric convergence rate is also attained by the squared 2-Wasserstein distance, KL divergence, and χ^2-divergence, so long as P achieves finite input-output χ^2 mutual information across the additive white Gaussian noise (AWGN) channel. We show that this condition is always met for subgaussian P in the low signal-to-noise ratio (SNR) regime. This fast convergence is remarkable since classical (unconvolved) empirical approximation suffers from the so-called curse of dimensionality. For instance, the empirical 1-Wasserstein distance $\mathbb{E}\mathsf{W}_1(\hat{P}_{S^n}, P)$ is known to decay at most as $n^{-\frac{1}{d}}$ [1], which is sharp for all $d > 2$ (see [2] and [3] for sharper results). Convolving P and $\hat{P}_{S^n}$ with $\mathcal{N}_\sigma$ improves the rate from $n^{-\frac{1}{d}}$ to $c^d n^{-1/2}$ (or $c^d n^{-1}$ for squared distances). The latter has a milder

This work was partially supported by the MIT-IBM Watson AI Lab. The work of Z. Goldfeld and Y. Polyanskiy was also supported in part by the National Science Foundation CAREER award under grant agreement CCF-12-53205, by the Center for Science of Information (CSoI), an NSF Science and Technology Center under grant agreement CCF-09-39370, and a grant from Skoltech–MIT Joint Next Generation Program (NGP). The work of Z. Goldfeld was also supported by the Rothschild postdoc fellowship. The work of J. Niles-Weed was partially supported by the Josephine de Kármán Fellowship.

This work was presented in part at the 2018 IEEE International Conference on the Science of Electrical Engineering (ICSEE-2018), Eilat, Israel, and in part at the 2019 IEEE International Symposium on Information Theory (ISIT-2019), Paris, France.

Z. Goldfeld is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850 US (e-mail: goldfeld@cornell.edu). K. Greenewald is with the MIT-IBM Watson AI Lab, Cambridge, MA 02142 US (email: kristjan.h.greenewald@ibm.com). J. Niles-Weed is with the Courant Institute of Mathematical Sciences and Center for Data Science, New York University, New York, NY 10003 US (email: jnw@cims.nyu.edu). Y. Polyanskiy is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 US (e-mail: yp@mit.edu).

[1] Here, δ_{S_i} stands for the Dirac measure at S_i.

dependence on d and can be dominated with practical choices of n, even in relatively high dimensions.

The χ^2-divergence $\mathbb{E}\chi^2\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right)$ presents a particularly curious behavior: it converges as[2] $\frac{1}{n}$ for low SNR and possibly diverges when SNR is high. To demonstrate a diverging scenario we construct a subgaussian P for which $\mathbb{E}\chi^2\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) = \infty$ whenever the subgaussian constant is greater than or equal to $\sqrt{2}\sigma$. We also show that, when the χ^2-divergence is infinite, the convergence rate of the squared 2-Wasserstein distance and KL divergence are strictly slower than the parametric rate. All of these empirical approximation results are gathered in Section II.

B. Differential Entropy Estimation Under Smoothing

We then apply these empirical approximation results to study the estimation of $h(S+Z)$, where $S \sim P$ is an arbitrary (continuous, discrete, or mixed) $\mathbb{R}^d$-valued random variable and $Z \sim \mathcal{N}_\sigma$ is an isotropic Gaussian. The differential entropy is estimated using n i.i.d. samples S^n from P and assuming σ is known. To investigate the decision-theoretic fundamental limit, we consider the minimax absolute-error risk

$$\mathcal{R}^\star(n,\sigma,\mathcal{F}_d) \triangleq \inf_{\hat{h}} \sup_{P\in\mathcal{F}_d} \mathbb{E}\left| h(P*\mathcal{N}_\sigma) - \hat{h}(S^n,\sigma)\right|, \quad (1)$$

where $\mathcal{F}_d$ is a nonparametric class of distributions and $\hat{h}$ is the estimator. The sample complexity $n^\star(\eta,\sigma,\mathcal{F}_d)$ is the smallest number of samples needed for estimating $h(P*\mathcal{N}_\sigma)$ within an additive gap η. We aim to understand whether having access to 'clean' samples of S can improve estimation performance (theoretically and empirically) compared to when only 'noisy' samples of $S+Z$ are available and the distribution of Z is unknown.

Our results establish the plug-in estimator as minimax rate-optimal for the considered problem. Defining $\mathsf{T}_\sigma(P) \triangleq h(P*\mathcal{N}_\sigma)$ as the functional of interest, the plug-in estimator is $\mathsf{T}_\sigma\big(\hat{P}_{S^n}\big) = h\big(\hat{P}_{S^n} * \mathcal{N}_\sigma\big)$. Plug-in techniques are suboptimal for vanilla discrete (Shannon) and differential entropy estimation (see [4] and [5], respectively). Nonetheless, we show that $h\big(\hat{P}_{S^n} * \mathcal{N}_\sigma\big)$ attains the parametric estimation rate of $O_{\sigma,d}(n^{-1/2})$ when P is subgaussian, establishing the optimality of the plug-in.

We use the χ^2 empirical approximation result to prove the parametric risk convergence rate when P has bounded support. The result is then extended to (unbounded) subgaussian P via a separate argument. Specifically, we first bound the risk by a *weighted* TV distance between $P*\mathcal{N}_\sigma$ and $\hat{P}_{S^n}*\mathcal{N}_\sigma$. This bound is derived by linking the two measures via the maximal TV-coupling. The subgaussianity of P and the smoothing introduced by the Gaussian convolution are used to bound the weighted TV distance by a $e^{O(d)}n^{-1/2}$ term, with all constants explicitly characterized. Notably, while the convergence with n is parametric, the derived rates still depends exponentially on d though the prefactor.

A natural next question is whether the exponential dependence on d is necessary. Answering in the affirmative, we prove that any good estimator of $h(P*\mathcal{N}_\sigma)$, within an additive gap η, has a sample complexity $n^\star(\eta,\sigma,\mathcal{F}_d) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma)$ is positive and monotonically decreasing in σ. The proof relates the estimation of $h(P*\mathcal{N}_\sigma)$ to estimating the discrete entropy of a distribution supported on a capacity-achieving codebook for an AWGN channel. Existing literature (e.g., [6], [7]) implies that the discrete problem has sample complexity exponential in d (because this the growth rate of the codebook's size), which is then carried over to the original problem to establish the result.

Finally, we focus on the practical estimation of $h(P*\mathcal{N}_\sigma)$. While the above results give necessary and sufficient conditions on the number of samples needed to drive the estimation error below a desired threshold, these are worst-case bounds. In practice, the unknown distribution P may not follow the minimax rates, and the resulting estimation error could be smaller. As a guideline for setting n in practice, we derive a lower bound on the bias of the plug-in estimator that scales as $\log\left(2^d n^{-1}\right)$. Our last step is to propose an efficient implementation of the plug-in estimator based on Monte Carlo (MC) integration. As the estimator amounts to computing the differential entropy of a *known* Gaussian mixture, MC integration allows a simple and efficient computation. We bound the mean squared error (MSE) of the computed value by $c^{(\mathsf{MC})}_{\sigma,d}(n\cdot n_{\mathsf{MC}})^{-1}$, where n is the number of centers in the mixture[3], n_{MC} is the number of MC samples, and $c^{(\mathsf{MC})}_{\sigma,d} = \Theta(d)$ is explicitly characterized. The proof uses the Gaussian Poincaré inequality to reduce the analysis to that of the log-mixture distribution gradient. Several simulations (including an estimation experiment over a small deep neural network (DNN) for a 3-dimensional spiral classification task) illustrate the superiority of the ad hoc plug-in approach over existing general-purpose estimators, both in the rate of error decay and scalability with dimension.

C. Related Differential Entropy Estimation Results

General-purpose differential entropy estimators can be used in the considered setup by estimating $h(S+Z)$ using 'noisy' samples of $S+Z$ (generated from the available samples of S). There are two prevailing approaches for estimating the nonsmooth differential entropy functional: (i) based on kernel density estimators (KDEs) [8]–[10]; and (ii) using k nearest neighbor (kNN) techniques [11]–[19] (see also [20], [21] for surveys). Many performance analyses of such estimators restrict attention to nonparametric classes of smooth and compactly supported densities that are bounded away from zero (although the support may be unknown [9], [10]). Since the density associated with $P*\mathcal{N}_\sigma$ violates these assumptions, such results do not apply in our setup. The work of Tsybakov and van der Meulen [13] accounted for densities with unbounded support and exponentially decaying tails for $d=1$, but we are interested in the high-dimensional scenario.

Two recent works weakened or dropped the boundedness from below assumption in the high-dimensional setting, providing general-purpose estimators whose risk bounds are

[2] Recall χ^2 is a squared distance.

[3] The number of centers is the number of samples used for estimation.

valid in our setup. In [5], a KDE-based differential entropy estimator that also combines best polynomial approximation techniques was proposed. Assuming subgaussian densities with unbounded support, Theorem 2 of [5] bounded the estimation risk by[4] $O\left(n^{-\frac{s}{s+d}}\right)$, where s is a Lipschitz smoothness parameter assumed to satisfy $0 < s \leq 2$. While the result is applicable for our setup when P is compactly supported or subgaussian, the convergence rate for large d is roughly $n^{-\frac{1}{d}}$. This rate deteriorates quickly with dimension and is unable to exploit the smoothness of $P * \mathcal{N}_\sigma$ due to the $s \leq 2$ restriction.[5] This is to be expected because the results of [5] account for a wide class of density functions, including highly non-smooth ones.

In [19], a weighted-KL (wKL) estimator (in the spirit of [15]) was studied for smooth densities. In a major breakthrough, the authors proved that with a careful choice of weights the estimator is asymptotically efficient,[6] under certain assumptions on the densities' speed of decay to zero (which captures $P*\mathcal{N}_\sigma$ when, e.g., P is compactly supported). Despite its $O(n^{-1/2})$ risk convergence rate, however, the empirical performance of the estimator seems lacking (at least in our experiments, which use the code provided in [19]). As shown in Section V, the plug-in estimator achieves superior performance even in rather simple scenarios of moderate dimension. The empirical performance of the estimator from [19] may originate from the dependence of its estimation risk on the dimension d, which was not characterized therein.

D. Relation to Information Flows in Deep Neural Networks

The considered differential entropy estimation problem is closely related to that of estimating information flows in DNNs. There has been a recent surge of interest in measuring the mutual information between selected groups of neurons in a DNN [23]–[28], partially driven by the Information Bottleneck (IB) theory [29], [30]. Much of the focus centers on the mutual information $I(X;T)$ between the input feature X and a hidden activity vector T. However, as explained in [28], this quantity is vacuous in deterministic DNNs[7] and becomes meaningful only when a mechanism for discarding information (e.g., noise) is integrated into the system. Such a noisy DNN framework was proposed in [28], where each neuron adds a small amount of Gaussian noise (i.i.d. across neurons) after applying the activation function. While the injection of noise alleviates the degeneracy of $I(X;T)$, the concatenation of Gaussian noises and nonlinearities makes this mutual information impossible to compute analytically or even evaluate numerically. Specifically, the distribution of T (marginal or conditioned on X) is highly convoluted and thus the appropriate mode of operation becomes treating it as unknown, belonging to some nonparametric class.

Herein, we lay the groundwork for estimating $I(X;T)$ (or any other mutual information between layers) over real-world DNN classifiers. To achieve this, the estimation of $I(X;T)$ is reduced to the problem of differential entropy estimation under Gaussian convolutions described above. Specifically, in a noisy DNN each hidden layer can be written as $T = S + Z$, where S is a deterministic function of the previous layer and Z is a centered isotropic Gaussian vector. The DNN's generative model enables sampling S, while the distribution of Z is known since the noise is injected by design. Estimating mutual information over noisy DNNs thus boils down to estimating $h(T) = h(S+Z)$ from samples of S, which is a main focus in this work.

Outline: The remainder of this paper is organized as follows. Section II analyzes the convergence of various statistical distances between $P*\mathcal{N}_\sigma$ and its Gaussian-smoothed empirical approximation. In Section III we set up the differential entropy estimation problem and state our main results. Section IV presents applications of the considered estimation problem, focusing on mutual information estimation over DNNs. Simulation results are given in Section V, and proofs are provided in Section VI. The main insights from this work and potential future directions are discussed in Section VII.

Notation: Logarithms are with respect to (w.r.t.) base e. For an integer $k \geq 1$, we set $[k] \triangleq \left\{i \in \mathbb{Z} \middle| 1 \leq i \leq k\right\}$. $\|x\|$ is the Euclidean norm in $\mathbb{R}^d$, and I_d is the $d \times d$ identity matrix. We use $\mathbb{E}_P$ for an expectation w.r.t. a distribution P, omitting the subscript when P is clear from the context. For a continuous $X \sim P$ with PDF p, we interchangeably use $h(X)$, $h(P)$ and $h(p)$ for its differential entropy. The n-fold product extension of P is denoted by $P^{\otimes n}$. The convolution of two distributions P and Q on $\mathbb{R}^d$ is $(P*Q)(\mathcal{A}) = \int\int \mathbb{1}_{\mathcal{A}}(x+y)\,\mathsf{d}P(x)\,\mathsf{d}Q(y)$, where $\mathbb{1}_{\mathcal{A}}$ is the indicator of the Borel set $\mathcal{A}$. We use $\mathcal{N}_\sigma$ for the isotropic Gaussian measure of parameter σ, and denotes its PDF by φ_σ.

II. Smooth Empirical Approximation

This section studies the convergence rate of $\delta(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)$ for different statistical distances $\delta(\cdot,\cdot)$, when P is a K-subgaussian distribution, as defined next, and $\hat{P}_{S^n}$ is the empirical measure associated with $S^n \sim P^{\otimes n}$.

Definition 1 (Subgaussian Distribution) *A d-dimensional distribution P is K-subgaussian, for $K > 0$, if $X \sim P$ satisfies*

$$\mathbb{E}\Big[\exp\big(\alpha^T(X - \mathbb{E}X)\big)\Big] \leq \exp\big(0.5K^2\|\alpha\|^2\big), \quad \forall \alpha \in \mathbb{R}^d. \tag{2}$$

In words, the above requires that every one-dimensional projection of X be subgaussian in the traditional scalar sense. When $\big(X - \mathbb{E}X\big) \in \mathcal{B}(0,R) \triangleq \left\{x \in \mathbb{R}^d \middle| \|x\| \leq R\right\}$, (2) holds with $K = R$.

When $\delta(\cdot,\cdot)$ is the TV or the 1-Wasserstein distance, the distance between the convolved distributions converges at the rate $n^{-1/2}$ for all K and d. However, when $\delta(\cdot,\cdot)$ is the KL divergence or squared 2-Wasserstein distance (both are squared distances), convergence at rate n^{-1} happens only when K

[4] Multiplicative polylogarithmic factors are overlooked in this restatement.

[5] Such convergence rates are typical in estimating $h(p)$ under boundedness or smoothness conditions on p. Indeed, the results cited above (applicable in our framework or otherwise) as well as many others bound the estimation risk as $O\Big(n^{-\frac{\alpha}{\beta+d}}\Big)$, where α, β are constants that may depend on s and d.

[6] in the sense of, e.g., [22, p. 367].

[7] i.e., DNNs that, for fixed parameters, define a deterministic mapping from input to output.

is sufficiently small or P has bounded support. Interestingly, when $\delta(\cdot,\cdot)$ is the χ^2-divergence, two very different behaviors are observed. For low SNR (and in particular when P has bounded support), $\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right)$ converges as n^{-1}. However, when SNR is high, we construct a subgaussian P for which the expected χ^2-divergence is infinite.

Another way to summarize the results of this section is through the following curious dichotomy. This dichotomy highlights the central role of the χ^2-divergence in Gaussian-smoothed empirical approximation. To state it, let $Y = S + Z$, where $S \sim P$ and $Z \sim \mathcal{N}_\sigma$ are independent. Denote the joint distribution of (S, Y) by $P_{S,Y}$, and let $P_S = P$ and $P_Y = P * \mathcal{N}_\sigma$ be their marginals. Setting $I_{\chi^2}(S;Y) \triangleq \chi^2\left(P_{S,Y}\|P_S \otimes P_Y\right)$ as the χ^2 mutual information between S and Y, we have:

1) If P_S is K-subgaussian with $K < \frac{\sigma}{2}$ then $I_{\chi^2}(S;Y) < \infty$. If $K > \sqrt{2}\sigma$ then $I_{\chi^2}(S;Y) = \infty$ for some distributions P_S.
2) Assume $I_{\chi^2}(S;Y) < \infty$. Then $\mathbb{E}_{P^{\otimes n}}\delta(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma) = O\left(n^{-1}\right)$ if δ is the KL or χ^2-divergence. If δ is the TV or the 1-Wasserstein distance, the convergence rate is $O(n^{-1/2})$. If, in addition, P_S is subgaussian (with any constant), then the rate is also $O(n^{-1})$ if δ is the square of the 2-Wasserstein distance.
3) Assume $I_{\chi^2}(S;Y) = \infty$. Then $\mathbb{E}_{P^{\otimes n}}\delta(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma) = \omega(n^{-1})$ if δ is the KL divergence or the squared 2-Wasserstein distance. If δ is the χ^2-divergence, it is infinite.

All the above are stated or immediately implied by the results to follow.

A. 1-Wasserstein Distance

The 1-Wasserstein distance between μ and ν is given by $\mathsf{W}_1(\mu,\nu) \triangleq \inf \mathbb{E}\|X - Y\|$, where the infimum is taken over all couplings of μ and ν, i.e., joint distributions $P_{X,Y}$ whose marginals satisfy $P_X = \mu$ and $P_Y = \nu$.

Proposition 1 (Smooth W_1 Approximation) *Fix $d \geq 1$, $\sigma > 0$ and $K > 0$. For any K-subgaussian distribution P, we have*

$$\mathbb{E}_{P^{\otimes n}}\mathsf{W}_1(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma) \leq c^{(\mathsf{W}_1)}_{\sigma,d,K}\frac{1}{\sqrt{n}}, \tag{3}$$

where $c^{(\mathsf{W}_1)}_{\sigma,d,K}$ is given in (8).

Proof: We can assume without loss of generality that P_S has mean 0. We start with the following upper bound [31, Theorem 6.15]:

$$\mathsf{W}_1(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma) \leq \int_{\mathbb{R}^d} \|z\| \left|r_{S^n}(z) - q(z)\right| \mathsf{d}z, \tag{4}$$

where r_{S^n} and q are the densities associated with $\hat{P}_{S^n} * \mathcal{N}_\sigma$ and $P * \mathcal{N}_\sigma$, respectively. This inequality follows by coupling $\hat{P}_{S^n} * \mathcal{N}_\sigma$ and $P * \mathcal{N}_\sigma$ via the maximal TV-coupling.

Let $f_a : \mathbb{R}^d \to \mathbb{R}$ be the PDF of $\mathcal{N}\left(0, \frac{1}{2a}\mathrm{I}_d\right)$, for $a > 0$ specified later. The Cauchy-Schwarz inequality implies

$$\mathbb{E}_{P^{\otimes n}} \int_{\mathbb{R}^d} \|z\| \left|r_{S^n}(z) - q(z)\right| \mathsf{d}z \tag{5}$$

$$\leq \left(\int_{\mathbb{R}^d} \|z\|^2 f_a(z)\, \mathsf{d}z\right)^{\frac{1}{2}} \left(\mathbb{E}_{P^{\otimes n}} \int_{\mathbb{R}^d} \frac{\left(q(z) - r_{S^n}(z)\right)^2}{f_a(z)} \mathsf{d}z\right)^{\frac{1}{2}}.$$

The first term is the expected squared Euclidean norm of $\mathcal{N}\left(0, \frac{1}{2a}\mathrm{I}_d\right)$, which equals $\frac{d}{2a}$.

For the second integral, note that $r_{S^n}(z)$ is a sum of i.i.d. terms with expectation $q(z)$. This implies

$$\mathbb{E}_{P^{\otimes n}}\left(q(z) - r_{S^n}(z)\right)^2 = \mathsf{var}_{P^{\otimes n}}\left(\frac{1}{n}\sum_{i=1}^n \varphi_\sigma(z - S_i)\right)$$
$$= \frac{1}{n}\mathsf{var}_P\left(\varphi_\sigma(z - S)\right) \leq \frac{c_1^2}{n}\mathbb{E}_P e^{-\frac{1}{\sigma^2}\|z-S\|^2},$$

where $c_1 = (2\pi\sigma^2)^{-d/2}$. Consequently, we have

$$\int_{\mathbb{R}^d} \mathbb{E}_{P^{\otimes n}} \frac{\left(q(z) - r_{S^n}(z)\right)^2}{f_a(z)} \mathsf{d}z \leq \frac{c_1}{n 2^{d/2}}\mathbb{E}\frac{1}{f_a(S + Z/\sqrt{2})}, \tag{6}$$

where $S \sim P$ and $Z \sim \mathcal{N}(0, \sigma^2\mathrm{I}_\mathrm{d})$ are independent.

Setting $c_2 \triangleq \left(\frac{\pi}{a}\right)^{\frac{d}{2}}$, we have $\left(f_a(z)\right)^{-1} = c_2 \exp\left(a\|z\|^2\right)$. Since S is K-subgaussian and Z is σ-subgaussian, $S + Z/\sqrt{2}$ is $(K + \sigma/\sqrt{2})$-subgaussian. Following (6), for any $0 < a < \frac{1}{2(K+\sigma/\sqrt{2})^2}$, we have [32, Remark 2.3]

$$\frac{c_1}{n2^{d/2}}\mathbb{E}\frac{1}{f_a(S + Z/\sqrt{2})} = \frac{c_1 c_2}{n2^{d/2}}\mathbb{E}\exp\left(a\left\|S + Z/\sqrt{2}\right\|^2\right)$$
$$\leq \frac{c_1 c_2}{n2^{d/2}}\exp\left(\left(K + \sigma/\sqrt{2}\right)^2 ad + \frac{(K + \sigma/\sqrt{2})^4 a^2 d}{1 - 2(K + \sigma/\sqrt{2})^2 a}\right), \tag{7}$$

where the last inequality uses the subgaussianity of $S + Z/\sqrt{2}$ and Definition 1. Setting $a = \frac{1}{4(K+\sigma/\sqrt{2})^2}$, we combine (5)-(7) to obtain the result

$$\mathbb{E}_{P^{\otimes n}}\mathsf{W}_1(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)$$
$$\leq \sigma\sqrt{2d}\left(\frac{1}{\sqrt{2}} + \frac{K}{\sigma}\right)^{\frac{d}{2}+1} e^{\frac{3d}{16}}\frac{1}{\sqrt{n}}. \tag{8}$$

■

Remark 1 (Smooth W_1 for Bounded Support) *A better constant is attainable if attention is restricted to the bounded support case. It was shown in [33] that analyzing $\mathbb{E}_{P^{\otimes n}}\mathsf{W}_1(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma)$ directly for* $\mathrm{supp}(P) \subseteq [-1,1]^d$, *one can obtain the bound $\frac{2^{d+2}\sqrt{d}}{\min\{1,\sigma^d\}}n^{-1/2}$.*

B. Total Variation Distance

The TV distance between μ and ν is $\|\mu - \nu\|_{\mathsf{TV}} \triangleq \sup_{A\in\mathcal{F}}|\mu(A) - \nu(A)|$, where $\mathcal{F}$ is the sigma-algebra. When μ and ν have densities, say f and g, respectively, the TV distance reduces to $\frac{1}{2}\int|f(z) - g(z)|\,\mathsf{d}z$.

Proposition 2 (Smooth Total Variation Approximation) *Fix $d \geq 1$, $\sigma > 0$ and $K > 0$. For any K-subgaussian distribution P, we have*

$$\mathbb{E}_{P^{\otimes n}}\left\|\hat{P}_{S^n} * \mathcal{N}_\sigma - P * \mathcal{N}_\sigma\right\|_{\mathsf{TV}} \leq c^{(\mathsf{TV})}_{\sigma,d,K}\frac{1}{\sqrt{n}}, \tag{9}$$

where $c^{(\mathsf{TV})}_{\sigma,d,K}$ is given in (10).

Proof: Noting that $\mathbb{E}_{P^{\otimes n}}\left\|\hat{P}_{S^n}*\mathcal{N}_\sigma-P*\mathcal{N}_\sigma\right\|_{\mathsf{TV}}=\frac{1}{2}\mathbb{E}_{P^{\otimes n}}\int\left|r_{S^n}(z)-q(z)\right|\mathsf{d}z$, we may apply the Cauchy-Schwarz inequality similarly to (5). The only difference now is that the first integral sums up to 1 (rather than being a Gaussian moment). Repeating steps (6)-(7) we obtain

$$\mathbb{E}_{P^{\otimes n}}\left\|\hat{P}_{S^n}*\mathcal{N}_\sigma-P*\mathcal{N}_\sigma\right\|_{\mathsf{TV}}\le\left(\frac{1}{\sqrt{2}}+\frac{K}{\sigma}\right)^{\frac{d}{2}}e^{\frac{3d}{16}}\frac{1}{\sqrt{n}},\tag{10}$$

as desired. ■

C. χ^2-Divergence

The χ^2-divergence $\chi^2(\mu\|\nu)\triangleq\int\left(\frac{\mathsf{d}\mu}{\mathsf{d}\nu}-1\right)^2\mathsf{d}\nu$ presents perhaps the most surprising behavior of all the considered distances. When the signal-to-noise ratio (SNR) $\frac{K}{\sigma}<\frac{1}{2}$, we prove that $\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)$ converges as n^{-1} for all dimensions. However, if $K\ge\sqrt{2}\sigma$, then there exists K-subgaussian distributions P such that $\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)=\infty$ even in $d=1$. Our results rely on the following identity.

Lemma 1 ($\boldsymbol{\chi^2}$-Divergence and Mutual Information) *Let $S\sim P$ and $Y=S+Z$, with $Z\sim\mathcal{N}_\sigma$ independent of S. Then*

$$\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)=\frac{1}{n}I_{\chi^2}(S;Y)\,.$$

Proof: Recall that $r_{S^n}(z)$ is a sum of i.i.d. terms with expectation $q(z)$. This yields

$$\begin{aligned}&\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)\\&\quad=\mathbb{E}_{P^{\otimes n}}\int_{\mathbb{R}^d}\frac{\big(r_{S^n}(z)-q(z)\big)^2}{q(z)}\mathsf{d}z\\&\quad=\frac{1}{n}\left(\int_{\mathbb{R}^d}\mathbb{E}_P\frac{(\varphi_\sigma(z-S)-q(z))^2}{q(z)}\mathsf{d}z\right)\\&\quad=\frac{1}{n}I_{\chi^2}(S;Y)\,.\end{aligned}\tag{11}$$

■

Lemma 1 implies that $\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)=O(n^{-1})$ if and only if $I_{\chi^2}(S;Y)<\infty$. When $I_{\chi^2}(S;Y)=\infty$, $\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)$ diverges for all n. It therefore suffices to examine conditions under which $I_{\chi^2}(S;Y)$ is finite.

1) Convergence at Low SNR and Bounded Support: We start by stating and proving the convergence results.

Proposition 3 (Smooth $\boldsymbol{\chi^2}$ Approximation) *Fix $d\ge1$ and $\sigma>0$. If P is K-subgaussian with $K<\frac{\sigma}{2}$, then $I_{\chi^2}(S;Y)\le c^{(\chi^2)}_{\sigma,d,K}<\infty$, where $c^{(\chi^2)}_{\sigma,d,K}$ is given in* (13)*. Consequently,*

$$\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)\le c^{(\chi^2)}_{\sigma,d,K}\frac{1}{n}\,.$$

Proof: Denote by $\mathcal{N}(x,\sigma^2\mathrm{I}_d)$ an isotropic Gaussian of entrywise variance σ^2 centered at x. Then by the convexity of the χ^2-divergence,

$$\begin{aligned}I_{\chi^2}(S;Y)&=\mathbb{E}_P\chi^2\left(\mathcal{N}(S,\sigma^2\mathrm{I}_d)\middle\|\mathbb{E}_P\mathcal{N}(\tilde{S},\sigma^2\mathrm{I}_d)\right)\\&\le\mathbb{E}_{P^{\otimes2}}\chi^2\left(\mathcal{N}(S,\sigma^2\mathrm{I}_d)\middle\|\mathcal{N}(\tilde{S},\sigma^2\mathrm{I}_d)\right)\\&=\mathbb{E}_{P^{\otimes2}}e^{\frac{1}{\sigma^2}\|S-\tilde{S}\|^2}\,,\end{aligned}\tag{12}$$

where the last equality follows from the closed form expression for the χ^2-divergence between Gaussians [34].

Since $S-\tilde{S}$ is $\sqrt{2}K$-subgaussian, the RHS above converges if $K<\frac{\sigma}{2}$ and gives the following bound [32, Remark 2.3]

$$I_{\chi^2}(S;Y)\le\exp\left(2d\left(\frac{K}{\sigma}\right)^2\frac{\sigma^2-2K^2}{\sigma^2-4K^2}\right)\,.\tag{13}$$

The second claim follows from Lemma 1. ■

The proof of Proposition 3 immediately implies χ^2 convergence for any compactly supported P.

Corollary 1 (Smooth $\boldsymbol{\chi^2}$ for Bounded Support) *If P has a bounded support with diameter $D\triangleq\sup_{x,y\in\mathrm{supp}(P)}\|x-y\|$, then $I_{\chi^2}(S;Y)\le\exp\left(\frac{D^2}{\sigma^2}\right)$. Consequently,*

$$\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)\le\exp\left(\frac{D^2}{\sigma^2}\right)\cdot\frac{1}{n},\tag{14}$$

Proof: Follows by inserting $\|S-\tilde{S}\|\le D$ into (12). ■

2) Diverging Example: This section shows that for $K\ge\sqrt{2}\sigma$, there exist K-subgaussian distributions P for which $I_{\chi^2}(S;Y)=\infty$.

Let $d=1$ and without loss of generality assume $\sigma=1$. Furthermore, for simplicity of the proof we set $K=\sqrt{2}\sigma$, since the resulting counterexample will apply for any $K\ge\sqrt{2}\sigma$ (recalling that any $\sqrt{2}\sigma$-subgaussian distribution is K-subgaussian for any $K\ge\sqrt{2}\sigma$).

Let $\epsilon=\frac{1}{2K^2}=\frac{1}{4}$ and define the sequence $\{r_k\}_{k=0}^\infty$ by $r_0=0$, $r_1=1$ and $r_k=\frac{r_{k-1}}{1-\sqrt{2\epsilon}}$, for $k\ge2$. Let P be discrete distribution with $\mathrm{supp}(P)=\{r_k\}_{k=0}^\infty$ given by

$$P=\sum_{k=0}^\infty p_k\delta_{r_k},\tag{15a}$$

where δ_x is the Dirac measure at x. We make P K-subgaussian by setting

$$p_k=\begin{cases}2\sqrt{\frac{\epsilon}{\pi}}e^{-\epsilon r_k^2} & k\ge1\\ 1-\sum_{k=1}^\infty p_k & k=0.\end{cases}\tag{15b}$$

Note then that since $\min_{k\ge1}|r_k-r_{k-1}|=1$ and $p_k=2\varphi_K(r_k)$, we get that $\sum_{k=1}^\infty p_k<1$; the remainder of the probability is allocated to $r_0=0$. As stated in the next proposition, $I_{\chi^2}(S;Y)$ diverges when $S\sim P$ as constructed above (which, in turn implies that the χ^2 smoothed empirical approximation is also infinite). This stands in contrast to the classic KL mutual information, which is always finite over an AWGN channel for inputs with a bounded second moment.

Proposition 4 (χ^2 Diverging Example) *For P as in (15) and $\epsilon = 1/4$, we have that*

$$I_{\chi^2}(S;Y) = \infty\,. \tag{16}$$

Consequently

$$\mathbb{E}_{P^{\otimes n}} \chi^2\left(\hat{P}_{S^n} * \mathcal{N}_1 \middle\| P * \mathcal{N}_1\right) = \infty, \quad \forall n \in \mathbb{N}.$$

The proof is given in Appendix D. Intuitively, the constructed P has infinitely many atoms at sufficiently large distance from each other such that the tail contribution at r_k from any $j \neq k$ component of the mixture $P * \mathcal{N}_1$ is negligible. Note that we grow r_k exponentially to counter the exponentially shrinking p_k weights. Since $\mathrm{supp}(P) = \mathbb{N} \cup \{0\}$, for any finite n there are infinitely many atoms which were not sampled in S^n. Since they are sufficiently well-separated, each of these unsampled atoms contributes a constant value to the considered χ^2-divergence, which consequently becomes infinite.

D. 2-Wasserstein Distance and Kullback-Leibler Divergence

One can leverage the above χ^2 results to obtain analogous bounds for KL divergence and for W_2. The squared 2-Wasserstein distance between μ and ν is $\mathsf{W}_2^2(\mu,\nu) \triangleq \inf \mathbb{E}\|X - Y\|^2$, where the infimum is taken over all couplings of μ and ν. The KL divergence is given by $\mathsf{D}_{\mathsf{KL}}(\mu\|\nu) \triangleq \int \log\left(\frac{\mathsf{d}\mu}{\mathsf{d}\nu}\right) \mathsf{d}\mu$.

The behavior of the 2-Wasserstein distance and KL divergence are governed by $I_{\chi^2}(S;Y)$. If $I_{\chi^2}(S;Y) < \infty$, then the KL divergence converges at the rate $O(n^{-1})$. If additionally P is K-subgaussian (for any $K < \infty$), then the squared 2-Wasserstein distance also converges as $O(n^{-1})$. On the other hand, if $I_{\chi^2}(S;Y) = \infty$, then both the KL divergence and the squared 2-Wasserstein distance are $\omega(n^{-1})$ in expectation.

1) Parametric convergence when $I_{\chi^2}(S;Y) < \infty$: Our bounds on the χ^2-divergence immediately imply analogous bounds for the KL divergence.

Proposition 5 (Smooth KL Divergence Approximation) *Fix $d \geq 1$ and $\sigma > 0$. We have*

$$\mathbb{E}_{P^{\otimes n}} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) \leq \frac{1}{n} I_{\chi^2}(S;Y)\,. \tag{17}$$

In particular, if P is K-subgaussian with $K < \frac{\sigma}{2}$, then

$$\mathbb{E}_{P^{\otimes n}} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) \leq c^{(\chi^2)}_{\sigma,d,K} \frac{1}{n} \tag{18a}$$

and if P is supported on a set of diameter D, then

$$\mathbb{E}_{P^{\otimes n}} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) \leq \exp\left(\frac{D^2}{\sigma^2}\right) \cdot \frac{1}{n}. \tag{18b}$$

Proof: The first claim follows directly from Lemma 1 because $D_{\mathsf{KL}}(\mu\|\nu) \leq \log\left(1 + \chi^2(\mu\|\nu)\right)$ for any two probability measures μ and ν. Proposition 3 and Corollary 1 then imply the subsequent claims. ■

To obtain bounds for the 2-Wasserstein distance, we leverage a transport-entropy inequality that connects KL divergence and W_2. We have the following.

Proposition 6 (Smoothed W_2^2 Approximation) *Let $d \geq 1$, $\sigma > 0$ and $K < \infty$. If P is a K-subgaussian distribution and $I_{\chi^2}(S;Y) < \infty$, then*

$$\mathbb{E}_{P^{\otimes n}} \mathsf{W}_2^2(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma) = O\left(\frac{1}{n}\right)\,. \tag{19}$$

In particular, this holds if $K < \frac{\sigma}{2}$. More explicitly, if P is supported on a set of diameter D, then

$$\mathbb{E}_{P^{\otimes n}} \mathsf{W}_2^2(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma) \leq c^{(\mathsf{W}_2)}_{\sigma,d,D} \cdot \frac{1}{n}\,, \tag{20}$$

where $c^{(\mathsf{W}_2)}_{\sigma,d,D}$ is given in (23).

Proof: The subgaussianity of P implies [35, Lemma 5.5] that $\mathbb{E}_P e^{\varepsilon\|S\|^2} < \infty$ for $\varepsilon > 0$ sufficiently small. Therefore, [36, Theorem 1.2] implies that $P * \mathcal{N}_\sigma$ satisfies a log-Sobolev inequality with some constant $C_{P,\sigma}$, depending on P and σ. This further means [37], [38] that $P*\mathcal{N}_\sigma$ satisfies the transport-entropy inequality

$$\mathsf{W}_2^2(Q, P * \mathcal{N}_\sigma) \leq C_{P,\sigma} \mathsf{D}_{\mathsf{KL}}(Q\|P * \mathcal{N}_\sigma) \tag{21}$$

for all probability measures Q. Combining this inequality with Proposition 5 yields the first claim.

If P is supported on a set of diameter D, we have the following more explicit bound:

$$\mathsf{W}_2^2(Q, P * \mathcal{N}_\sigma) \leq c' \sqrt{d} \sigma^2 \left(1 + \frac{D^2}{4\sigma^2}\right) e^{\frac{D^2}{\sigma^2}} \mathsf{D}_{\mathsf{KL}}(Q\|P * \mathcal{N}_\sigma) \tag{22}$$

for an absolute constant c' and any probability measure Q on $\mathbb{R}^d$. Applying Proposition 5 yields

$$\begin{aligned}\mathbb{E}_{P^{\otimes n}} &\mathsf{W}_2^2(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma) \\ &\leq c' \frac{\sqrt{d}\sigma^2}{n} \left(1 + \frac{D^2}{4\sigma^2}\right) \exp\left(\frac{2D^2}{\sigma^2}\right).\end{aligned} \tag{23}$$

■

Remark 2 (W_2^2 Speedup in One Dimension) *The convergence of (unsmoothed) empirical measures in the squared 2-Wasserstein distance suffers from the curse of dimensionality, converging at rate $n^{-2/d}$ when d is large [1], [39]. Proposition 6, however, shows that when smoothed with Gaussian kernels the convergence rate improves to n^{-1} for all d and low SNR ($K < \frac{\sigma}{2}$). Interestingly, the Gaussian smoothing speeds up the convergence rate even for d = 1. For instance, Theorem 7.11 from [40] shows that $\mathbb{E}_{P^{\otimes n}} \mathsf{W}_2^2(\hat{P}_{S^n}, P) \asymp n^{-1/2}$ whenever the support of P is not an interval in $\mathbb{R}$, which is slower than the n^{-1} attained under Gaussian smoothing. Even for the canonical case when P is Gaussian, Corollary 6.14 from [40] shows that $\mathbb{E}_{P^{\otimes n}} \mathsf{W}_2^2(\hat{P}_{S^n}, P) \asymp \frac{\log\log n}{n}$.*

2) Slower convergence when $I_{\chi^2}(S;Y) = \infty$: Unlike the χ^2-divergence, it is easy to see that the 2-Wasserstein distance and KL divergence between the convolved measures are always finite when P is subgaussian. However, when $I_{\chi^2}(S;Y) = \infty$, the rate of convergence of the KL divergence and the squared 2-Wasserstein distance is strictly slower than parametric.

Proposition 7 (Smooth KL Divergence vs. Parametric) *If $I_{\chi^2}(S;Y) = \infty$, for $S \sim P$ and $Y = S + Z$, with $Z \sim \mathcal{N}_\sigma$ independent of S, then*

$$\mathbb{E}_{P^{\otimes n}} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) = \omega\left(\frac{1}{n}\right). \tag{24}$$

For examples of K-subgaussian $S \sim P$ distributions with $I_{\chi^2}(S;Y) = \infty$ see Proposition 4.

Proof: By rescaling, we assume that $\sigma = 1$. Let $S^n \sim P^{\otimes n}$, $Z \sim \mathcal{N}_1$ and $W \sim \mathsf{Unif}([n])$ be independent random variable. Defining $V = S_W + Z$, the proof of Lemma 5 in Appendix C establishes that V has law $P * \mathcal{N}_1$ and that the conditional distribution of V given $S^n = s^n$ is $\hat{P}_{s^n} * \mathcal{N}_1$. This implies

$$\begin{aligned}
\mathbb{E}_{P^{\otimes n}} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_1 \middle\| P * \mathcal{N}_1\right) &= I(S^n; V) \\
&\overset{(a)}{=} \sum_{i=1}^n h(S_i | S^{i-1}) - h(S_i | V, S^{i-1}) \\
&\overset{(b)}{\geq} \sum_{i=1}^n h(S_i) - h(S_i | V) \\
&\overset{(c)}{=} n I(S_1; V),
\end{aligned} \tag{25}$$

where (a) uses $I(S^n;V) = h(S^n) - h(S^n|V)$ and the entropy chain rule, (b) is since S_i are independent (first term) and because conditioning cannot increase entropy (second term), while (c) follows because S_i are identically distributed and since $P_{V|S_i}$ does not depend on i.

Conditioned on $S_1 = s_1$, the random variable V has law $\frac{1}{n}\delta_{s_1} * \mathcal{N}_\sigma + \frac{n-1}{n} P * \mathcal{N}_1$. Therefore

$$I(S_1;V) = \mathbb{E}_P \mathsf{D}_{\mathsf{KL}}\left(\frac{1}{n}\delta_{S_1} * \mathcal{N}_1 + \frac{n-1}{n} P * \mathcal{N}_1 \middle\| P * \mathcal{N}_1\right). \tag{26}$$

Let $S \sim P$ and $Y = S + Z$, and denote by P_S and P_Y the distributions of S and Y, respectively. By Fatou's lemma,

$$\begin{aligned}
&\liminf_{n \to \infty} n \mathbb{E}_{P^{\otimes n}} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_1 \middle\| P * \mathcal{N}_1\right) \\
&\geq \liminf_{n \to \infty} \mathbb{E}_{P_S} n^2 \mathsf{D}_{\mathsf{KL}}\left(\frac{1}{n}\delta_S * \mathcal{N}_1 + \frac{n-1}{n} P * \mathcal{N}_1 \middle\| P * \mathcal{N}_1\right) \\
&= \liminf_{n \to \infty} \mathbb{E}_{P_S} n^2 \mathsf{D}_{\mathsf{KL}}\left(\frac{1}{n} P_{Y|S} + \frac{n-1}{n} P_Y \middle\| P_Y\right) \\
&= \liminf_{n \to \infty} n^2 \mathsf{D}_{\mathsf{KL}}\left(\frac{1}{n} P_{S,Y} + \frac{n-1}{n} P_S \otimes P_Y \middle\| P_S \otimes P_Y\right) \\
&\quad \overset{(a)}{\geq} \mathbb{E}_{P_S \otimes P_Y}\left(\frac{\mathrm{d}P_{S,Y}}{\mathrm{d}P_S \otimes P_Y} - 1\right)^2 \\
&\quad = I_{\chi^2}(S;Y) \\
&\quad = \infty,
\end{aligned} \tag{27}$$

where (a) follows by [41, Proposition 4.2]. We conclude that $\mathbb{E}_{P^{\otimes n}} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_1 \middle\| P * \mathcal{N}_1\right) = \omega(n^{-1})$. ■

We likewise obtain an analogous claim for W_2, showing that the squared 2-Wasserstein distance converges as $\omega(n^{-1})$ when smoothed by any Gaussian with strictly smaller variance.

Corollary 2 (Smooth W_2^2 Slower than Parametric) *If $I_{\chi^2}(S;Y) = \infty$, for $S \sim P$ and $Y = S + Z$, with $Z \sim \mathcal{N}_\sigma$ independent of S, then for any $\tau < \sigma$,*

$$\mathbb{E}_{P^{\otimes n}} \mathsf{W}_2^2(\hat{P}_{S^n} * \mathcal{N}_\tau, P * \mathcal{N}_\tau) = \omega\left(\frac{1}{n}\right). \tag{28}$$

For examples of K-subgaussian $S \sim P$ distributions with $I_{\chi^2}(S;Y) = \infty$ see Proposition 4.

Proof: We assume as above that $\sigma = 1$. Let $S^n \sim P^{\otimes n}$, define $T_i \triangleq \frac{K}{\sqrt{2}} S_i$, and Let $\lambda \triangleq \sqrt{1 - \tau^2} > 0$. If $P_{R_n,R}$ is any coupling between $\hat{P}_{S^n} * \mathcal{N}_\tau$ and $P * \mathcal{N}_\tau$, then the joint convexity of the KL divergence implies that

$$\begin{aligned}
&\mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) \\
&\quad = \mathsf{D}_{\mathsf{KL}}\left(\mathbb{E}_{P_{R_n}} \mathcal{N}(R_n, \lambda^2 \mathrm{I}_d) \middle\| \mathbb{E}_{P_R} \mathcal{N}(R, \lambda^2 \mathrm{I}_d)\right) \\
&\quad \leq \mathbb{E}_{P_{R_n,R}} \mathsf{D}_{\mathsf{KL}}\left(\mathcal{N}(R_n, \lambda^2 \mathrm{I}_d) \middle\| \mathcal{N}(R, \lambda^2 \mathrm{I}_d)\right) \\
&\quad = \mathbb{E}_{P_{R_n,R}} \frac{1}{2\lambda^2} \|R_n - R\|^2,
\end{aligned} \tag{29}$$

where the last step uses the explicit expression for the KL divergence between isotropic Gaussians. Taking an infimum over all valid couplings yields

$$\mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_1 \middle\| P * \mathcal{N}_1\right) \leq \frac{1}{2\lambda^2} \mathsf{W}_2^2(\hat{P}_{S^n} * \mathcal{N}_\tau, P * \mathcal{N}_\tau). \tag{30}$$

Proposition 7 then implies

$$\begin{aligned}
\mathbb{E}_{P^{\otimes n}} \mathsf{W}_2^2(\hat{P}_{S^n} * \mathcal{N}_\tau, &P * \mathcal{N}_\tau) \\
&\geq 2\lambda^2 \mathbb{E}_{P^{\otimes S^n}} \mathsf{D}_{\mathsf{KL}}\left(\hat{P}_{S^n} * \mathcal{N}_1 \middle\| P * \mathcal{N}_1\right) \\
&= \omega\left(\frac{1}{n}\right),
\end{aligned}$$

which concludes the proof. ■

E. Open Questions

We list here some open questions that remain unanswered by the above. First, recall we focused on bounds of the form:

$$\mathbb{E}_{P^{\otimes n}} \delta\left(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right) \leq C(d) \frac{1}{n}.$$

What is the correct dependence of $C(d)$ on dimension? For $\delta = \mathsf{W}_1$ we proved a bound with $C(d) = e^{O(d)}$ for all subgaussian P. Similarly, for $\delta = \mathsf{W}_2^2$ we have shown bounds with $C(d) = e^{O(d)}$ (for small-variance subgaussian P) and $C(d) = \sqrt{d} e^{O(D^2)}$ (for P supported on a set of diameter D). What is the sharp dependence on dimension? Does it change as a function of the subgaussian constant?

A second, and perhaps more interesting, direction is to understand the rate of convergence of W_2^2 in cases when it is $\omega(n^{-1})$. A proof similar to Proposition 1 can be used to show that for subgaussian P (with any constant), we have

$$\mathbb{E}_{P^{\otimes n}} \mathsf{W}_2^2\left(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\right) = O\left(\frac{1}{\sqrt{n}}\right).$$

Is this ever sharp? What rates are possible in the range between $\omega(n^{-1})$ and $O(n^{-1/2})$?

Heuristically, one may think that since W_1 converges as $n^{-1/2}$, then some truncation argument should be able to recover $\mathsf{W}_2^2 \lesssim \frac{\mathrm{polylog}(n)}{n}$. Rigorizing this reasoning requires, however, analyzing the distance distribution between $A \sim \hat{P}_{S^n} * \mathcal{N}_\sigma$ and $B \sim P * \mathcal{N}_\sigma$ under the optimal W_1-coupling. The TV-coupling that was used in Proposition 1 will not work here because under it we have $\mathbb{P}\big(\|A-B\| > \Omega(1)\big) = \Omega(n^{-1/2})$, which results in the $n^{-1/2}$ rate for W_2^2.

Finally, as we saw, the finiteness of $I_{\chi^2}(S;Y)$ is a sufficient condition for many of the above empirical measure convergence results. When $S \sim P$ is K-subgaussian with $K < \frac{\sigma}{2}$, Proposition 3 shows that $I_{\chi^2}(S;Y) < \infty$ always holds. However, for $K \geq \sqrt{2}\sigma$, there exist K-subgaussian distribution for which $I_{\chi^2}(S;Y) = \infty$ (Proposition 4). Characterizing the sharp threshold at which $I_{\chi^2}(S;Y)$ may diverge is another open question.

III. Differential Entropy Estimation

Our main application of the Gaussian smoothed empirical approximation questions is the estimation of $h(P * \mathcal{N}_\sigma)$, based on samples $S^n \sim P^{\otimes n}$ and knowledge of σ. The d-dimensional distribution P is unknown and belongs to some nonparametric class. We first consider the class $\mathcal{F}_d$ of all distributions P with $\mathrm{supp}(P) \subseteq [-1,1]^d$.[8] The second class of interest is $\mathcal{F}_{d,K}^{(\mathsf{SG})}$, which contains all K-subgaussian distributions (see Definition 1).

A. Lower Bounds on Risk

The sample complexity for estimating $h(P * \mathcal{N}_\sigma)$ over the class $\mathcal{F}_d$ is $n^\star(\eta,\sigma,\mathcal{F}_d) \triangleq \min\big\{n \in \mathbb{N} : \mathcal{R}^\star(n,\sigma,\mathcal{F}_d) \leq \eta\big\}$, where $\mathcal{R}^\star(n,\sigma,\mathcal{F}_d)$ is defined in (1). As claimed next, the sample complexity of any estimator is exponential in d.

Theorem 1 (Exponential Sample Complexity) *The following claims hold:*

1) *Fix $\sigma > 0$. There exist $d_0(\sigma) \in \mathbb{N}$, $\eta_0(\sigma) > 0$ and $\gamma(\sigma) > 0$ (monotonically decreasing in σ), such that for all $d \geq d_0(\sigma)$ and $\eta < \eta_0(\sigma)$, we have $n^\star(\eta,\sigma,\mathcal{F}_d) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$.*
2) *Fix $d \in \mathbb{N}$. There exist $\sigma_0(d), \eta_0(d) > 0$, such that for all $\sigma < \sigma_0(d)$ and $\eta < \eta_0(d)$, we have $n^\star(\eta,\sigma,\mathcal{F}_d) = \Omega\left(\frac{2^d}{\eta d}\right)$.*

Theorem 1 is proven in Section VI-A, based on channel coding arguments. For instance, the proof of Part 1 relates the estimation of $h(P * \mathcal{N}_\sigma)$ to discrete entropy estimation of a distribution supported on a capacity-achieving codebook for a peak-constrained AWGN channel. Since the codebook size is exponential in d, discrete entropy estimation over the codebook within a small gap $\eta > 0$ is impossible with less than order of $\frac{2^{\gamma(\sigma)d}}{\eta d}$ samples [6], [7]. The exponent $\gamma(\sigma)$ is monotonically decreasing in σ, implying that larger σ values are favorable for estimation. The 2nd part of the theorem relies on a similar argument but for a d-dimensional AWGN channel and an input constellation that comprises the vertices of the d-dimensional hypercube $[-1,1]^d$.

Remark 3 (Complexity for Restricted Distribution Classes) *Restricting $\mathcal{F}_d$ by imposing smoothness or lower-boundedness assumptions on the distributions in the class would not alleviate the exponential dependence on d from Theorem 1. For instance, consider convolving any $P \in \mathcal{F}_d$ with $\mathcal{N}_{\frac{\sigma}{\sqrt{2}}}$, i.e., replacing each P with $Q = P * \mathcal{N}_{\frac{\sigma}{\sqrt{2}}}$. These Q distributions are smooth, but if one could accurately estimate $h\big(Q * \mathcal{N}_{\frac{\sigma}{\sqrt{2}}}\big)$ over the convolved class, then $h(P * \mathcal{N}_\sigma)$ over $\mathcal{F}_d$ could have been estimated as well. Therefore, Theorem 1 applies also for the class of such smooth Q distributions.*

The next propositions shows that the absolute-error risk attained by any estimator of $h(P * \mathcal{N}_\sigma)$ decays no faster than $n^{-1/2}$.

Proposition 8 (Risk Lower Bound) *For any $\sigma > 0,\ d \geq 1$, we have*

$$\mathcal{R}^\star\left(n,\sigma,\mathcal{F}_{d,K}^{(\mathsf{SG})}\right) = \Omega\left(\frac{1}{\sqrt{n}}\right). \tag{31}$$

That proposition states the so-called parametric lower bound on the absolute-error estimation risk. Under (the square root of the) quadratic loss, the result trivially follows from the Cramér-Rao lower bound. For completeness, Appendix A provides a simple proof for the absolute-error loss considered herein, based on the Hellinger modulus [42].

B. Upper Bound on Risk

We establish the minimax-rate optimality of the plug-in estimator by showing its risk converges as $n^{-1/2}$. Our risk bounds provide explicit constants (in terms of σ, K and d). These constants depend exponentially on dimension, in accordance to the results of Theorem 1. Recall that given a collection of samples $S^n \sim P^{\otimes n}$, the estimator is $h(\hat{P}_{S^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{S^n} = \frac{1}{n}\sum_{i=1}^n \delta_{S_i}$ is the empirical measure.

A risk bound for the bounded support case is presented first. Although a special case of Theorem 3, where the subgaussian class $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ is considered, we state the bounded support result separately since it gives a cleaner bound with a better constant.

Theorem 2 (Plug-in Risk Bound - Bounded Support Class) *Fix $\sigma > 0$ and $d \geq 1$. For any n, we have*

$$\sup_{P \in \mathcal{F}_d} \mathbb{E}_{P^{\otimes n}} \left| h(P * \mathcal{N}_\sigma) - h(\hat{P}_{S^n} * \mathcal{N}_\sigma) \right| \leq \tilde{C}_{\sigma,d} \frac{1}{\sqrt{n}}, \tag{32}$$

where $\tilde{C}_{\sigma,d} = O_\sigma(c^d)$, for a numerical constant c, is explicitly characterized in (62).

The proof (given in Section VI-B) relies on the χ^2-divergence $n^{-1/2}$ convergence rate from Corollary 1. Specifically, we relate the differential entropy estimation error to $\mathbb{E}_{P^{\otimes n}} \chi^2\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right)$ using χ^2 variational representation. The result then follows by controlling certain variance terms and using Corollary 1.

We next bound the estimation risk when $P \in \mathcal{F}_{d,K}^{(\mathsf{SG})}$.

[8] One may consider any other class of compactly supported distributions.

Theorem 3 (Plug-in Risk Bound - Subgaussian Class) *Fix $\sigma > 0$ and $d \geq 1$. For any n, we have*

$$\sup_{P\in\mathcal{F}_{d,K}^{(\mathsf{SG})}} \mathbb{E}_{P^{\otimes n}} \left| h(P*\mathcal{N}_\sigma) - h(\hat{P}_{S^n}*\mathcal{N}_\sigma)\right| \leq C_{\sigma,d,K}\frac{1}{\sqrt{n}}, \tag{33}$$

where $C_{\sigma,d,K} = O_{\sigma,K}(c^d)$, for a numerical constant c, is explicitly characterized in (69)*.*

The proof of Theorem 3 is given in Section VI-C. While the result follows via arguments similar to the bounded support case (namely, through the χ^2 subgaussian bound from Proposition 3), this method only covers the regime $\frac{K}{\sigma} < \frac{1}{2}$. To prove Theorem 3 without restricting σ and K, we resort to a different argument. Using the maximal TV-coupling, we bound the estimation risk by a weighted TV distance between $P*\mathcal{N}_\sigma$ and $\hat{P}_{S^n}*\mathcal{N}_\sigma$. The smoothing induced by the Gaussian convolutions allows us to control this TV distance by a $e^{O(d)}n^{-1/2}$. Several things to note about the result are the following:

1) The theorem does not require any smoothness conditions on the distributions in $\mathcal{F}_{d,K}^{(\mathsf{SG})}$. This is achievable due to the inherent smoothing introduced by the convolution with the Gaussian density. Specifically, while the differential entropy $h(q)$ is not a smooth functional of the underlying density q in general, our functional is $\mathsf{T}_\sigma(P) \triangleq h(P*\mathcal{N}_\sigma)$, which is smooth.
2) The above smoothness also allows us to avoid any assumptions on P being bounded away from zero. So long as P has subgaussian tails, the distribution may be arbitrary.

Remark 4 (Knowledge of Noise Parameter) *Our original motivation for this work is the noisy DNN setting, where additive Gaussian noise is injected into the system to enable tracking "information flows" during training (see [28]). In this setting, the parameter σ is known and the considered observation model reflects this. However, an interesting scenario is when σ is unknown. To address this, first note that samples from P contain no information about σ. Hence, in the setting where σ is unknown, presumably samples of both $S \sim P$ and $S+Z \sim P*\mathcal{N}_\sigma$ would be available. Under this alternative model, estimating σ can be done immediately by comparing the empirical variance of S and $S+Z$. This empirical proxy would converge as $O\left((nd)^{-1/2}\right)$, implying that for large enough dimension, the empirical σ can be substituted into our entropy estimator (in place of the true σ) without affecting the $O\left(c^d n^{-1/2}\right)$ convergence rate.*

C. Bias Lower Bound

To have a guideline as to the smallest number of samples needed to avoid biased estimation, we present the following lower bound on the estimator's bias $\sup_{P\in\mathcal{F}_d}\left|h(P*\mathcal{N}_\sigma) - \mathbb{E}_{P^{\otimes n}}h(\hat{P}_{S^n}*\mathcal{N}_\sigma)\right|$.

Theorem 4 (Bias Lower Bound) *Fix $d \geq 1$ and $\sigma > 0$, and let $\epsilon \in \left(1-\left(1-2Q\left(\frac{1}{2\sigma}\right)\right)^d, 1\right]$, where Q is the Q-function.*[9] *Set $k_\star \triangleq \left\lfloor \frac{1}{\sigma Q^{-1}\left(\frac{1}{2}\left(1-(1-\epsilon)^{\frac{1}{d}}\right)\right)}\right\rfloor$, where Q^{-1} is the inverse of the Q-function. By the choice of ϵ, clearly $k_\star \geq 2$, and we have*

$$\sup_{P\in\mathcal{F}_d}\left|h(P*\mathcal{N}_\sigma) - \mathbb{E}_{P^{\otimes n}}h(\hat{P}_{S^n}*\mathcal{N}_\sigma)\right| \geq \log\left(\frac{k_\star^{d(1-\epsilon)}}{n}\right) - H_b(\epsilon). \tag{34}$$

Consequently, the bias cannot be less than a given $\delta > 0$ so long as $n \leq k_\star^{d(1-\epsilon)}\cdot e^{-(\delta+H_b(\epsilon))}$.

The theorem is proven in Section VI-D. Since $H_b(\epsilon)$ shrinks with ϵ, for sufficiently small ϵ values, the lower bound from (34) essentially shows that the our estimator will not have negligible bias unless $n > k_\star^{d(1-\epsilon)}$ is satisfied. The condition $\epsilon > 1-\left(1-2Q\left(\frac{1}{2\sigma}\right)\right)^d$ is non-restrictive in any relevant regime of d and σ. For the latter, values we have in mind are inspired by [28], where noisy DNNs with parameter σ are studied. In that work, σ values are around 0.1, for which the lower bound on ϵ is at most 0.0057 for all dimensions up to at least $d = 10^4$. For example, when setting $\epsilon = 0.01$ (for which $H_b(0.01) \approx 0.056$), the corresponding $k_\star$ equals 3 for $d \leq 11$ and 2 for $12 \leq d \leq 10^4$. Thus, with these parameters, a negligible bias requires n to be at least $2^{0.99d}$.

D. Computing the Estimator

Evaluating the plug-in estimator $h(\hat{P}_{S^n}*\mathcal{N}_\sigma)$ requires computing the differential entropy of a d-dimensional n-mode Gaussian mixture ($\hat{P}_{S^n}*\mathcal{N}_\sigma$). Although it cannot be computed in closed form, this section presents a method for computing an arbitrarily accurate approximation via MC integration [43]. To simplify the presentation, we present the method for an arbitrary Gaussian mixture without referring to the notation of the estimation setup.

Let $g(t) \triangleq \frac{1}{n}\sum_{i=1}^n \varphi_\sigma(t-\mu_i)$ be a d-dimensional, n-mode Gaussian mixture, with centers $\{\mu_i\}_{i=1}^n \subset \mathbb{R}^d$. Let $C \sim \mathsf{Unif}\left(\{\mu_i\}_{i=1}^n\right)$ be independent of $Z \sim \mathcal{N}_\sigma$ and note that $V \triangleq C+Z \sim g$. First, rewrite $h(g)$ as follows:

$$\begin{aligned} h(g) = -\mathbb{E}\log g(V) &= -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\log g(\mu_i+Z)\Big|C=\mu_i\Big] \\ &= -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\log g(\mu_i+Z), \end{aligned} \tag{35}$$

where the last step uses the independence of Z and C. Let $\left\{Z_j^{(i)}\right\}_{\substack{i\in[n]\\ j\in[n_{\mathsf{MC}}]}}$ be $n\times n_{\mathsf{MC}}$ i.i.d. samples from φ_σ. For each $i\in[n]$, we estimate the i-th summand on the RHS of (35) by

$$\hat{I}_{\mathsf{MC}}^{(i)} \triangleq \frac{1}{n_{\mathsf{MC}}}\sum_{j=1}^{n_{\mathsf{MC}}}\log g\left(\mu_i + Z_j^{(i)}\right), \tag{36a}$$

which produces

$$\hat{h}_{\mathsf{MC}} \triangleq \frac{1}{n}\sum_{i=1}^n \hat{I}_{\mathsf{MC}}^{(i)} \tag{36b}$$

as the approximation of $h(g)$. Note that since g is a mixture of n Gaussians, it can be efficiently evaluated using off-the-shelf

[9] The Q-function is defined as $Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{t^2}{2}}\,dt$.

KDE software packages, many of which require only $O(\log n)$ operations on average per evaluation of g.

Define the MSE of $\hat{h}_{\mathsf{MC}}$ as

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \triangleq \mathbb{E}\left[\left(\hat{h}_{\mathsf{MC}} - h(g)\right)^2\right]. \tag{37}$$

We have the following bounds on the MSE.

Theorem 5 (MSE Bounds for the MC Estimator)

(i) *Bounded support: Assume $C \in [-1,1]^d$ almost surely, then*

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \leq \frac{2d(2+\sigma^2)}{\sigma^2}\frac{1}{n \cdot n_{\mathsf{MC}}}. \tag{38}$$

(ii) *Bounded moment: Assume $m \triangleq \mathbb{E}\|C\|_2^2 < \infty$, then*

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \leq \tag{39}$$
$$\frac{9d\sigma^2 + 8(2+\sigma\sqrt{d})m + 3(11\sigma\sqrt{d}+1)\sqrt{m}}{\sigma^2}\frac{1}{n \cdot n_{\mathsf{MC}}}.$$

The proof is given in Section VI-E. The bounds on the MSE scale only linearly with the dimension d, making σ^2 in the denominator often the dominating factor experimentally.

IV. Information Flow in Deep Neural Networks

A utilization of the developed theory is estimating the mutual information between selected groups of neurons in DNNs. Much attention was recently devoted to this task [23]–[28], partly motivated by the Information Bottleneck (IB) theory for DNNs [29], [30]. The theory tracks the mutual information pair $\big(I(X;T), I(Y;T)\big)$, where X is the DNN's input (i.e., feature), Y is the true label and T is the hidden representation vector. An interesting claim from [30] is that the mutual information $I(X;T)$ undergoes a so-called 'compression' phase during training. Namely, after an initial short 'fitting' phase (where $I(Y;T)$ and $I(X;T)$ both grow), $I(X;T)$ exhibits a slow long-term decrease, which is termed the 'compression' phase. According to [30], this phase explains the excellent generalization performance of DNNs.

The main caveat in the supporting empirical results from [30] (and the partially opposing results from the followup work [23]) is that in deterministic networks, where $T = f(X)$ with strictly monotone activations, $I(X;T)$ is either infinite (when the data distribution P_X is continuous) or a constant (when P_X is discrete[10]). As explained in [28], the reason [30] and [23] miss this fact stems from an inadequate application of a binning-based mutual information estimator for $I(X;T)$.

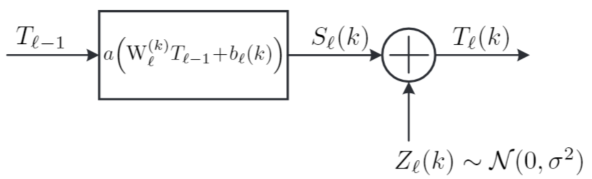

$T_{\ell-1}$
$a\left(\mathrm{W}_\ell^{(k)}T_{\ell-1}+b_\ell(k)\right)$
$S_\ell(k)$
$T_\ell(k)$
$Z_\ell(k) \sim \mathcal{N}(0,\sigma^2)$

Fig. 1: k-th noisy neuron in a fully connected or a convolutional layer ℓ with activation function a; $\mathrm{W}_\ell^{(k)}$ and $b_\ell(k)$ are the k-th row and the k-th entry of the weight matrix and the bias vector, respectively.

To fix this constant/infinite mutual information issue, [28] proposed the framework of noisy DNNs, in which each neuron adds a small amount of Gaussian noise (i.i.d. across all neurons) after applying the activation function. The injected noise makes the map $X \mapsto T$ a stochastic parameterized channel, and as a consequence, $I(X;T)$ is a finite quantity that depends on the network's parameters. Although the primary purpose of the noise injection in [28] was to ensure that $I(X;T)$ depends on the system parameters, experimentally it was found that the network's performance is optimized at non-zero noise variance, thus providing a natural to select this parameter. In the following, we first define noisy DNNs and then show that estimating $I(X;T)$, $I(Y;T)$ or any other mutual information term between layers of a noisy DNN can be reduced to differential entropy estimation under Gaussian convolutions. The reduction relies on a sampling procedure that leverages the DNN's generative model.

A. Noisy DNNs and Mutual Information between Layers

We start by describing the noisy DNN setup from [28]. Let $(X,Y) \sim P_{X,Y}$ be a feature-label pair, where $P_{X,Y}$ is the (unknown) true distribution of (X,Y), and $\big\{(X_i,Y_i)\big\}_{i=1}^n$ be n i.i.d. samples from $P_{X,Y}$.

Consider an $(L+1)$-layered (fixed / trained) noisy DNN with layers $T_0, T_1, \ldots, T_L$, input $T_0 = X$ and output $T_L = \hat{Y}$ (i.e., an estimate of Y). For each $\ell \in [L-1]$, the ℓ-th hidden layer is given by $T_\ell = S_\ell + Z_\ell$, where $S_\ell \triangleq f_\ell(T_{\ell-1})$ with $f_\ell : \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell}$ being a deterministic function of the previous layer and $Z_\ell \sim \mathcal{N}\big(0, \sigma^2\mathrm{I}_{d_\ell}\big)$ being the noise injected at layer ℓ. The functions $f_1, f_2, \ldots, f_L$ can represent any type of layer (fully connected, convolutional, max-pooling, etc.). For instance, $f_\ell(t) = a(\mathrm{W}_\ell t + b_\ell)$ for a fully connected or a convolutional layer, where a is the activation function which operates on a vector component-wise, $\mathrm{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ is the weight matrix and $b_\ell \in \mathbb{R}^{d_\ell}$ is the bias. For fully connected layers W_ℓ is arbitrary, while for convolutional layers W_ℓ is Toeplitz. Fig. 1 shows a neuron in a noisy DNN.

The noisy DNN induces a stochastic map from X to the rest of the network, described by the conditional distribution $P_{T_1,\ldots,T_L|X}$. The joint distribution of the tuple $(X,Y,T_1,\ldots,T_L)$ is $P_{X,Y,T_1,\ldots,T_L} \triangleq P_{X,Y}P_{T_1,\ldots,T_L|X}$ under which $Y - X - T_1 - \ldots - T_L$ forms a Markov chain. For any $\ell \in [L-1]$, consider the mutual information between the hidden layer and the input (see Remark 6 for an account of $I(Y;T_\ell)$):

$$\begin{aligned} I(X;T_\ell) &= h(T_\ell) - h(T_\ell|X) \\ &= h(P_{T_\ell}) - \int dP_X(x) h(P_{T_\ell|X=x}). \end{aligned} \tag{40}$$

[10] The mapping from the discrete values of X to T is almost always (except for a measure-zero set of weights) injective whenever the nonlinearities are, thereby causing $I(X;T) = H(X)$ for any hidden layer T, even if T consists of a single neuron.

Since P_{T_ℓ} and $P_{T_\ell|X}$ have a highly complicated structure (due to the composition of Gaussian noises and nonlinearities), this mutual information cannot be computed analytically and must be estimated. Based on the expansion from (40), an estimator of $I(X;T_\ell)$ is constructed by estimating the unconditional and each of the conditional differential entropy terms, while approximating the expectation by an empirical average. As explained next, all these entropy estimation tasks are instances of our framework of estimating $h(P * \mathcal{N}_\sigma)$ based on samples from P and knowledge of σ.

B. From Differential Entropy to Mutual Information

Recall that $T_\ell = S_\ell + Z_\ell$, where $S_\ell \sim P_{S_\ell} = P_{f_\ell(T_{\ell-1})}$ and $Z_\ell \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_{d_\ell})$ are independent. Thus,

$$h(P_{T_\ell}) = h(P_{S_\ell} * \mathcal{N}_\sigma) \tag{41a}$$

and

$$h(P_{T_\ell|X=x_i}) = h(P_{S_\ell|X=x_i} * \mathcal{N}_\sigma). \tag{41b}$$

The DNN's forward pass enables sampling from P_{S_ℓ} and $P_{S_\ell|X}$ as follows:

1) Unconditional Sampling: To generate the sample set from P_{S_ℓ}, feed each X_i, for $i \in [n]$, into the DNN and collect the outputs it produces at the $(\ell-1)$-th layer. The function f_ℓ is then applied to each collected output to obtain $S_\ell^n \triangleq \{S_{\ell,1}, S_{\ell,2}, \ldots, S_{\ell,n}\}$, which is the a set of n i.i.d. samples from P_{S_ℓ}.
2) Conditional Sampling Given X: To generate i.i.d. samples from $P_{S_\ell|X=x_i}$, for $i \in [n]$, we feed x_i into the DNN n times, collect outputs from $T_{\ell-1}$ corresponding to different noise realizations, and apply f_ℓ on each. Denote the obtained samples by $S_\ell^n(X_i)$.[11]

The knowledge of σ and together with the samples S_ℓ^n and $S_\ell^n(X_i)$ can be used to estimate the unconditional and the conditional entropies, from (41a) and (41b), respectively.

For notational simplicity, we henceforth omit the layer index ℓ. Based on the above sampling procedure we construct an estimator $\hat{I}\big(X^n, \hat{h}\big)$ of $I(X;T)$ using a given estimator $\hat{h}(A^n, \sigma)$ of $h(P * \mathcal{N}_\sigma)$ for P supported inside $[-1,1]^d$ (i.e., a tanh / sigmoid network), based on i.i.d. samples $A^n = \{A_1, \ldots, A_n\}$ from P and knowledge of σ. Assume that $\hat{h}$ attains

$$\sup_{P \in \mathcal{F}_d} \mathbb{E}_{P^{\otimes n}} \left| h(P * \mathcal{N}_\sigma) - \hat{h}(A^n, \sigma) \right| \leq \Delta_{\sigma,d}(n). \tag{42}$$

An example of such an $\hat{h}$ is the estimator $h(\hat{P}_{A^n} * \mathcal{N}_\sigma)$. The corresponding $\Delta_{\sigma,d}(n)$ term is given in Theorem 3. Our estimator for the mutual information is

$$\hat{I}_{\mathsf{Input}}\left(X^n, \hat{h}, \sigma\right) \triangleq \hat{h}(S^n, \sigma) - \frac{1}{n} \sum_{i=1}^{n} \hat{h}\big(S^n(X_i), \sigma\big). \tag{43}$$

The absolute-error estimation risk of $\hat{I}_{\mathsf{Input}}\left(X^n, \hat{h}, \sigma\right)$ is bounded in the following proposition, proven in Section VI-F.

Proposition 9 (Input–Hidden Layer Mutual Information) *For the above described estimation setting, we have*

$$\sup_{P_X} \mathbb{E} \left| I(X;T) - \hat{I}_{\mathsf{Input}}\left(X^n, \hat{h}, \sigma\right) \right| \leq 2\Delta_{\sigma,d}(n) + \frac{d \log\left(1 + \frac{1}{\sigma^2}\right)}{4\sqrt{n}}.$$

The quantity $\frac{1}{\sigma^2}$ is the SNR between S and Z. The larger σ is the easier estimation becomes, since the noise smooths out the complicated P_X distribution. Also note that the dimension of the ambient space in which X lies does not appear in the absolute-risk bound. The bound depends only on the dimension of T (through $\Delta_{\sigma,d}$). This happens because the blurring effect caused by the noise enables uniformly lower bounding $\inf_x h(T|X=x)$ and thereby controlling the variance of the estimator for each conditional entropy. This reduces the impact of X on the estimation error to that of an empirical average converging to its expected value with rate $n^{-1/2}$.

Remark 5 (Subgaussian Class and Noisy ReLU DNNs) *We provide performance guarantees for the plug-in estimator also over the more general class $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ of distributions with subgaussian marginals. This class accounts for the following important cases:*

1) *Distributions with bounded support, which correspond to noisy DNNs with bounded nonlinearities. This case is directly studied through the bounded support class $\mathcal{F}_d$.*
2) *Discrete distributions over a finite set, which is a special case of bounded support.*
3) *Distributions P of a random variable S that is a hidden layer of a noisy ReLU DNN, so long as the input X to the network is itself subgaussian. To see this recall that linear combinations of independent subgaussian random variables are also subgaussian. Furthermore, for any (scalar) random variable A, we have that $\big|\mathsf{ReLU}(A)\big| = \big|\max\{0, A\}\big| \leq |A|$, almost surely. Each layer in a noisy* ReLU *DNN is a coordinate-wise* ReLU *applied to a linear transformation of the previous layer plus a Gaussian noise. Consequently, for a d-dimensional hidden layer S and any $i \in [d]$, one may upper bound $\big\|S(i)\big\|_{\psi_2}$ by a constant, provided that the input X is coordinate-wise subgaussian. This constant depends on the network's weights and biases, the depth of the hidden layer, the subgaussian norm of the input, and the noise variance.*

In the context of estimation of mutual information over DNNs, the input distribution is typically taken as uniform over the dataset [23], [28], [30], adhering to case (2).

Remark 6 (Hidden Layer–Label Mutual Information) *Another quantity of interest is the mutual information between the hidden layer and the true label (see, e.g., [30]). For $(X,Y) \sim P_{X,Y}$, and a hidden layer T in a noisy DNN with input X, the joint distribution of (X,Y,S,T) is $P_{X,Y}P_{S,T|X}$,*

[11] The described sampling procedure is valid for any layer $\ell \geq 2$. For $\ell = 1$, S_1 coincides with $f_1(X)$ but the conditional samples are undefined. Nonetheless, noting that for the first layer $h(T_1|X) = h(Z) = \frac{d}{2}\log(2\pi e \sigma^2)$, we see that no estimation of the conditional entropy is needed. The mutual information estimator given in (43) is modified by replacing the subtracted term with $h(Z)$.

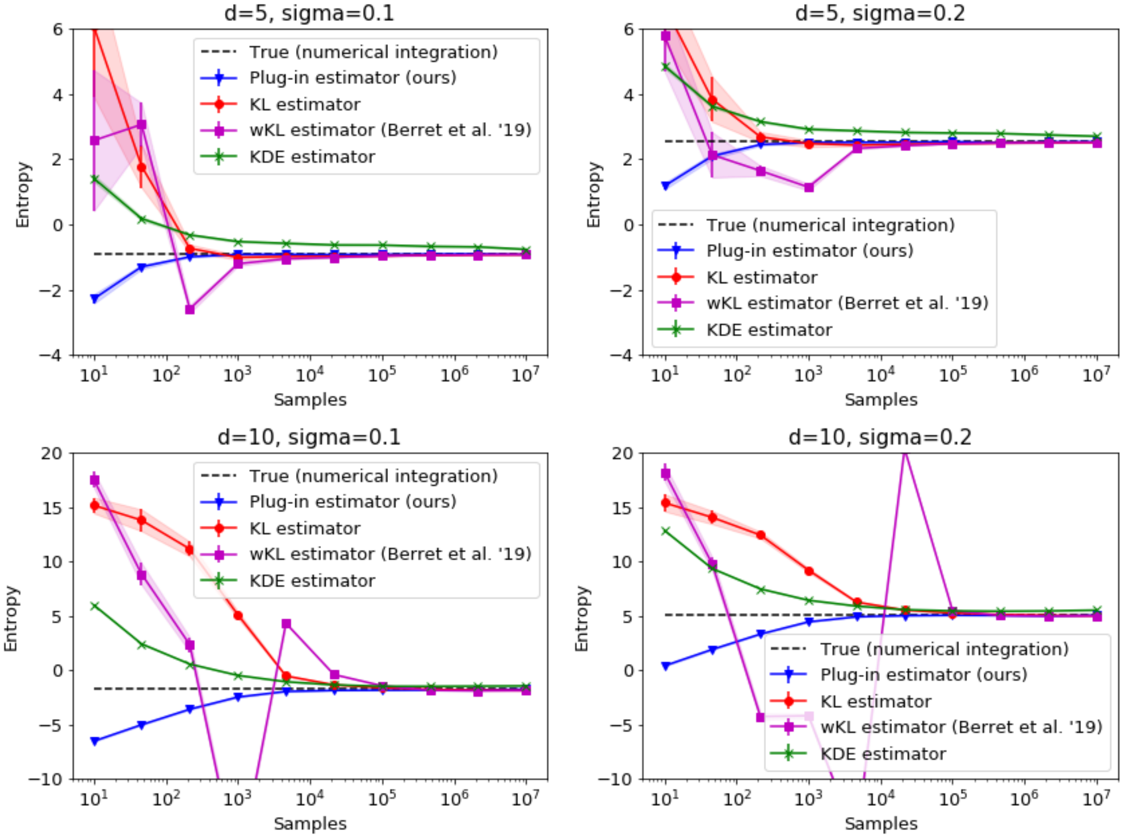


Fig. 2: Estimation results comparing the plug-in estimator to: (i) a KDE-based method [8]; (ii) the KL estimator [11]; and (iii) the wKL estimator [19]. The differential entropy of $S+Z$ is estimated, where S is a truncated d-dimensional mixture of 2^d Gaussians and $Z\sim\mathcal{N}(0,\sigma^2\mathrm{I}_d)$. Results are shown as a function of n, for $d=5,10$ and $\sigma=0.1,0.5$. Error bars are one standard deviation over 20 random trials. The $h(P*\mathcal{N}_\sigma)$ estimator presents faster convergence rates, improved stability and better scalability with dimension compared to the two competing methods.

under which $Y-X-(S,T)$ forms a Markov chain.[12] *The mutual information of interest is then*

$$I(Y;T)=h(P_S*\mathcal{N}_\sigma)-\sum_{y\in\mathcal{Y}}P_Y(y)h(P_{S|Y=y}*\mathcal{N}_\sigma),\quad (44)$$

where $\mathcal{Y}$ is the (known and) finite set of labels. Just like for $I(X;T)$, estimating $I(Y;T)$ reduces to differential entropy estimation under Gaussian convolutions. Namely, an estimator for $I(Y;T)$ can be constructed by estimating the unconditional and each of the conditional differential entropy terms in (44)*, while approximating the expectation by an empirical average. There are several required modifications for estimating $I(Y;T)$ as compared to $I(X;T)$. Most notably is the procedure for sampling from $P_{S|Y=y}$, which results in a sample set whose size is random (Binomial). In appendix B, the estimation of $I(Y;T)$ is described in detail and a corresponding risk bound is derived.*

This section shows that the performance in estimating mutual information depends on our ability to estimate $h(P*\mathcal{N}_\sigma)$. In Section V we present experimental results for $h(P*\mathcal{N}_\sigma)$, when P is induced by a DNN.

V. Simulations

We present empirical results illustrating the convergence of the plug-in estimator compared to several competing methods: (i) the KDE-based estimator of [8]; (ii) and kNN Kozachenko-Leonenko (KL) estimator [11]; and (iii) the recently developed wKL estimator from [19]. These competing methods are general-purpose estimators of the differential entropy $h(Q)$ based on i.i.d. samples from Q. Such methods are applicable for estimating $h(P*\mathcal{N}_\sigma)$ by sampling $\mathcal{N}_\sigma$ and adding the noise values to the samples from P.

A. Simulations for Differential Entropy Estimation

1) P with Bounded Support: Convergence rates in the bounded support regime are illustrated first. We set P as a mixture of Gaussians truncated to have support in $[-1,1]^d$. Before truncation, the mixture consists of 2^d Gaussian components with means at the 2^d corners of $[-1,1]^d$ and standard deviations 0.02. This produces a distribution that is, on one hand, complicated (2^d mixtures) while, on the other hand, is still simple to implement. The entropy $h(P*\mathcal{N}_\sigma)$ is estimated for various values of σ.

Fig. 2 shows estimation results as a function of n, for $d=5,10$ and $\sigma=0.1,0.2$. The KL and plug-in estimators

[12] In fact, the Markov chain is $Y-X-S-T$ since $T=S+Z$, but this is inconsequential here.

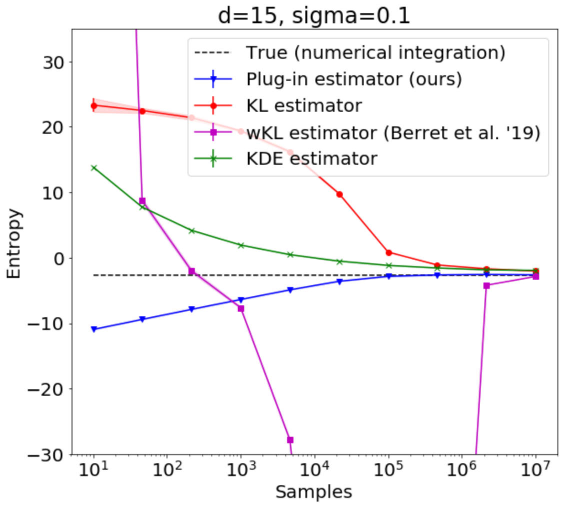


Fig. 3: Estimation results comparing the plug-in estimator to: (i) a KDE-based method [8]; (ii) the KL estimator [11]; and (iii) the wKL estimator [19]. Here P is an untruncated d-dimensional mixture of 2^d Gaussians and $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. Results are shown as a function of n, for $d = 5, 10$ and $\sigma = 0.01, 0.1, 0.5$. Error bars are one standard deviation over 20 random trials.

require no tuning parameters; for wKL we used the default weight setting in the publicly available software. We stress that the KDE estimate is highly unstable and, while not shown here, the estimated value is very sensitive to the chosen kernel width. The kernel width (varying with both d and n) for the KDE estimate was chosen by generating a variety of different Gaussian mixture constellations of moderately different cardinalities and optimizing the kernel width for good performance across regimes (evaluated by comparing finite sample estimates to the large-sample entropy estimate).[13] As seen in Fig. 2, the KDE, KL and wKL estimators converge slowly, at a rate that degrades with increased d, underperforming the plug-in estimator. Finally, we note that in accordance to the explicit risk bound from (69), the absolute error increases with larger d and smaller σ.

2) *P with Unbounded Support:* In Fig. 3, we show the convergence rates in the unbounded support regime by considering the same setting with $d = 15$ but without truncating the 2^d-mode Gaussian mixture. The fast convergence of the plug-in estimator is preserved, outperforming the competing methods. Notice that the performance of the wKL estimator from [19] (whose asymptotic efficiency was established therein) deteriorates in this relatively high-dimensional setup. This may be a result of the dependence of its estimation error on d, which was not characterized in [19].

B. Monte Carlo Integration

Fig. 4 illustrates the convergence of the MC integration method for computing the plug-in estimator. The figure shows the root-MSE (RMSE) as a function of MC samples n_{MC}, for the truncated 2^d Gaussian mixture distribution with $n = 10^4$ (which corresponds to the number of modes in the Gaussian mixture $\hat{P}_{S^n} * \mathcal{N}_\sigma$ whose entropy approximates $h(P * \mathcal{N}_\sigma)$), $d = 5, 10, 15$, and $\sigma = 0.01, 0.1$. Note the error decays approximately as $n_{\mathsf{MC}}^{1/2}$ in accordance with Theorem 5, and that the convergence does not vary excessively for different d and σ values.

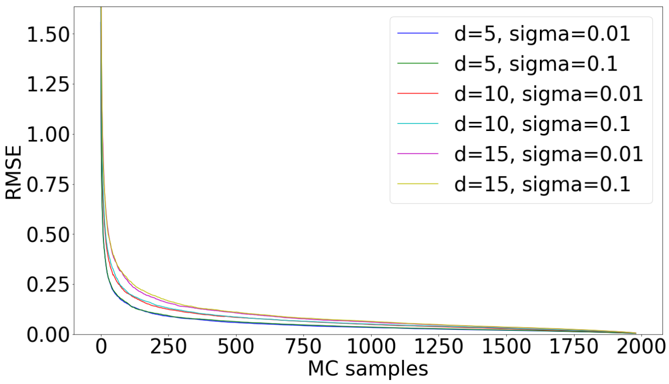
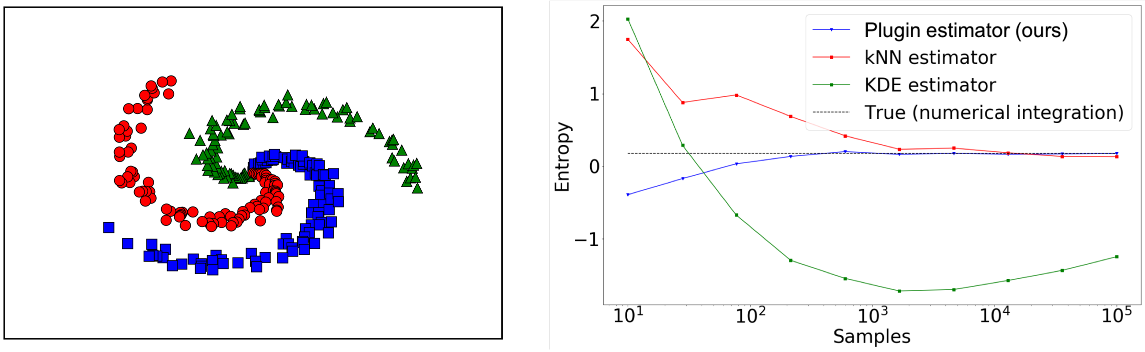


Fig. 4: Convergence of the Monte Carlo integrator computation of the proposed estimator. Shown is the decay of the RMSE as the number of Monte Carlo samples increases, for a variety of σ and d values. The MC integrator is computing the $h(P * \mathcal{N}_\sigma)$ estimate of the entropy of $S + Z$ where S is a truncated d-dimensional mixture of 2^d Gaussians and $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. The number of samples of S used by $h(P * \mathcal{N}_\sigma)$ is 10^4.

C. Estimation in a Noisy Deep Neural Network

We next illustrate entropy estimation in a noisy DNN. The dataset is a 2-dimensional 3-class spiral (shown in Fig. 5(a)). The network has 3 fully connected layers of sizes 8-9-10, with tanh activations and $\mathcal{N}(0, \sigma^2)$ Gaussian noise added to the output of each neuron, where $\sigma = 0.2$. We estimate the entropy of the output of the 10-dimensional third layer in the network trained to achieve 98% classification accuracy. Estimation results are shown in Fig. 5(b), comparing the

[13] This gives an unrealistic advantage to KDE method, but we prefer this to uncertainty about baseline performance that stems from inferior kernel width selection.



(a)

(b)

Fig. 5: 10-dimensional entropy estimation in a 3-layer neural network trained on the 2-dimensional 3-class spiral dataset shown on the left. Estimation results for the plug-in estimator compared to general-purpose kNN and KDE methods are shown on the right. The differential entropy of $S+Z$ is estimated, where S is the output of the third (10-dimensional) layer. Results are shown as a function of samples n with $\sigma = 0.2$.

plug-in estimator to the KDE and KL estimators; the wKL estimator from [19] is omitted due to its poor performance in this experiment. As before, the plug-in estimate converges faster than the competing methods illustrating its efficiency for entropy and mutual information estimation over noisy DNNs. The KDE estimate, which performed quite well in the synthetic experiments, underperform here. In our companion work [28], additional examples of mutual information estimation in DNNs based on the proposed estimator are provided.

D. Reed-Muller Codes over AWGN Channels

We next consider data transmission over an AWGN channel using a binary phase-shift keying (BPSK) modulation of a Reed-Muller code. A Reed-Muller code $\mathsf{RM}(r,m)$ of parameters $r,m \in \mathbb{N}$, where $0 \leq r \leq m$, encodes messages of length $k = \sum_{i=0}^{r} \binom{m}{i}$ into 2^m-lengthed binary codewords. Let $\mathcal{C}_{\mathsf{RM}(r,m)}$ be set of BPSK modulated sequences corresponding to $\mathsf{RM}(r,m)$ (with 0 and 1 mapped to -1 and 1, respectively). The number of bits reliably transmittable over the 2^m-dimensional AWGN with noise $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_{2^m})$ is given by

$$I(S;S+Z) = h(S+Z) - 2^{m-1}\log(2\pi e \sigma^2), \tag{45}$$

where $S \sim \mathsf{Unif}(\mathcal{C}_{\mathsf{RM}(r,m)})$ and Z are independent. Despite $I(S;S+Z)$ being a well-behaved function of σ, an exact computation of this quantity is infeasible.

Our estimator readily estimates $I(S;S+Z)$ from samples of S. Results for the Reed-Muller codes $\mathsf{RM}(4,4)$ and $\mathsf{RM}(5,5)$ (containing 2^{16} and 2^{32} codewords, respectively) are shown in Fig. 6 for various values of σ and n. Fig. 6(a) shows our estimate of $I(S;S+Z)$ for an $\mathsf{RM}(4,4)$ code as a function of σ, for different values of n. As expected, the plug-in estimator converges faster when σ is larger. Fig. 6(b) shows the estimated $I(S;S+Z)$ for $S \sim \mathsf{Unif}(\mathcal{C}_{\mathsf{RM}(5,5)})$ and $\sigma = 2$, with the KDE and KL estimates based on samples of $(S+Z)$ shown for comparison. Our method significantly outperforms the competing general-purpose methods (with the wKL estimator being again omitted due to its instability in this high-dimensional ($d = 32$) setting).

Remark 7 (AWGN with Input Constraint) *When* $\mathrm{supp}(P)$ *lies inside a ball of radius* $\sqrt{d}$*, the subgaussian constant* K *is proportional to* d*, and the bound from* (33) *scales like* $d^{d/2}n^{-1/2}$*. This corresponds to the popular setup of an AWGN channel with an input constraint.*

Remark 8 (Calculating the Ground Truth) *To compute the true value of* $I(S;S+Z)$ *in Fig. 6(b) (dashed red line) we used our MC integrator and the fact the Reed-Muller code is known. Specifically, the distribution of* $S+Z$ *is a Gaussian mixture, whose differential entropy we compute via the expression from* (36)*. Convergence of the computed value was ensured using Theorem 5.*

VI. Proofs for Section III

A. Proof of Theorem 1

1) Part 1: Consider a AWGN channel $Y = X + N$, where the input X is bound to a peak constraint $X \in [-1,1]$, almost surely, and $N \sim \mathcal{N}(0,\sigma^2)$ is an AWGN independent of X. The capacity (in nats) of this channel is

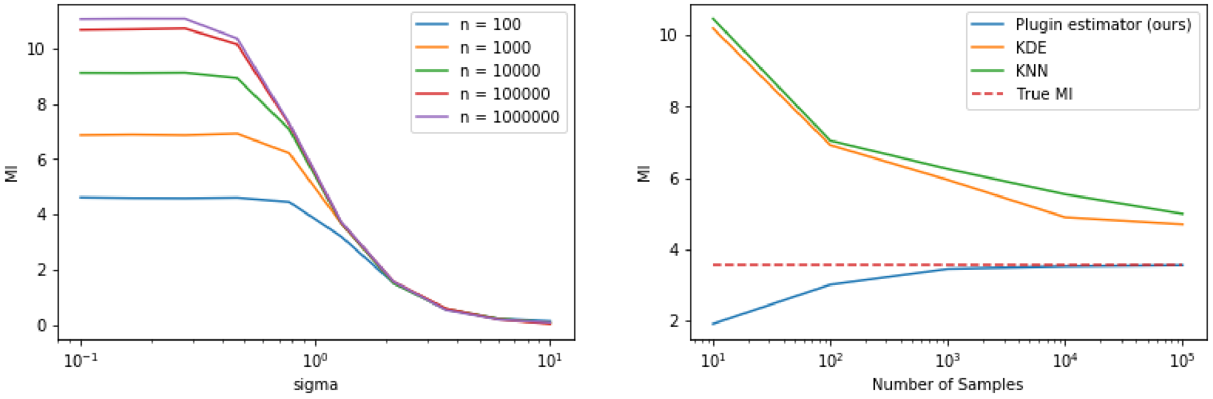
$$\mathsf{C}_{\mathsf{AWGN}}(\sigma) = \max_{X\sim P:\ P\in\mathcal{F}_d} I(X;Y), \tag{46}$$

which is positive for any $\sigma < \infty$. The positivity of capacity implies the following [44]: for any rate $0 < R < \mathsf{C}_{\mathsf{AWGN}}(\sigma)$, there exists a sequence of block codes (with blocklength d) of that rate, with an exponentially decaying (in d) maximal probability of error. More precisely, for any $\epsilon \in \big(0, \mathsf{C}_{\mathsf{AWGN}}(\sigma)\big)$, there exists a codebook $\mathcal{C}_d \subset [-1,1]^d$ of size $|\mathcal{C}_d| \doteq e^{d(\mathsf{C}_{\mathsf{AWGN}}(\sigma)-\epsilon)}$ and a decoding function $\psi_d : \mathbb{R}^d \to [-1,1]^d$ such that

$$\mathbb{P}\Big(\psi_d(Y^d) = c\Big| X^d = c\Big) \geq 1 - e^{-\epsilon^2 d}, \quad \forall c \in \mathcal{C}_d, \tag{47}$$

where $X^d \triangleq (X_1, X_2, \ldots, X_d)$ and $Y^d \triangleq (Y_1, Y_2, \ldots, Y_d)$ are the channel input and output sequences, respectively. The sign $\doteq$ stands for equality in the exponential scale, i.e., $a_k \doteq b_k$ means that $\lim_{k\to\infty} \frac{1}{k}\log\frac{a_k}{b_k} = 0$.

Since (47) ensures an exponentially decaying error probability for any $c \in \mathcal{C}_d$, we also have that the error probability



(a)

(b)

Fig. 6: Estimating $I(S; S+Z)$, where S comes from a BPSK modulated Reed-Muller and $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$: (a) Estimated $I(S; S+Z)$ as a function of σ, for different n values, for the $\mathsf{RM}(4,4)$ code. (b) Plug-in, KDE and KL $I(S; S+Z)$ estimates for the $\mathsf{RM}(5,5)$ code and $\sigma = 2$ as a function of n. Shown for comparison are the curves for the kNN and KDE estimators based on noisy samples of $S+Z$ as well as the true value (dashed).

induced by a randomly selected codeword is exponentially small. Namely, let X^d be a discrete random variable with any distribution P over the codebook $\mathcal{C}_d$. We have

$$\mathbb{P}\big(X^d \neq \psi_d(Y^d)\big) = \sum_{c \in \mathcal{C}_d} P(c) \mathbb{P}\Big(\psi_d(c + N^d) \neq c \Big| X^d = c\Big) \\ \leq e^{-\epsilon^2 d}. \tag{48}$$

Based on (48), Fano's inequality implies

$$H\left(X^d \middle| \psi_d(Y^d)\right) \leq H_b\left(e^{-\epsilon^2 d}\right) + e^{-\epsilon^2 d} \log|\mathcal{C}_d| \triangleq \delta^{(1)}_{\sigma,d}, \tag{49}$$

where $H_b(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$, for $\alpha \in [0,1]$, is the binary entropy function. Although not explicit in our notation, the dependence of $\delta^{(1)}_{\sigma,d}$ on σ is through ϵ. Note that $\lim_{d\to\infty} \delta^{(1)}_{\sigma,d} = 0$, for all $\sigma > 0$, because $\log|\mathcal{C}_d|$ grows only linearly with d and $\lim_{q\to 0} H_b(q) = 0$.

This further gives

$$\begin{aligned} I\left(X^d; Y^d\right) &= H\left(X^d\right) - H\left(X^d \middle| Y^d\right) \\ &\overset{(a)}{\geq} H\left(X^d\right) - H\left(X^d \middle| \psi_d(Y^d)\right) \\ &\overset{(b)}{\geq} H\left(X^d\right) - \delta^{(1)}_{\sigma,d}, \end{aligned} \tag{50}$$

where (a) follows because $H(A|B) \leq H\big(A\big|f(B)\big)$ for any pair of random variables (A,B) and any deterministic function f, while (b) uses (49).

Non-negativity of discrete entropy also implies $I(X^d; Y^d) \leq H(X^d)$, which means that $H(X^d)$ and $I(X^d; Y^d)$ become arbitrarily close as d grows:

$$\Big| H(X^d) - I(X^d; Y^d) \Big| \leq \delta^{(1)}_{\sigma,d}. \tag{51}$$

This means that any good estimator (within an additive gap) of $H(X^d)$ over the class of distributions $\big\{P \big| \operatorname{supp}(P) = \mathcal{C}_d\big\} \subseteq \mathcal{F}_d$ is also a good estimator of the mutual information. Using the well-known lower bound on the sample complexity of discrete entropy estimation in the large alphabet regime (see, e.g., [6, Corollary 10] or [7, Proposition 3]), we have that estimating $H(X^d)$ within a small additive gap $\eta > 0$ requires at least

$$\Omega\left(\frac{|\mathcal{C}_d|}{\eta \log|\mathcal{C}_d|}\right) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right), \tag{52}$$

where $\gamma(\sigma) \triangleq \mathsf{C}_{\mathsf{AWGN}}(\sigma) - \epsilon > 0$ is independent of d.

We relate the above back to the considered differential estimation setup by noting that

$$\begin{aligned} I(X^d; Y^d) &= h(X^d + N^d) - h(N^d) \\ &= h(X^d + N^d) - \frac{d}{2}\log_2(2\pi e \sigma^2). \end{aligned} \tag{53}$$

Letting $S \sim P$ and noting that $Z \overset{\mathcal{D}}{=} N^d$, where $\overset{\mathcal{D}}{=}$ denotes equality in distribution, we have $h(X^d + N^d) = h(S+Z)$. Assuming in contradiction that there exists an estimator of $h(S+Z)$ that uses $o\big(2^{\gamma(\sigma)d}/(\eta d)\big)$ samples and achieves an additive gap $\eta > 0$ over $\big\{P \big| \operatorname{supp}(P) = \mathcal{C}_d\big\}$, implies that $H(X^d)$ can be estimated from these samples within gap $\eta + \delta^{(1)}_{\sigma,d}$. This follows from (51) by taking the estimator of $h(S+Z)$ and subtracting the constant $\frac{d}{2}\log_2(2\pi e \sigma^2)$. We arrive at a contradiction.

2) *Part 2:* Fix $d \geq 1$ and consider a d-dimensional AWGN channel $Y = X + N$, with input X and noise $N \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. Let $\mathcal{C} = \{-1, 1\}^d$ and consider the set of all (discrete) distributions P with $\operatorname{supp}(P) = \mathcal{C}$. For $X \sim P$, with P being an arbitrary distribution from the aforementioned set, and any mapping $\psi_{\mathcal{C}} : \mathbb{R}^d \to \mathcal{C}$, Fano's inequality gives

$$H(X|Y) \leq H\big(X\big|\psi_{\mathcal{C}}(Y)\big) \leq H_b\big(\mathsf{P}_{\mathsf{e}}(\mathcal{C})\big) + \mathsf{P}_{\mathsf{e}}(\mathcal{C}) \cdot \log|\mathcal{C}|, \tag{54}$$

where $\mathsf{P}_{\mathsf{e}}(\mathcal{C}) \triangleq \mathbb{P}\Big(\psi_{\mathcal{C}}(Y) \neq X\Big)$ is the error probability. We choose $\psi_{\mathcal{C}}$ as the maximum likelihood decoder: upon observing $y \in \mathbb{R}^d$ it returns the closest point in $\mathcal{C}$ to y. Namely, $\psi_{\mathcal{C}}$ returns $c \in \mathcal{C}$ if and only if y falls inside the unique orthant that contains c. We have:

$$\mathsf{P}_{\mathsf{e}}(\mathcal{C}) = \sum_{c \in \mathcal{C}} P(c) \mathbb{P}\Big(\psi_{\mathcal{C}}(c + Z) \neq c \Big| X = c\Big)$$

$$= 1 - \left(1 - Q\left(\frac{1}{\sigma}\right)\right)^d \triangleq \epsilon_{\sigma,d}, \tag{55}$$

where Q is the Q-function. Together, (54) and (55) give $H(X|Y) \le H_b(\epsilon_{\sigma,d}) + \epsilon_{\sigma,d} d \log 2 \triangleq \delta^{(2)}_{\sigma,d}$. Note that for any $d \ge 1$, $\lim_{\sigma\to 0} \delta^{(2)}_{\sigma,d} = 0$ exponentially fast in $\frac{1}{\sigma^2}$ (this follows from the large x approximation of $Q(x)$). Similarly to (51), the above implies that

$$\Big| H(X) - I(X;Y) \Big| \le \delta^{(2)}_{\sigma,d}. \tag{56}$$

Thus, any good estimator (within an additive gap η) of $H(X)$ within the class of X distributions P with $\mathrm{supp}(P) = \mathcal{C}$, can be used to estimate $I(X;Y)$ within an $\eta + \delta^{(2)}_{\sigma,d}$ gap.

Now, for σ small enough $\epsilon_{\sigma,d}$, and consequently $\delta^{(2)}_{\sigma,d}$ are arbitrarily close to zero. Hence we may again use lower bounds on the sample complexity of discrete entropy estimation. Like in the proof of Theorem 1, setting $S \sim P$, any estimator of $h(S+Z)$ within a small gap η produces an estimator of $H(X)$ (through $H(X) = h(S+Z) - \frac{d}{2}\log(2\pi e\sigma^2)$ and (56)) within an $\eta + \delta^{(2)}_{\sigma,d}$ gap. Therefore, for sufficiently small $\sigma > 0$ and $\eta > 0$, any estimator of $h(S+Z)$ within a gap of η requires at least

$$\Omega\left(\frac{\mathrm{supp}(P)}{\big(\eta + \delta^{(2)}_{\sigma,d}\big)\log\big(\mathrm{supp}(P)\big)}\right) = \Omega\left(\frac{2^d}{\big(\eta + \delta^{(2)}_{\sigma,d}\big)d}\right) \tag{57}$$

samples. This concludes the proof.

B. Proof of Theorem 2

We start with the following lemma.

Lemma 2 *Let $U \sim P_U$ and $V \sim P_V$ be continuous random variables with densities p_U and p_V, respectively. If $\big|h(U)\big|, \big|h(V)\big| < \infty$, then*

$$\big|h(U) - h(V)\big| \le \max\left\{ \left|\mathbb{E}\log\frac{p_V(V)}{p_V(U)}\right|, \left|\mathbb{E}\log\frac{p_U(U)}{p_U(V)}\right| \right\}.$$

Proof: Recall the identity

$$\begin{aligned} h(U) - h(V) &\le h(U) - h(V) + D(P_U\|P_V) \\ &= \mathbb{E}\log\frac{p_V(V)}{p_V(U)} \le \left|\mathbb{E}\log\frac{p_V(V)}{p_V(U)}\right|. \end{aligned}$$

Reversing the roles of U and V in the above derivation establishes the second bound and completes the proof. ■

Recall now the variational characterization of the χ^2-divergence:

$$\chi^2(\mu\|\nu) = \sup_{g:\ \mathsf{var}_\nu(g)\le 1} \big|\mathbb{E}_\mu g - \mathbb{E}_\nu g\big|^2. \tag{58}$$

Combining this with Lemma 2, we obtain

$$\begin{aligned} &\big|h(U) - h(V)\big| \\ &\quad \le \max\Big\{ \sqrt{\mathsf{var}_{P_V}\big(\log p_V(V)\big)\chi^2(P_U\|P_V)}, \\ &\qquad\qquad \sqrt{\mathsf{var}_{P_V}\big(\log p_U(V)\big)\chi^2(P_U\|P_V)} \Big\}. \end{aligned} \tag{59}$$

Setting $P_V = P * \mathcal{N}_\sigma$ and $P_U = \hat{P}_{S^n} * \mathcal{N}_\sigma$, the next lemma is useful in controlling the variance terms. To state it recall that q and r_{S^n} are the PDFs of $P * \mathcal{N}_\sigma$ and $\hat{P}_{S^n} * \mathcal{N}_\sigma$, respectively, and set $\tilde{q} \triangleq \frac{q}{c_1}$ and $\tilde{r}_{S^n} \triangleq \frac{r}{c_1}$ for $c_1 = (2\pi\sigma^2)^{-d/2}$.

Lemma 3 *Let $S \sim P$. For all $z \in \mathbb{R}^d$ it holds that*

$$\mathbb{E}_{P^{\otimes n}}\big(\log \tilde{r}_{S^n}(z)\big)^2 \le \frac{1}{4\sigma^4}\mathbb{E}_P\|z - S\|^4 \tag{60a}$$

$$\big(\log \tilde{q}(z)\big)^2 \le \frac{1}{4\sigma^4}\mathbb{E}_P\|z - S\|^4. \tag{60b}$$

Proof: We prove (60a); the proof of (60b) is similar and therefore omitted. The map $x \mapsto (\log x)^2$ is convex on $[0,1]$. For any fixed s^n, let $\hat{S} \sim \hat{P}_{s^n}$. Jensen's inequality gives

$$\begin{aligned} \big(\log \tilde{r}_{s^n}(z)\big)^2 &= \left(\log \mathbb{E}_{\hat{P}_{s^n}} \exp\left(-\frac{\|z - \hat{S}\|^2}{2\sigma^2}\right)\right)^2 \\ &\le \mathbb{E}_{\hat{P}_{s^n}} \frac{\|z - \hat{S}\|^4}{4\sigma^4}. \end{aligned}$$

Taking an outer expectation w.r.t. $S^n \sim P^{\otimes n}$ yields

$$\mathbb{E}_{P^{\otimes n}}\big(\log \tilde{r}_{S^n}(z)\big)^2 \le \mathbb{E}_{P^{\otimes n}}\mathbb{E}_{\hat{P}_{S^n}}\frac{\|z - \hat{S}\|^4}{4\sigma^4} = \frac{\mathbb{E}_P\|z - S\|^4}{4\sigma^4}.$$

■

Let $Y = S' + Z$, where $S' \sim P$ and $Z \sim \mathcal{N}_\sigma$ are independent. Since variance is translation invariant, we get

$$\begin{aligned} \mathsf{var}_{P*\mathcal{N}_\sigma}\big(\log q(Y)\big) &= \mathsf{var}_{P*\mathcal{N}_\sigma}\big(\log \tilde{q}(Y)\big) \\ &\le \frac{1}{4\sigma^4}\mathbb{E}\|Z + S' - S\|^4 \\ &\le \frac{\sigma^2 d(2+d)(2+\sigma^2) + 8d^2}{4\sigma^4}. \end{aligned} \tag{61}$$

When combined with Proposition 3, the above bound takes care of the first term in (59).

For the second term, we apply Cauchy-Schwartz and treat the expected values of $\mathsf{var}_{P_V}\big(\log p_U(V)\big)$ and $\chi^2(P_U\|P_V)$ separately. For the variance, using (60a) and an argument similar to (61) we get the same bound therein. The expected χ^2-square divergence in both arguments of the maximum in (59) is bounded using Corollary 1. Combining the pieces, for any $P \in \mathcal{F}_d$, we obtain

$$\begin{aligned} &\mathbb{E}_{P^{\otimes n}}\big|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{S^n} * \mathcal{N}_\sigma)\big| \\ &\qquad \le 2\sqrt{\frac{\sigma^2 d(2+d)(2+\sigma^2) + 8d^2}{4\sigma^4}}e^{\frac{2d}{\sigma^2}} \cdot \frac{1}{\sqrt{n}}. \end{aligned} \tag{62}$$

Remark 9 *An alternative proof of the parametric estimation rate was given in [33] using the 1-Wasserstein distance instead of χ^2-square. Specifically, one may invoke [45, Proposition 5] to reduce the analysis of $\mathbb{E}_{P^{\otimes n}}\big|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{S^n} * \mathcal{N}_\sigma)\big|$ to that of $\mathbb{E}_{P^{\otimes n}}\mathsf{W}_1\Big(\hat{P}_{S^n} * \mathcal{N}_\sigma, P * \mathcal{N}_\sigma\Big)$. Then, using [31, Theorem 6.15] and the bounded support assumption, the parametric risk convergence rate follows with the constant $\frac{\sqrt{d}\cdot 2^{d+2}}{\min\{1,\sigma^d\}}$.*

C. Proof of Theorem 3

Starting from Lemma 2, we again focus on bounding the maximum of the two expected log ratios. The following lemma allows converting $\left|\mathbb{E}\log\frac{p_V(V)}{p_V(U)}\right|$ and $\left|\mathbb{E}\log\frac{p_U(U)}{p_U(V)}\right|$ into forms that are more convenient to analyze.

Lemma 4 *Let $U \sim P_U$ and $V \sim P_V$ be continuous random variables with PDFs p_U and p_V, respectively. For any measurable function $g : \mathbb{R}^d \to \mathbb{R}$*

$$\big|\mathbb{E}g(U) - \mathbb{E}g(V)\big| \leq \int \big|g(z)\big| \cdot \big|p_U(z) - p_V(z)\big| \, \mathsf{d}z \ .$$

Proof: We couple P_U and P_V via the maximal TV coupling[14]. Specifically, let $(P_U - P_V)_+$ and $(P_U - P_V)_-$ are the positive and negative parts of the signed measure $(P_U - P_V)$. Define $(P_U \wedge P_V) \triangleq P_U - (P_U - P_V)_+$, and let $(\mathrm{Id}, \mathrm{Id})_\sharp (P_U \wedge P_V)$ be the push-forward measure of $P_U \wedge P_V$ through $(\mathrm{Id}, \mathrm{Id}) : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$. Letting $\alpha \triangleq \frac{1}{2}\int |p_U(x) - p_V(x)| dx$ (note that $\int \mathsf{d}(P_U - P_V)_+ = \int \mathsf{d}(P_U - P_V)_- = \alpha$), the maximal TV coupling is give by

$$\pi \triangleq (\mathrm{Id}, \mathrm{Id})_\sharp (P_U \wedge P_V) + \frac{1}{\alpha}(P_U - P_V)_+ \otimes (P_U - P_V)_- . \quad (63)$$

Jensen's inequality implies $\big|\mathbb{E}g(U) - \mathbb{E}g(V)\big| \leq \mathbb{E}_\pi \big|g(U) - g(V)\big|$ and we proceed as

$$\begin{aligned}
&\mathbb{E}_\pi \big|g(U) - g(V)\big| \\
&\leq \frac{1}{\alpha}\int \Bigg(\Big(\big|g(u)\big| + \big|g(v)\big|\Big)\big(p_U(u) - p_V(u)\big)_+ \\
&\qquad\qquad \cdot \big(p_U(v) - p_V(v)\big)_- \Bigg) \mathsf{d}u \, \mathsf{d}v \\
&= \int \big|g(u)\big|\big(p_U(u) - p_V(u)\big)_+ \mathsf{d}u \\
&\qquad + \int \big|g(v)\big|\big(p_U(v) - p_V(v)\big)_- \mathsf{d}v \\
&= \int \big|g(z)\big|\Big(\big(p_U(z) - p_V(z)\big)_+ + \big(p_U(z) - p_V(z)\big)_-\Big) \mathsf{d}z \\
&= \int \big|g(z)\big| \cdot \big|p_U(z) - p_V(z)\big| \, \mathsf{d}z. \qquad (64)
\end{aligned}$$

■

Fix any $P \in \mathcal{F}_{d,K}^{(\mathsf{SG})}$ and assume that $\mathbb{E}_P S = 0$. This assumption comes with no loss of generality since both the target functional $h(P*\mathcal{N}_\sigma)$ and the plug-in estimator are translation invariant. Note that $\big|h(P * \mathcal{N}_\sigma)\big|, \big|h(\hat{P}_{S^n} * \mathcal{N}_\sigma)\big| < \infty$. Combining Lemmata 2 and 4, we a.s. have

$$\begin{aligned}
&\big|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{S^n} * \mathcal{N}_\sigma)\big| \\
&\leq \max\Bigg\{ \int \big|\log \tilde{r}_{S^n}(z)\big| \cdot |q(z) - r_{S^n}(z)| \, \mathsf{d}z, \\
&\qquad\qquad \int \big|\log \tilde{q}(z)\big| \cdot \big|q(z) - r_{S^n}(z)\big| \, \mathsf{d}z \Bigg\}, \qquad (65)
\end{aligned}$$

where, as before, q and r_{S^n} are the PDFs of $P * \mathcal{N}_\sigma$ and $\hat{P}_{S^n} * \mathcal{N}_\sigma$, respectively, while $\tilde{q} \triangleq \frac{q}{c_1}$ and $\tilde{r}_{S^n} \triangleq \frac{r_{S^n}}{c_1}$, for $c_1 = (2\pi\sigma^2)^{-d/2}$.

Recalling that $\mathbb{E}[\max\{|X|, |Y|\}] \leq \mathbb{E}|X| + \mathbb{E}|Y|$, for any random variable X, Y, we now bound $\int \big|\log \tilde{r}_{S^n}(z)\big|\big|p_U(z) - p_V(z)\big| \, \mathsf{d}z$. The bound for the other integral is identical and thus omitted. Let $f_a : \mathbb{R}^d \to \mathbb{R}$ be the PDF of $\mathcal{N}\left(0, \frac{1}{2a}\mathrm{I}_d\right)$, for $a > 0$ specified later. The Cauchy-Schwarz inequality implies

$$\begin{aligned}
&\left(\mathbb{E}_{P^{\otimes n}} \int \big|\log \tilde{r}_{S^n}(z)\big|\big|q(z) - r_{S^n}(z)\big| \, \mathsf{d}z\right)^2 \leq \\
&\int \mathbb{E}_{P^{\otimes n}} \big(\log \tilde{r}_{S^n}(z)\big)^2 f_a(z) \, \mathsf{d}z \cdot \int \mathbb{E}_{P^{\otimes n}} \frac{\big(q(z) - r_{S^n}(z)\big)^2}{f_a(z)} \mathsf{d}z. \qquad (66)
\end{aligned}$$

Using Lemma 3, we bound the first integral as

$$\begin{aligned}
&\int \mathbb{E}_{P^{\otimes n}} \big(\log \tilde{r}_{S^n}(z)\big)^2 f_a(z) \, \mathsf{d}z \\
&\qquad \leq \int \frac{\mathbb{E}\|z - S\|^4}{4\sigma^4} \frac{\exp\big(-a\|z\|^2\big)}{\sqrt{\pi^d a^{-d}}} \mathsf{d}z \\
&\qquad \overset{(a)}{\leq} \frac{2}{\sigma^4}\mathbb{E}\|S\|^4 + \frac{2}{\sigma^4}\int \|z\|^4 \frac{\exp\big(-a\|z\|^2\big)}{\sqrt{\pi^d a^{-d}}} \mathsf{d}z \\
&\qquad \overset{(b)}{\leq} \frac{32K^4 d^2}{\sigma^4} + \frac{1}{2\sigma^4 a^2} d(d+2)
\end{aligned}$$

where (a) follows from the triangle inequality, and (b) uses the K-subgaussianity of S [35, Lemma 5.5]

To bound the second integral, we repeat steps (6)-(7) from the proof of Proposition 1. Specifically, we have $\mathbb{E}_{P^{\otimes n}}\big(q(z) - r_{S^n}(z)\big)^2 \leq \frac{c_1^2}{n}\mathbb{E}e^{-\frac{1}{\sigma^2}\|z-S\|^2}$, because $r_{S^n}(z)$ is a sum of i.i.d. random variables with $\mathbb{E}_{P^{\otimes n}} r_{S^n}(z) = q(z)$. This gives

$$\int \mathbb{E}_{P^{\otimes n}} \frac{\big(q(z) - r_{S^n}(z)\big)^2}{f_a(z)} \mathsf{d}z \leq \frac{c_1}{n 2^{d/2}} \mathbb{E}\frac{1}{f_a(S + Z/\sqrt{2})}, \qquad (67)$$

for independent $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I_d})$ and $S \sim P$. Recalling that $\big(f_a(z)\big)^{-1} = c_2 \exp\big(a\|z\|^2\big)$, for $c_2 \triangleq \left(\frac{\pi}{a}\right)^{\frac{d}{2}}$, the subgaussianity of S and Z implies

$$\begin{aligned}
&\frac{c_1}{n 2^{d/2}} \mathbb{E}\frac{1}{f_a(S + Z/\sqrt{2})} \leq \\
&\frac{c_1 c_2}{n 2^{d/2}} \exp\left(\big(K + \sigma/\sqrt{2}\big)^2 ad + \frac{(K + \sigma/\sqrt{2})^4 a^2 d}{1 - 2(K + \sigma/\sqrt{2})^2 a}\right), \qquad (68)
\end{aligned}$$

where $0 < a < \frac{1}{2(K+\sigma/\sqrt{2})^2}$.

Setting $a = \frac{1}{4(K+\sigma/\sqrt{2})^2}$, we combine (66)-(68) to obtain the result (recalling that the second integral from (65) is bounded exactly as the first). For any $P \in \mathcal{F}_{d,K}^{(\mathsf{SG})}$ we have

$$\begin{aligned}
&\left(\mathbb{E}_{P^{\otimes n}}\big|h(P * \mathcal{N}_\sigma) - h(\hat{P}_{S^n} * \mathcal{N}_\sigma)\big|\right)^2 \leq \\
&\frac{64\big(2d^2K^4 + d(d+2)(K + \sigma/\sqrt{2})^4\big)}{\sigma^4} \left(\left(\frac{1}{\sqrt{2}} + \frac{K}{\sigma}\right) e^{\frac{3}{8}}\right)^d \frac{1}{n} \ . \qquad (69)
\end{aligned}$$

D. Proof of Theorem 4

First note that since $h(q)$ is concave in q and because $\mathbb{E}_{P^{\otimes n}} \hat{P}_{S^n} = P$, we have

$$\mathbb{E}_{P^{\otimes n}} h(\hat{P}_{S^n} * \varphi_\sigma) \leq h(P * \mathcal{N}_\sigma), \qquad (70)$$

for all $P \in \mathcal{F}_d$. Now, let $W \sim \mathsf{Unif}([n])$ be independent of (S^n, Z) and define $Y = S_W + Z$. We have the following

[14] This coupling attains maximal probability of the event $\{U = V\}$.

lemma, whose proof is found in Appendix C.

Lemma 5 *For any $P \in \mathcal{F}_d$, we have*

$$h(P*\mathcal{N}_\sigma) - \mathbb{E}_{P^{\otimes n}} h(\hat{P}_{S^n}*\mathcal{N}_\sigma) = I(S^n;Y). \tag{71}$$

Using the lemma, we have

$$\sup_{P\in\mathcal{F}_d}\left|h(P*\mathcal{N}_\sigma) - \mathbb{E}_{P^{\otimes n}} h(P*\mathcal{N}_\sigma)\right| = \sup_{P\in\mathcal{F}_d} I(S^n;Y), \tag{72}$$

where the right hand side is the mutual information between n i.i.d. random samples S_i from P and the random vector $Y = S_W + Z$, formed by choosing one of the S_i's at random and adding Gaussian noise.

To obtain a lower bound on the supremum, we consider the following P. Partition the hypercube $[-1,1]^d$ into k^d equal-sized smaller hypercubes, each of side length k. Denote these smaller hypercubes as $\mathsf{C}_1, \mathsf{C}_2, \ldots, \mathsf{C}_{k^d}$ (the order does not matter). For each $i \in [k^d]$ let $c_i \in \mathsf{C}_i$ be the centroid of the hypercube C_i. Let $\mathcal{C} \triangleq \{c_i\}_{i=1}^{k^d}$ and choose P as the uniform distribution over $\mathcal{C}$.

By the mutual information chain rule and the non-negativity of discrete entropy, we have

$$\begin{aligned} I(S^n;Y) &= I(S^n;Y,S_W) - I(S^n;S_W|Y) \\ &\overset{(a)}{\geq} I(S^n;S_W) - H(S_W|Y) \\ &= H(S_W) - H(S_W|S^n) - H(S_W|Y), \end{aligned} \tag{73}$$

where step (a) uses the independence of (S^n, W) and Z. Clearly $H(S_W) = \log|\mathcal{C}|$, while $H(S_W|S^n) \leq H(S_W, W|S^n) \leq H(W) = \log n$, via the independence of W and S^n. For the last (subtracted) term in (73) we use Fano's inequality to obtain

$$H(S_W|Y) \leq H\big(S_W\big|\psi_{\mathcal{C}}(Y)\big) \leq H_b\big(\mathsf{P}_\mathsf{e}(\mathcal{C})\big) + \mathsf{P}_\mathsf{e}(\mathcal{C}) \cdot \log|\mathcal{C}|, \tag{74}$$

where $\psi_{\mathcal{C}} : \mathbb{R}^d \to \mathcal{C}$ is a function for decoding S_W from Y and $\mathsf{P}_\mathsf{e}(\mathcal{C}) \triangleq \mathbb{P}\big(S_W \neq \psi_{\mathcal{C}}(Y)\big)$ is the probability that $\psi_{\mathcal{C}}$ commits an error.

Fano's inequality holds for any decoding function $\psi_{\mathcal{C}}$. We choose $\psi_{\mathcal{C}}$ as the maximum likelihood decoder, i.e., upon observing a $y \in \mathbb{R}^d$ it returns the closest point to y in $\mathcal{C}$. Denote by $\mathcal{D}_i \triangleq \psi_{\mathcal{C}}^{-1}(c_i)$ the decoding region on c_i, i.e., the region $\left\{y \in \mathbb{R}^d \middle| \psi_{\mathcal{C}}(y) = c_i\right\}$ that $\psi_{\mathcal{C}}$ maps to c_i. Note that $\mathcal{D}_i = \mathsf{C}_i$ for all $i \in [k^d]$ for which C_i doesn't intersect with the boundary of $[-1,1]^d$. The probability of error for the decoder $\psi_{\mathcal{C}}$ is bounded as:

$$\begin{aligned} \mathsf{P}_\mathsf{e}(\mathcal{C}) &= \frac{1}{k^d}\sum_{i=1}^{k^d} \mathbb{P}\Big(\psi_{\mathcal{C}}(c_i+Z) \neq c_i \Big| S_W = c_i\Big) \\ &= \frac{1}{k^d}\sum_{i=1}^{k^d} \mathbb{P}\big(c_i + Z \notin \mathcal{D}_i\big) \\ &\overset{(a)}{\leq} \mathbb{P}\left(\|Z\|_\infty > \frac{2/k}{2}\right) \\ &\overset{(b)}{=} 1 - \left(1 - 2Q\left(\frac{1}{k\sigma}\right)\right)^d, \end{aligned} \tag{75}$$

where (a) holds since the C_i have sides of length $2/k$ and the error probability is largest for $i \in [k^d]$ such that C_i is in the interior of $[-1,1]^d$. Step (b) follows from independence and the definition of the Q-function.

Taking $k = k_\star$ in (75) as given in the statement of the theorem gives the desired bound $\mathsf{P}_\mathsf{e}(\mathcal{C}) \leq \epsilon$. Collecting the pieces and inserting back to (73), we obtain

$$I(S^n;Y) \geq \log\left(\frac{k_\star^{d(1-\epsilon)}}{n}\right) - H_b(\epsilon). \tag{76}$$

Together with (72) this concludes the proof.

E. Proof of Theorem 5

Denote the joint distribution of (C, Z, V) by $P_{C,Z,V}$. Marginal or conditional distributions are denoted as usual by keeping only the relevant subscripts. Lowercase p denotes a probability mass function (PMF) or a PDF depending on whether the random variable in the subscript is discrete or continuous. In particular, p_C is the PMF of C, $p_{C|V}$ is the conditional PMF of C given V, while $p_Z = \varphi_\sigma$ and $p_V = g$ are the PDFs of Z and V, respectively.

First observe that the estimator is unbiased:

$$\mathbb{E}\hat{h}_{\mathsf{MC}} = -\frac{1}{n \cdot n_{\mathsf{MC}}}\sum_{i=1}^{n}\sum_{j=1}^{n_{\mathsf{MC}}} \mathbb{E}\log g\left(\mu_i + Z_j^{(i)}\right) = h(g). \tag{77}$$

Therefore, the MSE expands as

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) = \frac{1}{n^2 \cdot n_{\mathsf{MC}}}\sum_{i=1}^{n} \mathsf{var}\Big(\log g(\mu_i + Z)\Big). \tag{78}$$

We next bound the variance of $\log g(\mu_i + Z)$ via the Gaussian Poincaré inequality (with Poincaré constant σ^2). For each $i \in [n]$, we have

$$\mathsf{var}\Big(\log g(\mu_i + Z)\Big) \leq \sigma^2\mathbb{E}\Big[\big\|\nabla \log g(\mu_i + Z)\big\|^2\Big]. \tag{79}$$

We proceed with separate derivations of (38) and (39).

1) MSE for Bounded Support: Since $\|C\|_2 \leq \sqrt{d}$ almost surely, Proposition 3 from [45] implies

$$\big\|\nabla \log g(v)\big\|_2 \leq \frac{\|v\| + \sqrt{d}}{\sigma^2}. \tag{80}$$

Inserting this into the Poincaré inequality and using $(a+b)^2 \leq 2a^2 + 2b^2$ we have,

$$\mathsf{var}\Big(\log g(\mu_i + Z)\Big) \leq \frac{2d(4+\sigma^2)}{\sigma^2}, \tag{81}$$

for each $i \in [n]$. Together with (78), this produces (38).

2) MSE for Bounded Second Moment: To prove (39), we use Proposition 2 from [45] to obtain

$$\big\|\nabla \log g(v)\big\| \leq \frac{1}{\sigma^2}\big(3\|v\| + 4\mathbb{E}\|C\|\big). \tag{82}$$

Using (79), the variance is bounded as

$$\mathsf{var}\Big(\log g(\mu_i + Z)\Big) \leq \frac{1}{\sigma^2}\mathbb{E}\Big[\big(3\|\mu_i + Z\| + 4\mathbb{E}\|C\|\big)^2\Big] \leq$$

$$\frac{1}{\sigma^2}\Big(9d\sigma^2+16m+24\sigma\sqrt{dm}+3\|\mu_i\|\left(3+9\sigma\sqrt{d}+8\sigma\sqrt{dm}\right)\Big), \tag{83}$$

where the last step uses Hölder's inequality $\mathbb{E}\|C\| \leq \sqrt{\mathbb{E}\|C\|^2}$. The proof of (39) is concluded by plugging (83) into the MSE expression from (78) and noting that $\frac{1}{n}\sum_{i=1}^{n}\|\mu_i\| \leq \sqrt{m}$.

F. Proof of Proposition 9

Fix P_X, define $g(x) \triangleq h(T|X=x) = h(P_{S|X=x} * \mathcal{N}_\sigma)$ and write

$$I(X;T) = h(T) - h(T|X) = h(P_S * \mathcal{N}_\sigma) - \mathbb{E}g(X). \tag{84}$$

Applying the triangle inequality to (43) we obtain

$$\begin{aligned}
&\mathbb{E}\left|I(X;T) - \hat{I}_{\mathsf{Input}}\left(X^n,\hat{h},\sigma\right)\right| \\
&\quad\leq \mathbb{E}\left|\hat{h}(S^n,\sigma) - h(P_S * \mathcal{N}_\sigma)\right| \\
&\qquad + \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{h}\big(S^n(X_i),\sigma\big) - \mathbb{E}g(X)\right| \\
&\quad\leq \underbrace{\mathbb{E}\left|\hat{h}(S^n,\sigma) - h(P_S * \mathcal{N}_\sigma)\right|}_{\text{(I)}} \\
&\qquad + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left|\hat{h}\big(S^n(X_i),\sigma\big) - g(X_i)\right|}_{\text{(II)}} \\
&\qquad + \underbrace{\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}g(X_i) - \mathbb{E}g(X)\right|}_{\text{(III)}}
\end{aligned} \tag{85}$$

By assumption (42) and because $P_S \in \mathcal{F}_d$, we have

$$\mathbb{E}\left|\hat{h}(S^n,\sigma) - h(P_S * \mathcal{N}_\sigma)\right| \leq \Delta_{\sigma,d}(n). \tag{86}$$

Similarly, for any fixed $X^n = x^n$, $P_{S|X=x_i} \in \mathcal{F}_d$, for all $i \in [n]$, and hence

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{h}(S^n(X_i),\sigma) - g(X_i)\right|\middle|X^n = x^n\right] \\
&\quad\overset{(a)}{=} \mathbb{E}\left|\hat{h}\big(S^n(x_i),\sigma\big) - h(P_{S|X=x_i} * \mathcal{N}_\sigma)\right| \\
&\quad\leq \Delta_{\sigma,d}(n),
\end{aligned} \tag{87}$$

where (a) is because for a fixed x_i, sampling from $P_{S|X=x_i}$ corresponds to drawing multiple noise realization for the previous layers of the DNN. Since these noises are independent of X, we may remove the conditioning from the expectation. Taking an expectation on both sides of (87) and applying the law of total expectation, we have

$$\text{(II)} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left|\hat{h}\big(S^n(x_i),\sigma\big) - g(X_i)\right| \leq \Delta_{\sigma,d}(n). \tag{88}$$

Turning to term (III), observe that $\big\{g(X_i)\big\}_{i=1}^{n}$ are i.i.d random variables. Hence

$$\frac{1}{n}\sum_{i=1}^{n}g(X_i) - \mathbb{E}g(X) \tag{89}$$

is the difference between an empirical average and the expectation. By monotonicity of moments we have

$$\begin{aligned}
\text{(III)}^2 &= \left(\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}g(X_i) - \mathbb{E}g(X)\right|\right)^2 \\
&\leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}g(X_i) - \mathbb{E}g(X)\right)^2\right] \\
&= \frac{1}{n}\mathsf{var}\big(g(X)\big) \\
&\leq \frac{1}{4n}\left(\sup_x h(P_{T|X=x}) - \inf_x h(P_{T|X=x})\right)^2.
\end{aligned} \tag{90}$$

The last inequality follows since $\mathsf{var}(A) \leq \frac{1}{4}(\sup A - \inf A)^2$ for any random variable A.

It remains to bound the supremum and infimum of $h(P_{T|X=x})$ uniformly in $x \in \mathbb{R}^{d_0}$. By definition $T = S + Z$, where S and Z are independent and $Z \sim \mathcal{N}(0,\sigma^2\mathrm{I}_d)$. Therefore, for all $x \in \mathbb{R}^{d_0}$

$$\begin{aligned}
h(P_{T|X=x}) &= h(S+Z|X=x) \\
&\geq h(S+Z|S,X=x) \\
&= h(Z) \\
&= \frac{d}{2}\log(2\pi e\sigma^2),
\end{aligned} \tag{91}$$

where we have used the independence of Z and (S,X) and the fact that conditioning cannot increase entropy. On the other hand, denoting the entries of T by $T \triangleq \big(T(k)\big)_{k=1}^{d}$, we can obtain an upper bound as

$$h(P_{T|X=x}) = h(T|X=x) \leq \sum_{k=1}^{d}h\big(T(k)\big|X=x\big), \tag{92}$$

since independent random variables maximize differential entropy. Now for any $k \in [d]$, we have

$$\mathsf{var}\big(T(k)\big|X=x\big) \leq \mathbb{E}\big[T^2(k)\big|X=x\big] \leq 1+\sigma^2, \tag{93}$$

because $S(k) \in [-1,1]$ almost surely. Since the Gaussian distribution maximizes differential entropy under a variance constraint, we have

$$h(P_{T|X=x}) \leq \frac{d}{2}\log\big(2\pi e(1+\sigma^2)\big). \tag{94}$$

for all $x \in \mathbb{R}^{d_0}$. Substituting the lower bound (91) and upper bound (94) into (90) gives

$$\text{(III)}^2 \leq \left(\frac{d\log\left(1+\frac{1}{\sigma^2}\right)}{4\sqrt{n}}\right)^2. \tag{95}$$

Inserting this along with (86) and (88) into the bound (85) bounds the expected estimation error as

$$\mathbb{E}\left|\hat{I}_{\mathsf{Input}}\left(X^n,\hat{h},\sigma\right)-I(X;T)\right|\leq 2\Delta_n+\frac{d\log\left(1+\frac{1}{\sigma^2}\right)}{4\sqrt{n}}. \tag{96}$$

Taking the supremum over P_X concludes the proof.

VII. Summary and Concluding Remarks

This work first explored the problem of empirical approximation under Gaussian smoothing in high dimensions. To quantify the approximation error, we considered various statistical distances, such as 1-Wasserstein, squared 2-Wasserstein, TV, KL divergence and χ^2-divergence. It was shown that when P is subgaussian, the 1-Wasserstein and the TV distances converge as $n^{-1/2}$. The parametric convergence rate is also attained by the KL divergence, squared 2-Wasserstein distance and χ^2-divergence, so long that the χ^2 mutual information $I_{\chi^2}(S;Y)$, for $Y=S+Z$ with $S\sim P$ independent of $Z\sim\mathcal{N}_\sigma$, is finite. The latter condition is always satisfied by K-subgaussian P distributions in the low SNR regime where $K<\frac{\sigma}{2}$. However, when SNR is high ($K>\sqrt{2}\sigma$), there exist K-subgaussian distributions P for which $I_{\chi^2}(S;Y)=\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n}*\mathcal{N}_\sigma\middle\|P*\mathcal{N}_\sigma\right)=\infty$. Whenever this happens, it was further established that the KL divergence and the squared 2-Wasserstein distance are $\omega(n^{-1})$. Whenever the parametric convergence rate of the smooth empirical measure is attained, it strikingly contrasts classical (unconvolved) results, e.g., for the Wasserstein distance, which suffer from the curse of dimensionality (see, e.g., [3, Theorem 1]).

The empirical approximation results were used to study differential entropy estimation under Gaussian smoothing. Specifically, we considered the estimation of $\mathsf{T}_\sigma(P)=h(P*\mathcal{N}_\sigma)$ based on i.i.d. samples from P and knowledge of the noise distribution $\mathcal{N}_\sigma$. It was shown that the absolute-error risk of the plug-in estimator over the bounded support and subgaussian classes converges as $e^{O(d)}n^{-1/2}$ (with the prefactor explicitly characterized). This established the plug-in estimator as minimax-rate optimal. The exponential dependence of the sample complexity on dimension was shown to be necessary. These results were followed by a bias lower bound of order $\log\left(2^d n^{-1}\right)$, as well as an efficient and provably accurate MC integration method for computing the plug-in estimator.

The considered differential entropy estimation framework enables studying information flows in DNNs [28]. In Section IV we showed how the mutual information between layers of a DNN reduces to estimating $h(P*\mathcal{N}_\sigma)$. An ad hoc estimator for $h(P*\mathcal{N}_\sigma)$ was important here because the general-purpose estimators (based on noisy samples from $P*\mathcal{N}_\sigma$) available in the literature are unsatisfactory for several (theoretical and/or practical) reasons. Most theoretical performance guarantees for such estimators are not valid in our setup, as they typically assume that the unknown density is positively lower bounded inside its compact support.

To the best of our knowledge, the only two works that provide convergence results that apply here are [5] and [19]. The rate derived for the KDE-based estimator in [5], however, effectively scales as $n^{-1/d}$ for large dimensions, which is too slow for practical purposes. Remarkably, [19] proposes a wKL estimator in the very smooth density regime that provably attains the parametric rate of estimation in our problem (e.g., when P is compactly supported). This result, however, does not characterize the dependence of that rate on d. Understanding this dependence is crucial in practice. Indeed, in Section V we show that, empirically, the performance of the wKL significantly deteriorates as d grows. In all our experiments, the plug-in estimator outperforms the wKL method from [19] (as well as all other generic estimator we have tested), converging faster with n and scaling better with d.

For future work, open questions regarding the smooth empirical measure convergence were listed in Section II-E. On top of that, there are appealing extensions of the differential estimation question to be considered. This includes non-Gaussian additive noise models or multiplicative Bernoulli noise (which corresponds to DNNs with dropout regularization). The question of estimating $h(P*\mathcal{N}_\sigma)$ when only samples from $P*\mathcal{N}_\sigma$ are available (yet the Gaussian convolution structure is known) is also attractive. This would, however, require a different technique to that employed herein. Our current method strongly relies on having 'clean' samples from P. Beyond this work, we see considerable virtue in exploring additional ad hoc estimation setups with exploitable structure that might enable improved estimation results.

Acknowledgement

The authors thank Yihong Wu for helpful discussions.

Appendix A
Proof of Proposition 8

We lower bound the minimax risk in the nonparametric estimation risk by a reduction to a parametric setup. Without loss of generality, assume $d=1$ (the risk of the one-dimensional estimation problem trivially lower bounds that of its d-dimensional counterpart). Recall that $\mathcal{F}_{1,K}^{(\mathsf{SG})}$ is the class of K-subgaussian measures on $\mathbb{R}$. Define $\mathcal{G}_K:=\{\mathcal{N}_\nu\}_{\nu\in[K/2,K]}$ as the collection of all centered Gaussian measures, each with variance ν^2. Noting that $\mathcal{G}\subset\mathcal{F}_{1,K}^{(\mathsf{SG})}$, we obtain

$$\mathcal{R}^\star\left(n,\sigma,\mathcal{F}_{1,K}^{(\mathsf{SG})}\right)\geq\inf_{\hat{h}}\sup_{P\in\mathcal{G}}\mathbb{E}\left|h(P*\mathcal{N}_\sigma)-\hat{h}(S^n,\sigma)\right|, \tag{97}$$

and henceforth focus on lower bounding the RHS.

Note that $h(P*\mathcal{N}_\sigma)=\frac{1}{2}\log(\nu^2+\sigma^2)+\frac{1}{2}\log(2\pi e)$, for $P=\mathcal{N}_\nu\in\mathcal{G}$. Thus, the estimation of $h(P*\mathcal{N}_\sigma)$, when $P\in\mathcal{G}$, reduces to estimating $\frac{1}{2}\log(\nu^2+\sigma^2)$ from samples of $\mathcal{N}_\nu$. Recall that $\nu\in[K/2,K]$ is considered unknown and σ known. This simple setting lands within the framework studied in [42], where lower bounds on the minimax absolute error in terms of the associated Hellinger modulus were derived. We follow the proof style of Corollary 3 therein.

Firstly, recall that the squared Hellinger distance between $\mathcal{N}(0,\nu_1^2)$ and $\mathcal{N}(0,\nu_0^2)$ is given by

$$\rho^2(\nu_1,\nu_0)=2\left(1-\sqrt{\frac{2\nu_1\nu_0}{\nu_1^2+\nu_0^2}}\right). \tag{98}$$

Since we are estimating $\frac{1}{2}\log(\nu^2+\sigma^2)$, for convenience we denote the parameter of interest as $\theta(\nu) := \frac{1}{2}\log(\nu^2+\sigma^2)$. The Hellinger modulus of $\theta(\nu_0)$ is defined as

$$\omega_q\left(\sqrt{\frac{\alpha}{n}};\theta(\nu_0)\right) = \sup_{\nu:\ \rho(\nu,\nu_0)\leq\sqrt{\frac{\alpha}{n}}} \left|\theta(\nu)-\theta(\nu_0)\right|. \qquad (99)$$

Based on Theorem 2 of [42], if $\omega_q\left(\sqrt{\frac{\alpha}{n}};\theta(\nu_0)\right)=\Omega(n^{-1/2})$ then the RHS of (97) is also $\Omega(n^{-1/2})$. We thus seek to bound this modulus.

We start by characterizing the set of ν values that lie in the Hellinger ball of radius $\sqrt{\frac{\alpha}{n}}$ around ν_0. From (98) and by defining $\xi=\left(1-\frac{\alpha}{2n}\right)^{-2}$, one readily verifies that $\rho(\nu,\nu_0)\leq\sqrt{\frac{\alpha}{n}}$ if and only if

$$\xi-\sqrt{\xi^2-1}\leq\frac{\nu}{\nu_0}\leq\xi+\sqrt{\xi^2-1}.$$

Equivalently, the feasible set of ν values satisfies

$$\begin{aligned}&\left|\log(\nu)-\log(\nu_0)\right|\\&\qquad\in\left[\log\left(\xi-\sqrt{\xi^2-1}\right),\log\left(\xi+\sqrt{\xi^2-1}\right)\right].\end{aligned}$$

One may check that $2\sqrt{\frac{\alpha}{2n}}$, for all positive $\alpha,n>0$ with $\alpha/n<1$, belongs to the above interval. Hence, $\nu_\star$ such that $\log(\nu_\star)=\log(\nu_0)+2\sqrt{\frac{a}{2n}}$ is feasible and we may substitute it into (99) to lower bound the modulus as follows:

$$\begin{aligned}\omega_q\left(\sqrt{\frac{a}{n}};\theta(\nu_0)\right)&\geq\left|\theta(\nu_\star)-\theta(\nu_0)\right|\\&=\frac{1}{2}\log(\nu_\star^2+\sigma^2)-\frac{1}{2}\log(\nu_0^2+\sigma^2)\\&\overset{(a)}{\geq}\frac{1}{1+\frac{\sigma^2}{\nu_0^2}}\log\left(\frac{\nu_\star}{\nu_0}\right)\\&\overset{(b)}{\geq}\frac{1}{1+\frac{4\sigma^2}{K^2}}\log\left(\frac{\nu_\star}{\nu_0}\right)\\&=2\frac{K^2}{K^2+4\sigma^2}\sqrt{\frac{\alpha}{2n}},\end{aligned}$$

where (a) is because for any $a,b,c>0$ with $b\geq a$ we have

$$\begin{aligned}\log(b+c)-\log(a+c)&=\int_a^b\frac{1}{x+c}dx\\&\geq\int_a^b\frac{1}{(1+c/a)x}dx\\&=\frac{1}{1+\frac{c}{a}}\log\frac{b}{a},\end{aligned}$$

and (b) follows since $\nu_0\geq K/2$.

Applying Theorem 2 of [42] to this bound on the Hellinger modulus implies that the best estimator of $h(P*\mathcal{N}_\sigma)$ over the class $\mathcal{G}_K$ achieves $\Omega(1/\sqrt{n})$ in absolute error.

Appendix B
Label and Hidden Layer Mutual Information

Consider the estimation of $I(Y;T)$, where Y is the true label and T is a hidden layer in a noisy DNN. For completeness, we first describe the setup (repeating some parts of Remark 6). Afterwards, the proposed estimator for $I(Y;T)$ is presented and an upper bound on the estimation error is stated and proven.

Let $(X,Y)\sim P_{X,Y}$ be a feature-label pair, whose distribution is unknown. Assume that $\mathcal{Y}\triangleq\mathrm{supp}(P_Y)$ is finite and known (as is the case in practice) and let $|\mathcal{Y}|=K$ be the cardinality of $\mathcal{Y}$. Let $\left\{(X_i,Y_i)\right\}_{i=1}^n$ be a set of n i.i.d. samples from $P_{X,Y}$, and T be a hidden layer in a noisy DNN with input X. Recall that $T=S+Z$, where S is a deterministic map of the previous layer and $Z\sim\mathcal{N}(0,\sigma^2\mathrm{I}_d)$. The tuple (X,Y,S,T) is jointly distributed according to $P_{X,Y}P_{S|X}P_{T|S}$, under which $Y-X-S-T$ forms a Markov chain. Our goal is to estimate the mutual information

$$I(Y;T)=h(P_S*\mathcal{N}_\sigma)-\sum_{y\in\mathcal{Y}}p_Y(y)h(P_{S|Y=y}*\mathcal{N}_\sigma), \quad (100)$$

based on a given estimator $\hat{h}$ of $h(P*\mathcal{N}_\sigma)$ that knows σ and uses i.i.d. samples from $P\in\mathcal{F}_d$. In (100), p_Y is the PMF associated with P_Y.

We first describe the sampling procedure for estimating each of the differential entropies from (100). For the unconditional entropy, P_S is sampled in the same manner described in Section IV-B for the estimation of $I(X;T)$. Denote the obtained samples by S^n. To sample from $P_{S|Y=y}$, for a fixed label $y\in\mathcal{Y}$, fix a sample set $\left\{(x_i,y_i)\right\}_{i=1}^n$ and consider the following. Define the set $\mathcal{I}_y\triangleq\left\{i\in[n]\middle|y_i=y\right\}$ and let $\mathcal{X}_y\triangleq\{x_i\}_{i\in\mathcal{I}_y}$ be the subset of features whose label is y; the elements of $\mathcal{X}_y$ are conditionally i.i.d. samples from $P_{X|Y=y}$. Now, feed each $x\in\mathcal{X}_y$ into the noisy DNN and collect the values induced at the layer preceding T. By applying the appropriate deterministic function on each of these samples we get a set of $n_y\triangleq|\mathcal{I}_y|$ i.i.d. samples from $P_{S|Y=y}$. Denote this sample set by $S^{n_y}(\mathcal{X}_y)$.

Similarly to Section IV-B, suppose we are given an estimator $\hat{h}(A^m,\sigma)$ of $h(P*\mathcal{N}_\sigma)$, for $P\in\mathcal{F}_d$, based on m i.i.d. samples $A^m=\{A_1,\ldots,A_m\}$ from P. Assume that $\hat{h}$ attains

$$\sup_{P\in\mathcal{F}_d}\mathbb{E}_{P^{\otimes m}}\left|h(P*\mathcal{N}_\sigma)-\hat{h}(A^m,\sigma)\right|\leq\Delta_{\sigma,d}(m). \quad (101)$$

Further assume that $\Delta_{\sigma,d}(m)<\infty$, for all $m\in\mathbb{N}$, and that $\lim_{m\to\infty}\Delta_{\sigma,d}(m)=0$, for any fixed σ and d (otherwise, $\hat{h}$ is not a good estimator and there is no hope using it for estimating $I(Y;T)$). Without loss of generality we may also assume that $\Delta_{\sigma,d}(m)$ is monotonically decreasing in m. Our estimator of $I(Y;T)$ is

$$\hat{I}_{\mathsf{Label}}\left(X^n,Y^n,\hat{h},\sigma\right)\triangleq\hat{h}(S^n,\sigma)-\sum_{y\in\mathcal{Y}}\hat{p}_{Y^n}(y)\hat{h}\big(S^{n_y}(\mathcal{X}_y),\sigma\big), \quad (102)$$

where $\hat{p}_{Y^n}(y)\triangleq\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{\{Y_i=y\}}$ is the empirical PMF associated with the labels Y^n. The following proposition bounds the expected absolute-error risk of $\hat{I}_{\mathsf{Label}}\left(X^n,Y^n,\hat{h},\sigma\right)$; the proof is given after the statement.

Proposition 10 (Label-Hidden Layer Mutual Information) *For the above described estimation setting, we have*

$$\sup_{P_{X,Y}:\ |\mathcal{Y}|=K}\mathbb{E}\left|I(Y;T)-\hat{I}_{\mathsf{Label}}\left(X^n,Y^n,\hat{h},\sigma\right)\right|\leq$$

$$\Delta_{\sigma,d}(n)+c^{(\mathsf{MI})}_{\sigma,d}\sqrt{\frac{K-1}{n}}+K\left(\Delta^{\star}_{\sigma,d}\cdot e^{-\frac{np_l^2}{8p_u}}+\Delta_{\sigma,d}\left(\frac{np_l}{2}\right)\right),$$

where

$$c^{(\mathsf{MI})}_{\sigma,d}\triangleq\frac{d}{2}\max\Big\{-\log(2\pi e\sigma^2),\log\big(2\pi e(1+\sigma^2)\big)\Big\} \quad (103a)$$

$$p_l\triangleq\min_{y\in\mathcal{Y}}p_Y(y) \quad (103b)$$

$$p_u\triangleq\max_{y\in\mathcal{Y}}p_Y(y) \quad (103c)$$

$$\Delta^{\star}_{\sigma,d}\triangleq\max_{n\in\mathbb{N}}\Delta_{\sigma,d}(n). \quad (103d)$$

The proof is reminiscent of that of Proposition 9, but with a few technical modifications accounting for n_y being a random quantity (as it depends on the number of Y_i-s that equal to y). To control n_y we use the concentration of the Binomial distribution about its mean.

Proof: Fix $P_{X,Y}$ with $|\mathcal{Y}|=K$, and use the triangle inequality to get

$$\begin{aligned}&\mathbb{E}\Big|I(Y;T)-\hat{I}_{\mathsf{Label}}\left(X^n,Y^n,\hat{h},\sigma\right)\Big|\\&\leq\underbrace{\mathbb{E}\Big|h(P_S*\mathcal{N}_\sigma)-\hat{h}(S^n,\sigma)\Big|}_{\text{(I)}}\\&\quad+\underbrace{\sum_{y\in\mathcal{Y}}\big|h(P_{S|Y=y}*\mathcal{N}_\sigma)\big|\mathbb{E}\big|p_Y(y)-\hat{p}_{Y^n}(y)\big|}_{\text{(II)}}\\&\quad+\underbrace{\sum_{y\in\mathcal{Y}}\mathbb{E}\bigg|\hat{p}_{Y^n}(y)\Big(h(P_{S|Y=y}*\mathcal{N}_\sigma)-\hat{h}\big(S^{n_y}(\mathcal{X}_y),\sigma^2\big)\Big)\bigg|}_{\text{(III)}},\end{aligned} \quad (104)$$

where we have added and subtracted $\sum_{y\in Y}\hat{p}_{Y^n}(y)h(P_{S|Y=y}*\mathcal{N}_\sigma)$ inside the original expectation.

Clearly, (I) is bounded by $\Delta_{\sigma,d}(n)$. For (II), we first bound the conditional differential entropies. For any $y\in\mathcal{Y}$, we have

$$\begin{aligned}h(P_{S|Y=y}*\mathcal{N}_\sigma)=h(S+Z|Y=y)&\geq h(S+Z|S,Y=y)\\&=\frac{d}{2}\log(2\pi e\sigma^2),\end{aligned} \quad (105)$$

where the last equality is since (Y,S) is independent of $Z\sim\mathcal{N}(0,\sigma^2\mathrm{I}_d)$. Furthermore,

$$\begin{aligned}h(P_{S|Y=y}*\mathcal{N}_\sigma)&\leq\sum_{k=1}^{d}h\big(S(k)+Z(k)\big|Y=y\big)\\&\leq\frac{d}{2}\log\big(2\pi e(1+\sigma^2)\big),\end{aligned} \quad (106)$$

where the first inequality is because independence maximizes differential entropy, while the second inequality uses $\mathsf{var}\big(S(k)+Z(k)\big|Y=y\big)\leq 1+\sigma^2$. Combining (105) and (106) we obtain

$$\begin{aligned}&\big|h(P_{S|Y=y}*\mathcal{N}_\sigma)\big|\\&\quad\leq c^{(\mathsf{MI})}_{\sigma,d}\triangleq\frac{d}{2}\max\Big\{-\log(2\pi e\sigma^2),\log\big(2\pi e(1+\sigma^2)\big)\Big\}.\end{aligned} \quad (107)$$

For the expected value in (II), monotonicity of moment gives

$$\begin{aligned}\mathbb{E}\big|p_Y(y)-\hat{p}_{Y^n}(y)\big|&\leq\sqrt{\mathsf{var}\big(p_{Y^n}(y)\big)}\\&=\sqrt{\frac{1}{n}\mathsf{var}\left(\mathbb{1}_{\{Y=y\}}\right)}\\&=\sqrt{\frac{p_Y(y)\big(1-p_Y(y)\big)}{n}}.\end{aligned} \quad (108)$$

Using (107) and (108) we bound Term (II) as follows:

$$\text{(II)}\leq\frac{c^{(\mathsf{MI})}_{\sigma,d}}{\sqrt{n}}\sum_{y\in\mathcal{Y}}\sqrt{p_Y(y)\big(1-p_Y(y)\big)}\leq c^{(\mathsf{MI})}_{\sigma,d}\sqrt{\frac{K-1}{n}}, \quad (109)$$

where the last step uses the Cauchy-Schwarz inequality.

For Term (III), we first upper bound $\hat{p}_{Y^n}(y)\leq 1$, for all $y\in\mathcal{Y}$, which leaves us to deal with the sum of expected absolute errors in estimating the conditional entropies. Fix $y\in\mathcal{Y}$, and notice that $n_y\sim\mathsf{Binom}\big(p_Y(y),n\big)$. Define $p_l\triangleq\min_{y\in\mathcal{Y}}p_Y(y)$ and $p_u\triangleq\max_{y\in\mathcal{Y}}p_Y(y)$ as in the statement of Proposition 10. Using a Chernoff bound for the Binomial distribution we have that for any $k\leq np_Y(y)$,

$$\begin{aligned}\mathbb{P}\Big(n_y\leq k\Big)&\leq\exp\left(-\frac{1}{2p_Y(y)}\cdot\frac{\big(np_Y(y)-k\big)^2}{n}\right)\\&\leq\exp\left(-\frac{1}{2p_u}\cdot\frac{\big(np_Y(y)-k\big)^2}{n}\right).\end{aligned}$$

Set $k^\star_y=n\left(p_Y(y)-\frac{1}{2}p_l\right)\in\big(0,np_Y(y)\big)$ into the above to get

$$\mathbb{P}\Big(n_y\leq k^\star_y\Big)\leq\exp\left(-\frac{np_l^2}{8p_u}\right). \quad (110)$$

Setting $\Delta^\star_{\sigma,d}\triangleq\max_{n\in\mathbb{N}}\Delta_{\sigma,d}(n)$, we note that $\Delta^\star_{\sigma,d}<\infty$ by hypothesis, and bound (III) as follows:

$$\begin{aligned}\text{(III)}&\leq\sum_{y\in\mathcal{Y}}\mathbb{E}\Big|h(P_{S|Y=y}*\mathcal{N}_\sigma)-\hat{h}\big(S^{n_y}(\mathcal{X}_y),\sigma^2\big)\Big|\\&\overset{(a)}{=}\sum_{y\in\mathcal{Y}}\mathbb{E}_{n_y}\left[\mathbb{E}\left[\Big|\hat{p}_{Y^n}(y)\Big(h(P_{S|Y=y}*\mathcal{N}_\sigma)-\hat{h}\big(S^{n_y}(\mathcal{X}_y),\sigma^2\big)\Big|\Big|n_y\right]\right]\\&\overset{(b)}{=}\sum_{y\in\mathcal{Y}}\mathbb{E}_{n_y}\Delta_{\sigma,d}(n_y)\\&\overset{(c)}{=}\sum_{y\in\mathcal{Y}}\mathbb{P}\Big(n_y\leq k^\star_y\Big)\mathbb{E}\big[\Delta_{\sigma,d}(n_y)\big|n_y\leq k^\star_y\big]\\&\qquad+\mathbb{P}\Big(n_y>k^\star_y\Big)\mathbb{E}\big[\Delta_{\sigma,d}(n_y)\big|n_y>k^\star_y\big]\\&\overset{(d)}{\leq}K\left(\Delta^\star_{\sigma,d}\cdot e^{-\frac{np_l^2}{8p_u}}+\Delta_{\sigma,d}\left(\frac{np_l}{2}\right)\right),\end{aligned} \quad (111)$$

where (a) and (c) use the law of total expectation, (b) is since for each fixed $n_y=k$, the expected differential entropy estimation error (inner expectation) is bounded by $\Delta_{\sigma,d}(k)$, while (d) relies on (110), the definition of $\Delta^\star_{\sigma,d}$ and the fact that $\Delta_{\sigma,d}(n)$ is monotonically decreasing with n along with $k^\star_y\geq\frac{np_l}{2}$, for all $y\in\mathcal{Y}$. Inserting (I) $\leq\Delta_{\sigma,d}(n)$ together with the bounds from (109) and (111) back into (104) and taking the supremum over all $P_{X,Y}$ with $|\mathcal{Y}|=K$ concludes the proof. ■

Appendix C
Proof of Lemma 5

We expand $I(S^n;Y) = h(Y) - h(Y|S^n)$. Let $T = S+Z \sim P * \mathcal{N}_\sigma$ and first note that for any measurable set $\mathcal{A}$,

$$\begin{aligned}\mathbb{P}(Y \in \mathcal{A}) &= \mathbb{P}\big(S_W + Z \in \mathcal{A}\big)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(S_i + Z \in \mathcal{A}) = \mathbb{P}(T \in \mathcal{A}).\end{aligned}$$

Thus, $h(Y) = h(P * \mathcal{N}_\sigma)$. It remains to show that $h(Y|S^n) = \mathbb{E}_{P^{\otimes n}} h(\hat{P}_{S^n} * \mathcal{N}_\sigma)$. Fix $S^n = s^n$ and consider

$$\begin{aligned}\mathbb{P}\big(Y \in \mathcal{A}\big|S^n = s^n\big) &= \mathbb{P}\big(S_W + Z \in \mathcal{A}\big|S^n = s^n\big)\\ &= \frac{1}{n}\mathbb{P}\big(s_i + Z \in \mathcal{A}\big),\end{aligned}$$

which implies that the density $p_{Y|S^n=s^n}$ equals the density of $\hat{P}_{s^n} * \mathcal{N}_\sigma$. Consequently, $h(Y|S^n = s^n) = h(\hat{P}_{s^n} * \mathcal{N}_\sigma)$, and by definition of conditional entropy $h(Y|S^n) = \mathbb{E}_{P^{\otimes n}} h(\hat{P}_{S^n} * \mathcal{N}_\sigma)$.

Appendix D
Proof of Proposition 4

We start from the derivation of (11), which shows that

$$\begin{aligned}\mathbb{E}_{P^{\otimes n}}\chi^2\left(\hat{P}_{S^n} * \mathcal{N}_\sigma \middle\| P * \mathcal{N}_\sigma\right) &= \frac{1}{n} I_{\chi^2}(S;Y)\\ &= \frac{1}{n}\left(\int_{\mathbb{R}^d}\frac{\mathbb{E}_P\varphi_\sigma^2(z-S)}{q(z)}\,\mathsf{d}z - 1\right)\end{aligned}$$

for $S \sim P$ and $Y = S + Z$, where $Z \sim \mathcal{N}_\sigma$ is independent of S. Recalling that without loss of generality $\sigma = 1$ and that $q(z) = \mathbb{E}_P\varphi_1(z-S)$, a sufficient condition for divergence in Proposition 4 is

$$\int_{\mathbb{R}}\frac{\mathbb{E}_P\varphi_{\frac{1}{\sqrt{2}}}(z-S)}{\mathbb{E}_P\varphi_1(z-S)}\,\mathsf{d}z = \infty. \tag{112}$$

Under the P from (15), the left-hand side (LHS) of (112) becomes

$$\begin{aligned}&\int_{\mathbb{R}}\frac{\sum_{k=0}^{\infty}p_k\varphi_{\frac{1}{\sqrt{2}}}(z-r_k)}{\sum_{k=0}^{\infty}p_k\varphi_1(z-r_k)}\,\mathsf{d}z\\ &= \sum_{k=0}^{\infty}\int_{\mathbb{R}}\frac{p_k\varphi_{\frac{1}{\sqrt{2}}}(z-r_k)}{p_k\varphi_1(z-r_k)+\sum_{j\neq k}p_j\varphi_1(z-r_j)}\,\mathsf{d}z\\ &= \sum_{k=0}^{\infty}\int_{\mathbb{R}}\frac{\varphi_{\frac{1}{\sqrt{2}}}(z-r_k)}{\varphi_1(z-r_k)}\frac{1}{1+\sum_{j\neq k}\frac{p_j}{p_k}\frac{\varphi_1(z-r_j)}{\varphi_1(z-r_k)}}\,\mathsf{d}z\\ &\geq \sum_{k=1}^{\infty}\int_{r_k-\frac{1}{100}}^{r_k+\frac{1}{100}}\left[\frac{\varphi_{\frac{1}{\sqrt{2}}}(z-r_k)}{\varphi_1(z-r_k)}\right.\\ &\qquad\left.\times\frac{1}{1+\sum\limits_{j=0}^{k-1}\frac{p_j}{p_k}\frac{\varphi_1(z-r_j)}{\varphi_1(z-r_k)}+\sum\limits_{j=k+1}^{\infty}\frac{p_j}{p_k}\frac{\varphi_1(z-r_j)}{\varphi_1(z-r_k)}}\right]\mathsf{d}z,\end{aligned} \tag{113}$$

where the inequality follows since the integrands are all nonnegative and the domain of integration has been reduced.

We now bound the sums in the second denominator of (113) for $k > 0$ and $z \in \left\{r_k - \frac{1}{100}, r_k + \frac{1}{100}\right\}$ (as indicated by the support of the outer sum and integral). First, consider the ratio $\frac{p_j}{p_k}\frac{\varphi_1(z-r_j)}{\varphi_1(z-r_k)}$. For $j = 0$ and $\epsilon \leq \frac{1}{4}$, we have

$$\begin{aligned}\frac{p_0}{p_k}\frac{\varphi_1(z)}{\varphi_1(z-r_k)} &\leq C\cdot\exp\left\{\epsilon r_k^2 + \frac{r_k^2 - 2\left(r_k - \frac{1}{100}\right)r_k}{2}\right\}\\ &= C\cdot\exp\left\{\left(\epsilon - \frac{1}{2}\right)r_k^2 + \frac{r_k}{100}\right\} \leq C,\end{aligned} \tag{114}$$

where C is a constant depending on K only, and in the last inequality is because $r_k \geq 1$, for $k \geq 1$. For $j > 0$, denoting $\alpha_\epsilon \triangleq 1 - \sqrt{2\epsilon}$, the bound becomes

$$\begin{aligned}&\frac{p_j}{p_k}\frac{\varphi_1(z-r_j)}{\varphi_1(z-r_k)} = \exp\left\{\epsilon(r_k^2 - r_j^2) + \frac{r_k^2 - r_j^2 - 2z(r_k - r_j)}{2}\right\}\\ &= \exp\left\{\epsilon r_k^2\left(1 - \alpha_\epsilon^{-2(j-k)}\right)\right.\\ &\qquad\left. + \frac{r_k^2\left(1 - \alpha_\epsilon^{-2(j-k)}\right) - 2zr_k\left(1 - \alpha_\epsilon^{-(j-k)}\right)}{2}\right\}.\end{aligned} \tag{115}$$

Using (114) and (115), for $\epsilon = \frac{1}{4}$ and $\alpha \triangleq \alpha_{\frac{1}{4}} = 1 - \frac{1}{\sqrt{2}}$, we have

$$\begin{aligned}&\sum_{j=0}^{k-1}\frac{p_j}{p_k}\frac{\varphi_1(z-r_j)}{\varphi_1(z-r_k)}\\ &\leq C + \sum_{j=1}^{k-1}\exp\left\{\epsilon r_k^2\left(1 - \alpha^{-2(j-k)}\right) + \frac{r_k^2}{2}\left(1 - \alpha^{-2(j-k)}\right)\right.\\ &\qquad\left. - r_k\left(r_k - \frac{1}{100}\right)\left(1 - \alpha^{-(j-k)}\right)\right\}\\ &\leq C + \sum_{j=1}^{k-1}\exp\left\{\epsilon r_k^2\left(1 - \alpha^{-2(j-k)}\right) + \frac{r_k^2}{2}\left(1 - \alpha^{-2(j-k)}\right)\right.\\ &\qquad\left. - r_k\left(r_k - \frac{1}{100}\right)\left(1 - \alpha^{-2(j-k)}\right)\min_j\left(\left(1 + \alpha^{-(j-k)}\right)^{-1}\right)\right\}\\ &= C + \sum_{j=1}^{k-1}\exp\left\{\left(\epsilon + \frac{1}{2} - \frac{1}{2 - \sqrt{1/2}}\right)r_k^2\left(1 - \alpha^{-2(j-k)}\right)\right.\\ &\qquad\left. + \frac{r_k}{100}\left(1 - \alpha^{-2(j-k)}\right)\right\}\\ &\leq C + k - 1,\end{aligned} \tag{116}$$

where the last inequality follows since $\epsilon = \frac{1}{4}$, $r_k^2 \geq r_k$, for $k \geq 1$, and $\frac{1}{4} + \frac{1}{2} - \frac{1}{2 - \sqrt{1/2}} + \frac{1}{100} < 0$.

Proceeding onto the series for $j \geq k+1$, we have

$$\sum_{j=k+1}^{\infty}\frac{p_j}{p_k}\frac{\varphi_1(z-r_j)}{\varphi_1(z-r_k)}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sum_{j=k+1}^{\infty} \exp\left\{ \frac{3}{4} r_k^2 \left(1-\alpha^{-2(j-k)}\right) \right. \\
&\qquad\qquad \left. - r_k \left(r_k + \frac{1}{100}\right)\left(1-\alpha^{-(j-k)}\right) \right\} \\
&\leq \sum_{j=k+1}^{\infty} \exp\left\{ -\frac{3}{4} r_k^2 \alpha^{-2(j-k)} + r_k \left(r_k + \frac{1}{100}\right) \alpha^{-(j-k)} \right\} \\
&\leq \sum_{j=k+1}^{\infty} \exp\left\{ -r_k^2 \alpha^{-(j-k)} \left(\frac{3\alpha^{-(j-k)}}{4} - \frac{101}{100} \right) \right\} \\
&\overset{(b)}{\leq} \sum_{j=k+1}^{\infty} \exp\left\{ -\frac{1}{4} r_k^2 \alpha^{-(j-k)} \right\} \\
&\leq \sum_{\ell=1}^{\infty} \exp\left\{ -\frac{1}{4} \alpha^{-\ell} \right\} \\
&\overset{(c)}{\leq} \frac{1}{4},
\end{aligned}
$$

where (a) uses (115) and the fact that $\left(1-\alpha^{-(j-k)}\right)$ is negative for $j > k$, (b) is since $\frac{3\alpha^{-t}}{4} - \frac{101}{100} \geq 1/4$ for all $t \geq 1$, and (c) follows by numerical computation and because the series converges by the ratio test.

Substituting these bounds into the LHS of (112), we get

$$
\begin{aligned}
&\int_{\mathbb{R}} \frac{\sum_{k=0}^{\infty} p_k \varphi_{\frac{1}{\sqrt{2}}}(z-r_k)}{\sum_{k=0}^{\infty} p_k \varphi_1(z-r_k)} \mathsf{d}z \\
&\quad \geq \sum_{k=1}^{\infty} \int_{r_k-\frac{1}{100}}^{r_k+\frac{1}{100}} \frac{\varphi_{\frac{1}{\sqrt{2}}}(z-r_k)}{\varphi_1(z-r_k)} \frac{1}{1+C+k-1+\frac{1}{4}} \mathsf{d}z \\
&\quad = \left(\int_{-\frac{1}{4}}^{\frac{1}{4}} \frac{\varphi_{\frac{1}{\sqrt{2}}}(z)}{\varphi_1(z)} \mathsf{d}z \right) \sum_{k=1}^{\infty} \frac{1}{k+C+\frac{1}{4}}. \qquad (117)
\end{aligned}
$$

The RHS above diverges because the integral is nonzero and $\sum_{k=1}^{\infty} \frac{1}{k+C+1/4}$ is a harmonic series.

References

[1] R. M. Dudley, "The speed of mean Glivenko-Cantelli convergence," *Ann. Math. Stats.*, vol. 40, no. 1, pp. 40–50, Feb. 1969.

[2] V. Dobrić and J. E. Yukich, "Asymptotics for transportation cost in high dimensions," *J. Theoretical Prob.*, vol. 8, no. 1, pp. 97–118, Jan. 1995.

[3] N. Fournier and A. Guillin, "On the rate of convergence in wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, pp. 707–738, 2015.

[4] L. Paninski, "Estimation of entropy and mutual information," *NEURAL COMPUTATION*, vol. 15, pp. 1191–1254, 2004.

[5] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over Lipschitz balls," *arXiv preprint arXiv:1711.02141*, Nov. 2017.

[6] G. Valiant and P. Valiant, "A CLT and tight lower bounds for estimating entropy." in *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 17, Nov. 2010, p. 9.

[7] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, June 2016.

[8] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and J. M. Robins, "Nonparametric von Mises estimators for entropies, divergences and mutual informations," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 397–405.

[9] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Ensemble estimation of information divergence †," *Entropy*, vol. 20, no. 8, 2018. [Online]. Available: http://www.mdpi.com/1099-4300/20/8/560

[10] ——, "Improving convergence of divergence functional ensemble estimators," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1133–1137.

[11] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[12] H. S. A. Kraskov and P. Grassberger, "Estimating mutual information," *Phys. rev. E*, vol. 69, no. 6, p. 066138, June 2004.

[13] A. B. Tsybakov and E. C. V. der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," *Scandinavian Journal of Statistics*, pp. 75–83, Mar. 1996.

[14] S. K, R. Raich, and A. O. Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4135–4159, Jul. 2012.

[15] K. Sricharan, D. Wei, and A. O. H. III, "Ensemble estimators for multivariate entropy estimation," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4374–4388, Jul. 2013.

[16] S. Singh and B. Póczos, "Finite-sample analysis of fixed-k nearest neighbor density functional estimators," in *Advances in Neural Information Processing Systems*, 2016, pp. 1217–1225.

[17] S. Delattre and N. Fournier, "On the Kozachenko–Leonenko entropy estimator," *Journal of Statistical Planning and Inference*, vol. 185, pp. 69–93, Jun. 2017.

[18] J. Jiao, W. Gao, and Y. Han, "The nearest neighbor information estimator is adaptively near minimax rate-optimal," *arXiv preprint arXiv:1711.08824*, 2017.

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

[20] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. V. der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, Jun. 1997.

[21] G. Biau and L. Devroye, *Lectures on the nearest neighbor method*. Springer, 2015.

[22] A. W. V. der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[23] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[24] J.-H. Jacobsen, A. Smeulders, and E. Oyallon, "i-RevNet: Deep invertible networks," *arXiv preprint arXiv:1802.07088*, 2018.

[25] K. Liu, R. A. Amjad, and B. C. Geiger, "Understanding individual neuron importance using information theory," *arXiv preprint arXiv:1804.06679*, 2018.

[26] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, "Entropy and mutual information in models of deep neural networks," *arXiv preprint arXiv:1805.09785*, 2018.

[27] G. Reeves, "Additivity of information in multilayer networks via additive gaussian noise transforms," in *Proc. 55th Annu. Allerton Conf. Commun., Control and Comput. (Allerton-2017)*, Monticello, Illinois, Oct. 2017, pp. 1064–1070.

[28] Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, "Estimating information flow in neural networks," Accepted to *the International Conference on Machine Learning (ICML-2019)*, vol. Long Beach, CA, US, Jun. 2019, arxiv link: https://arxiv.org/abs/1810.05728.

[29] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proceedings of the Information Theory Workshop (ITW)*, Jerusalem, Israel, Apr.-May 2015, pp. 1–5.

[30] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, arXiv:1703.00810 [cs.LG].

[31] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[32] D. Hsu, S. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electronic Communications in Probability*, vol. 17, 2012.

[33] J. Weed, "Sharper rates for estimating differential entropy under gaussian convolutions," Massachusetts Institute of Technology (MIT), Tech. Rep., Dec. 2018.

[34] F. Nielsen and R. Nock, "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 10–13, 2014.

[35] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[36] F.-Y. Wang and J. Wang, "Functional inequalities for convolution probability measures," *Ann. Inst. Henri Poincaré Probab. Stat.*, vol. 52, no. 2, pp. 898–914, 2016. [Online]. Available: https://doi.org/10.1214/14-AIHP659

[37] F. Otto and C. Villani, “Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality,” *J. Funct. Anal.*, vol. 173, no. 2, pp. 361–400, 2000. [Online]. Available: https://doi.org/10.1006/jfan.1999.3557

[38] S. G. Bobkov, I. Gentil, and M. Ledoux, “Hypercontractivity of Hamilton-Jacobi equations,” *Markov Process. Related Fields*, vol. 8, no. 2, pp. 233–235, 2002, inhomogeneous random systems (Cergy-Pontoise, 2001).

[39] E. Boissard and T. Le Gouic, “On the mean speed of convergence of empirical and occupation measures in wasserstein distance,” in *Annales de l’IHP Probabilités et statistiques*, vol. 50, no. 2, 2014, pp. 539–563.

[40] S. Bobkov and M. Ledoux, “One-dimensional empirical measures, order statistics and Kantorovich transport distances,” *preprint*, 2014.

[41] Y. Polyanskiy and Y. Wu, “Lecture notes on information theory,” *Lecture Notes 6.441, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA*, 2012–2017.

[42] J. Chen, “A general lower bound of minimax risk for absolute-error loss,” *Canadian Journal of Statistics*, vol. 25, no. 4, pp. 545–558, Dec. 1997.

[43] C. P. Robert, *Monte Carlo Methods*. Wiley Online Library, 2004.

[44] R. Gallager, *Information theory and reliable communication*. Springer, 1968, vol. 2.

[45] Y. Polyanskiy and Y. Wu, “Wasserstein continuity of entropy and outer bounds for interference channels,” *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 3992–4002, Jul. 2016.