Estimating Differential Entropy under Gaussian Convolutions

Ziv Goldfeld, Kristjan Greenewald and Yury Polyanskiy

Abstract

This paper studies the problem of estimating the differential entropy $h(S+Z)$, where S and Z are independent d-dimensional random variables with $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. The distribution of S is unknown, but n independently and identically distributed (i.i.d) samples from it are available. The question is whether having access to samples of S as opposed to samples of $S+Z$ can improve estimation performance. We show that the answer is positive.

More concretely, we first show that despite the regularizing effect of noise, the number of required samples still needs to scale exponentially in d. This result is proven via a random-coding argument that reduces the question to estimating the Shannon entropy on a $2^{O(d)}$-sized alphabet. Next, for a fixed d and $n \to \infty$, it is shown that a simple plugin estimator, given by the differential entropy of the empirical distribution from S convolved with the Gaussian density, achieves the loss of $O\left(\frac{(\log n)^{d/4}}{\sqrt{n}}\right)$. Note that the plugin estimator amounts here to the differential entropy of a d-dimensional Gaussian mixture, for which we propose an efficient Monte Carlo computation algorithm. At the same time, estimating $h(S+Z)$ via generic differential entropy estimators applied to samples from $S+Z$ would only attain much slower rates of order $O(n^{-1/d})$, despite the smoothness of P_{S+Z}.

As an application, which was in fact our original motivation for the problem, 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, minimax rates, mutual information.

I. Introduction

This work studies a new nonparametric and high-dimensional differential entropy estimation problem. The goal is to estimated the differential entropy $h(S+Z)$, based on samples of one random variable while knowing the distribution of the other. Specifically, let $S \sim P$ be an arbitrary (continuous / discrete / mixed) random variable with values in $\mathbb{R}^d$ and $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ be an independent isotropic Gaussian. Upon observing n i.i.d. samples $S^n \triangleq (S_1, \ldots, S_n)$ from P and assuming σ is known, we aim to estimate $h(S+Z) = h(P * \varphi_\sigma)$, where φ_σ

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.
This work will be presented in part at the 2018 IEEE International Conference on the Science of Electrical Engineering (ICSEE-2018), Eilat, Israel.
Z. Goldfeld and Y. Polyanskiy are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 US (e-mails: zivg@mit.edu, yp@mit.edu). K. Greenewald is with IBM Research, Cambridge, MA 02142 US (email: kristjan.h.greenewald@ibm.com)

denotes the probability density function (PDF) of a centered isotropic Gaussian with parameter σ.[1] To investigate the decision-theoretic fundamental limit, we consider the minimax absolute-error risk of differential entropy estimation:

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

where $\mathcal{F}_d$ is a nonparametric class of d-dimensional distributions and $\hat{h}$ is the estimator. The sample complexity $n_d^\star(\eta,\sigma)$ is the smallest number of samples for which estimation within an additive gap η is possible. The goal of this work is to study whether having access to 'clean' samples of S can improve estimation performance compared to the case where only 'noisy' samples of $S+Z$ are available and the distribution of Z is unknown.

A. Motivation

Our motivation to study the considered differential entropy estimation problem stems from mutual information estimation over deep neural networks (DNNs). There has been a recent surge of interest in estimating the mutual information between selected groups of neurons in a DNN [1]–[5], partially driven by the Information Bottleneck (IB) theory [6], [7]. Attention mostly focuses on the mutual information $I(X;T)$, between the input feature X and a hidden activity vector T. However, as explained in [5], this quantity is vacuous in deterministic DNNs[2] 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 [5], 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 renders $I(X;T)$ meaningful for studying deep learning, the concatenation of Gaussian noises and nonlinearities make 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 the appropriate mode of operation becomes treating it as unknown, belonging to some nonparametric class of distributions. This work sets the groundwork for estimating $I(X;T)$ (or any other mutual information between layers) DNN classifiers while providing theoretical guarantees that are not vacuous when d is relatively large.

To achieve this, we distill the estimation of $I(X;T)$ to the problem of differential entropy estimation under Gaussian convolutions described above. 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 by feeding data samples up the network; the distribution of Z is known since the noise is injected by design. Estimating mutual information over noisy DNNs thus boils down to the considered differential entropy estimation setup, which is the focus of this work.

B. Past Works for Unstructured Differential Entropy Estimation

General-purpose differential entropy estimators are applicable for estimating $h(S+Z)$ by accessing noisy i.i.d. samples of $S+Z$. However, the theoretical guarantees for unstructured differential entropy estimation commonly

[1]See the notation section at the end of the introduction for a precise definition of $P*\varphi_\sigma$ when P is discrete / continuous / mixed.

[2]i.e., DNNs that, upon fixing their parameters, define a deterministic map from input to output.

found in the literature are invalid for our framework. There are two prevailing approaches for estimating the nonsmooth differential entropy functional: the first relying on kernel density estimators (KDEs) [8]–[10], and the second using k nearest neighbor (kNN) techniques [11]–[18] (see also [19], [20] for surveys). However, performance analyses of these estimators often restrict attention to nonparametric classes of smooth densities that are bounded away from zero. Various works require uniform boundedness from zero [8], [9], [14], [21], [22], while others restrict it on average [13], [16], [23]–[25]. Since the convolved density $P * \varphi_\sigma$ can attain arbitrarily small values these results do not apply in the considered scenario.

To the best of our knowledge, the only two works that dropped the boundedness from zero assumption are [10] and [18], where the minimax risk of a KDE-based method and the Kozachenko-Leonenko (KL) entropy estimator [12] are, respectively, analyzed. These results assume that the densities are supported inside the unit hypercube, satisfy periodic boundary conditions and have (Lipschitz or Hölder) smoothness parameter $s \in (0, 2]$. The convolved density $P * \varphi_\sigma$ violates the two former assumptions. Notably, the analysis from [10] was also extended to densities supported on $\mathbb{R}^d$ that have sub-Gaussian tails. The derived upper bound on the minimax risk in this sub-Gaussian regime applies for our estimation setup when P is compactly supported or sub-Gaussian. However, the obtained risk convergence rate (overlooking some multiplicative polylogarithmic factors) is $O\left(n^{-\frac{s}{s+d}}\right)$, which quickly deteriorates with dimension d and in unable to fully exploit the smoothness of $P*\varphi_\sigma$ due to the $s \leq 2$ restriction. [3] Consequently, this risk bound is ineffective for evaluating the error of implemented estimators, even for moderate dimensions. We also note that all the above results include implicit constants that depend on d (possibly exponentially) that may (when combined with the weak decay with respect to n) significantly increase the number of samples required to achieve a desired estimation accuracy. We therefore ask if exploiting the explicit modeling of $P * \varphi_\sigma$, with the 'clean' samples from P and the knowledge of φ_σ, can improve estimation performance.

C. This Work

We begin the study of estimating $h(P*\varphi_\sigma)$ by showing that an exponential dependence of the sample complexity on dimension is unavoidable. Specifically, we prove that $n_d^\star(\eta, \sigma) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma)$ is a positive, monotonically decreasing function of σ. The proof relates the estimation of $h(P * \varphi_\sigma)$ to estimating the discrete entropy of a distribution over a capacity achieving codebook for the additive white Gaussian noise (AWGN) channel. Viewing $h(P * \varphi_\sigma)$ as a functional of P with parameter φ_σ, i.e., $\mathsf{T}_{\varphi_\sigma}(P) = h(P * \varphi_\sigma)$, we then analyze the performance of the plugin estimator. Specifically, based on the empirical measure $\hat{P}_{S^n} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{S_i}$, where δ_{S_i} is the Dirac measure associated with S_i, we consider the estimator

$$\hat{h}_{n,\sigma} \triangleq h\left(\hat{P}_{S^n} * \varphi_\sigma\right). \tag{2}$$

Note that $\hat{h}_{n,\sigma}$ approximates $h(P*\varphi_\sigma)$ via the differential entropy of a Gaussian mixture with centers at the sample points $\{S_i\}_{i=1}^n$. When P belongs to a class of compactly supported distributions on $\mathbb{R}^d$ (corresponding, for instance,

[3] This convergence rate is typical in unstructured differential entropy estimation. Indeed, all the results cited above (applicable in our estimation problem or otherwise) show that the estimation risk decays as $O\left(n^{-\frac{\alpha}{\beta+d}}\right)$, where α, β are constants that may depend on s and d.

to tanh/sigmoid DNNs) or when it has sub-Gaussian marginals (corresponding to ReLU DNNs with a sub-Gaussian input), we show that the minimax absolute-error risk $\hat{h}_{n,\sigma}$ is bounded by $c_{\sigma,d}\frac{(\log n)^{d/4}}{\sqrt{n}}$, with the constant $c_{\sigma,d}$ (that also depends on d exponentially) explicitly characterized. This convergence rate presents a significant improvement over the $O\left(n^{-\frac{\alpha s}{\beta s+d}}\right)$ rates derived for general-purpose differential entropy estimators. This is, of course, expected since $\hat{h}_{n,\sigma}$ is tailored for our particular setup, while generic KDE- or kNN-based estimators are not designed to exploit the $T = S + Z$ structure nor the 'clean' samples S^n.

Our proof exploits the $t\log\left(\frac{1}{t}\right)$ modulus of continuity for the map $x \mapsto x\log x$ to bound the absolute estimation error in terms of the pointwise mean squared error (MSE) of $\hat{P}_{S^n} * \varphi_\sigma$ as a proxy of the true density $P * \varphi_\sigma$. The analysis then reduces to integrating the modulus of continuity evaluated at the square root of the MSE bound. Functional optimization and concentration of measure arguments are used to control the integral and obtain the result. A similar result is derived for the nonparametric class of distributions P that have sub-Gaussian marginals. The bounded support and sub-Gaussian results essentially capture all cases of interest, and in particular, correspond respectively to DNNs with bounded nonlinearities and to unbounded nonlinearities with weight regularization.

We then focus on the practical implementation of $\hat{h}_{n,\sigma}$. While our performance guarantees give sufficient conditions on the number of samples needed to drive the estimation error below a desired threshold, these are worst-case result by definition. In practice, the unknown distribution P may not be one that follows the minimax rates, and the resulting decay of error could be faster. However, while the variance of the $\hat{h}_{n,\sigma}$ can be empirically evaluated using bootstrapping, there is no empirical test for the bias. We derive a lower bound on the bias of our estimator to have a guideline of the least number of samples needed for unbiased estimation. Our last step is to propose an efficient implementation of $\hat{h}_{n,\sigma}$ based on Monte Carlo (MC) integration. Since $\hat{h}_{n,\sigma}$ is simply the entropy of a known Gaussian mixture, MC integration using samples from this mixture allows a simple computation of $\hat{h}_{n,\sigma}$. We bound the MSE of the computed value that converge as $\frac{c^{(\mathsf{MC})}_{\sigma,d}}{n\cdot n_{\mathsf{MC}}}$, where n is the number of centers in the mixture[4], n_{MC} is the number of MC samples, and $c^{(\mathsf{MC})}_{\sigma,d}$ is an explicit constant that depends linearly on d. The proof leverages 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 DNN for a 3-dimensional spiral dataset classification) illustrate the gain of the ad-hoc $\hat{h}_{n,\sigma}$ estimator over its general-purpose counterparts, both in the rate of error decay and in its scalability with dimension.

The remainder of this paper is organized as follows. In Section II we set up the estimation problem, state our main results and discuss them. Section III presents applications of the considered estimation problem, focusing on mutual information estimation over DNNs. Simulation results are shown in Section IV-B, while Section V provides proofs. Our main insights from this work and appealing future directions are discussed in Section VI.

Notation: Throughout this work logarithms are with respect to the natural base. For an integer $k \geq 1$, we set $[k] \triangleq \left\{i \in \mathbb{Z} \middle| 1 \leq i \leq k\right\}$. For a real number $p \geq 1$, the L^p-norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|_p = \left(\sum_{i=1}^d x^p(i)\right)^{1/p}$, while $\|x\|_\infty = \max_{1\leq i\leq d}|x(i)|$. Matrices are denoted by non-italic letters, e.g., A; the d-dimensional identity matrix is I_d. We use calligraphic letters, such as $\mathcal{X}$, to denote sets. The cardinality of a finite set $\mathcal{X}$ is $|\mathcal{X}|$. Probability

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

distributions are denoted by uppercase letters such as P or Q. The support of a d-dimensional distribution P, denoted by $\mathrm{supp}(P)$, is the smallest set $\mathcal{R} \subseteq \mathbb{R}^d$ such that $P(\mathcal{R}) = 1$. If P is discrete, the corresponding probability mass function (PMF) is designated by p, i.e., $p(x) = P\big(\{x\}\big)$, for $x \in \mathrm{supp}(P)$. With some abuse of notation, the PDF associated with a continuous distribution is also denoted by p. Whether p is a PMF or a PDF is of no consequence for most of our results; whenever the distinction is important, the nature of p will be clarified. The n-fold product distribution associated with P is denoted by $P^{\otimes n}$. To highlight that an expectation or a probability measure is with respect to an underlying distribution P we write $\mathbb{E}_P$ or $\mathbb{P}_P$; if P has a PMF/density p, we use $\mathbb{E}_p$ or $\mathbb{P}_p$ instead. For a random variable $X \sim P$ and a deterministic function $f : \mathrm{supp}(P) \to \mathbb{R}$, we sometimes highlight that the expectation of f is with respect to the underlying distribution of X by writing $\mathbb{E}_X f$. For a continuous random variable X with density p, we interchangeably use $h(X)$ and $h(p)$ for its differential entropy.

Lastly, since our estimation setting considers the sum of independent random variables $S + Z$, we oftentimes deal with convolutions. For two probability measures μ and ν on $\mathbb{R}^d$, their convolution is defined by

$$(\mu * \nu)(\mathcal{A}) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}_{\mathcal{A}}(x + y) d\mu(x) d\nu(y),$$

where $\mathbb{1}_{\mathcal{A}}$ is the indicator of the Borel set $\mathcal{A}$. If $S \sim \mu$ and $Z \sim \nu$ are independent random variables, then $S + Z \sim \mu * \nu$. In this work, Z is always an isotropic Gaussian with parameter σ, whose PDF is denoted by φ_σ. The random variable S, however, may be discrete, continuous or mixed. Regardless of the nature of $S \sim P$, the random variable $S + Z$ is always continuous and its PDF is denoted by $P * \varphi_\sigma$. By the latter we mean $(P * \varphi_\sigma)(x) = \int_{\mathbb{R}^d} p(u)\varphi_\sigma(x - u)du = (p * \varphi_\sigma)(x)$, when P is continuous with density p. If P is discrete with PMF p, then $(P * \varphi_\sigma)(x) = \sum_{u:\ p(u)>0} p(u)\varphi_\sigma(x - u)$. For a mixed distribution P, Lebesgue's decomposition theorem allows to write $P * \varphi_\sigma$ as the sum of two expressions as above. Henceforth, we typically overlook the exact structure of $P * \varphi_\sigma$ only mentioning it when it is consequential.

II. Main Results

A. Preliminary Definitions

Let $\mathcal{F}_d$ be the set of distributions P with $\mathrm{supp}(P) \subseteq [-1, 1]^d$.[5] We also consider the class of distributions whose marginals are sub-Gaussian [26]. The sub-Gaussian norm is defined as $\|X\|_{\psi_2} \triangleq \sup_{p \geq 1} p^{-\frac{1}{2}} \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}}$, and we set $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ as the class of distributions P of a d-dimensional random variable $S = \big(S(1), \ldots, S(d)\big)$ whose coordinates satisfy $\big\|S(i)\big\|_{\psi_2} \leq K$, for all $i \in [d]$. The class $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ will be used in Section III to handle DNNs with unbounded activation functions, such as ReLUs. Clearly, for any $S \sim P$ with $\mathrm{supp}(P) \subseteq [-1, 1]^d$ we have $\big\|S(i)\big\|_{\psi_2} \leq 1$, for all $i \in [d]$, and therefore $\mathcal{F}_d \subseteq \mathcal{F}_{d,1}^{(\mathsf{SG})}$.

B. Lower Bounds on Risk

We give two converse claims showing that the sample complexity is exponential in d.

[5]Any support included in a compact subset of $\mathbb{R}^d$ would do. We focus on the case of $\mathrm{supp}(P) \subseteq [-1, 1]^d$ due to its correspondence to a noisy DNN with tanh nonlinearities.

Theorem 1 (Exponential Sample-Complexity) *The following claims are true:*

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

Theorem 1 is proven in Section V-B, based on channel coding arguments. For instance, the proof of Part 1) relates the estimation of $h(P * \varphi_\sigma)$ to the output sequence of a peak-constrained AWGN channel. Then, we show that estimating the entropy of interest is equivalent to estimating the entropy of a discrete random variable with some distribution over a capacity-achieving codebook. The positive capacity of the considered AWGN channel means that the size of this codebook is exponential in d. Therefore, (discrete) entropy estimation over the codebook within a small additive gap $\eta > 0$ cannot be done with less than order of $\frac{2^{\gamma(\sigma)d}}{\eta d}$ samples. Furthermore, the exponent $\gamma(\sigma)$ is monotonically decreasing in σ, implying that larger values of σ 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 1 (Exponential Sample Complexity for Restricted Classes of Distributions) *Note that 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 $\varphi_{\frac{\sigma}{2}}$, i.e., replacing each P with $Q = P * \varphi_{\frac{\sigma}{2}}$. These Q distributions are smooth, but if one could accurately estimate $h\big(Q * \varphi_{\frac{\sigma}{2}}\big)$ over the convolved class, then $h(P * \varphi_\sigma)$ over $\mathcal{F}_d$ would have been estimated as well. Therefore, an exponential sample complexity lower bound applies also for the class of such smooth Q distributions.*

Remark 2 (Critical Value of Noise Parameter) *We state Theorem 1 in asymptotic form for simplicity; the full bounds are found in the proof (Section V-B). We also note that, for any d, the critical $\sigma_0(d)$ value from the 2nd part can be extracted by following the constants through the proof (which relies on Proposition 3 from [27]). These critical values are not unreasonably small. For example for $d = 1$, a careful analysis gives that Theorem 1 holds for all $\sigma < 0.08$. This threshold on σ changes very slowly when increasing d due to the rapid decay of the Gaussian density. As a reference point, note that the per-neuron noise variance values used in the noisy DNNs from [5] ranged from 0.005 to 0.1.*

C. Upper Bound on Risk

This is our main section, where we analyze the performance of the $\hat{h}_{n,\sigma}$ estimator from (2). Recall that $\hat{h}_{n,\sigma} \triangleq h\left(\hat{P}_{S^n} * \varphi_\sigma\right)$, where $\hat{P}_{S^n} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{S_i}$ is the empirical measure associated with the data S^n. The following theorem shows that the expected absolute error of $\hat{h}_{n,\sigma}$ decays like $O\left(\frac{\mathsf{Polylog}(n)}{\sqrt{n}}\right)$ for all dimensions d. We provide explicit constants (in terms of σ and d), which present an exponential dependence on the dimension, in accordance to the results of Theorem 1.

Dimension	1	2	3	5	7	9	10	11	12
Risk bound	7.41×10^{-4}	0.00166	0.00369	0.0179	0.0856	0.402	0.869	1.87	4.02

TABLE I: Evaluation of the absolute-error risk bound from Theorem 2 (via the full formula (52)), for $\sigma = 1$, $n = 10^9$ and different dimensions. The bound produces satisfactory values up to $d = 10$ but quickly deteriorates for larger dimensions. The exponential growth of the bound with d is also evident from the table.

Theorem 2 (Absolute-Error Risk for Bounded Support) *Fix $\sigma > 0$ and $d \geq 1$. Then*

$$\sup_{P\in\mathcal{F}_d} \mathbb{E}_{S^n}\left|h(P*\varphi_\sigma) - \hat{h}_{n,\sigma}\right| \leq O_\sigma\left(\frac{(\log n)^{\frac{d}{4}}}{\sqrt{n}}\right). \tag{3}$$

The proof of Theorem 2 is given in Section V-C. While the theorem is stated in asymptotic form, a full expression, with all constants explicit, is given as part of the proof (see (52)). Table 1 evaluates this bound with $n = 10^9$ samples and up to $d = 10$. Several things to note about the result are the following:

1) The theorem does not assume any smoothness conditions on the distributions in $\mathcal{F}_d$. This is possible 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 $\mathsf{T}_{\varphi_\sigma}(P) \triangleq h(P*\varphi_\sigma)$, which is smooth.
2) The result does not rely on P being bounded away from zero. We circumvent the need for such an assumption by observing that although the convolved density $P*\varphi_\sigma$ can be arbitrarily close to zero, it is easily lower bounded inside $\mathcal{R}_n \triangleq [-1,1]^d + \mathcal{B}_d(0, \sigma\sqrt{(2+\epsilon)\log n})$ (i.e., a Minkowski sum of $[-1,1]^d$ with a d-dimensional sphere of radius $\sigma\sqrt{(2+\epsilon)\log n}$). The analysis inside the region exploits the $t\log\left(\frac{1}{t}\right)$ modulus of continuity for the map $x \mapsto x\log x$ combined with some functional optimization arguments; the integral outside the region is controlled using tail bounds for the Chi-squared distribution.

Theorem 2 provides convergence rates when estimating differential entropy (or mutual information) over DNNs with bounded activation functions, such as tanh or sigmoid. To account for networks with unbounded nonlinearities, such as the popular ReLU networks, the following theorem gives a more general result of estimation over the nonparametric class $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ of d-dimensional distributions with sub-Gaussian marginals.

Theorem 3 (Absolute-Error Risk for sub-Gaussian Distributions) *Fix $\sigma > 0$ and $d \geq 1$. Then*

$$\sup_{P\in\mathcal{F}_{d,K}^{(\mathsf{SG})}} \mathbb{E}_{S^n}\left|h(P*\varphi_\sigma) - \hat{h}_{n,\sigma}\right| \leq O_\sigma\left(\frac{(\log n)^{\frac{d}{4}}}{\sqrt{n}}\right). \tag{4}$$

The proof of Theorem 3 is given in Section V-D. Again, while (4) only states the asymptotic behavior of the risk, an explicit expression is given in (59) at the end of the proof. The derivation relies on the decomposition of the absolute-error and the technical lemmas employed in the proof of Theorem 2. The main difference is the analysis of the probability that $S + Z \sim P*\varphi_\sigma$ exceeds $\mathcal{R}_n$, which is taken here as the d-dimensional hypercube $\left[-c_1\sqrt{\log n}, c_1\sqrt{\log n}\right]^d$ with $c_1 = \sqrt{\frac{2(K+\sigma)^2}{e-1}}$.

D. Necessary Number of Samples for Unbiased Estimation

The results of the previous subsection are in minimax form, that is, they state worst-case convergence rates of the $\hat{h}_{n,\sigma}$ over a certain nonparametric class of distributions. In practice, the true distribution may very well not be one that attains these worst-case rates, and convergence may be faster. However, while the variance of $\hat{h}_{n,\sigma}$ can be empirically evaluated using bootstrapping, there is no empirical test for the bias. Even if multiple estimates of $h(P*\varphi_\sigma)$ via $\hat{h}_{n,\sigma}$ consistently produce similar values, this does not necessarily suggest that these values are close to the true $h(P*\varphi_\sigma)$. To have a guideline to the least number of samples needed to avoid biased estimation, we present the following lower bound on the estimator bias $\left|h(P*\varphi_\sigma) - \sup_{P\in\mathcal{F}_d} \mathbb{E}_{S^n}\hat{h}_{n,\sigma}\right|$.

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.*[6] *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 the bias of $\hat{h}_{n,\sigma}$ over the class $\mathcal{F}_d$ is bounded as*

$$\left|h(P*\varphi_\sigma) - \sup_{P\in\mathcal{F}_d} \mathbb{E}_{S^n}\hat{h}_{n,\sigma}\right| \geq \log\left(\frac{k_\star^{d(1-\epsilon)}}{n}\right) - H_b(\epsilon). \tag{5}$$

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 V-E. Since $H_b(\epsilon)$ shrinks with ϵ, for sufficiently small ϵ values the lower bound from (5) 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 [5], where noisy DNNs with parameter σ 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 $2^{0.99d}$, for any conceivably relevant dimension.

E. Computing the Estimator

Evaluating $\hat{h}_{n,\sigma}$ requires computing the differential entropy of a Gaussian mixture. Although it cannot be computed in closed form, this section presents a method for approximate computation via MC integration [28]. 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 $Z \sim \varphi_\sigma$ and note that $V \triangleq C + Z \sim g$. First, rewrite $h(g)$ as follows:

$$h(g) = -\mathbb{E}\log g(V) = -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\log g(\mu_i+Z)\Big|C=\mu_i\Big] = -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\log g(\mu_i+Z), \tag{6}$$

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

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 (6) by

$$\hat{I}_{\mathsf{MC}}^{(i)} \triangleq \frac{1}{n_{\mathsf{MC}}} \sum_{j=1}^{n_{\mathsf{MC}}} \log g\left(\mu_i + Z_j^{(i)}\right), \tag{7a}$$

which produces

$$\hat{h}_{\mathsf{MC}} \triangleq \frac{1}{n} \sum_{i=1}^{n} \hat{I}_{\mathsf{MC}}^{(i)} \tag{7b}$$

as our estimate of $h(g)$. Note that since g is a mixture of n Gaussians, it can be efficiently evaluated using off the shelf KDE software packages, many of which require only $O(\log n)$ operations on average per evaluation of g.

Define the mean squared error (MSE) of $\hat{h}_{\mathsf{MC}}$ as

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \triangleq \mathbb{E}\left[\left(\hat{h}_{\mathsf{MC}} - h(g)\right)^2\right]. \tag{8}$$

We have the following bounds on the MSE for tanh/sigmoid and ReLU networks, i.e., when the support or the second moment of C is bounded, respectively.

Theorem 5 (MSE Bounds for the MC Estimator)

(i) Assume $C \in [-1,1]^d$ almost surely (i.e., tanh / sigmoid networks), then

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \le \frac{1}{n \cdot n_{\mathsf{MC}}} \frac{2d(2+\sigma^2)}{\sigma^2}. \tag{9}$$

(ii) Assume $M_C \triangleq \mathbb{E}\|C\|_2^2 < \infty$ (e.g., ReLU networks with weight regularization), then

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \le \frac{1}{n \cdot n_{\mathsf{MC}}} \frac{9d\sigma^2 + 8(2+\sigma\sqrt{d})M_C + 3(11\sigma\sqrt{d}+1)\sqrt{M_C}}{\sigma^2}. \tag{10}$$

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

III. Applications for Deep Neural Networks

A main application of the developed theory is estimating the mutual information between selected groups of neurons in DNNs. Much attention was recently devoted to this task [1]–[5], mostly motivated by the Information Bottleneck (IB) theory for DNNs [6], [7]. The theory tracks the mutual information pair $\big(I(X;T), I(Y;T)\big)$, where X is the DNN's input (i.e., the feature), Y is the true label and T is the hidden activity. An intriguing claim from [7] is that the mutual information $I(X;T)$ undergoes a so-called 'compression' phase as the DNN's training progresses. Namely, after a short 'fitting' phase at the beginning of training (during which $I(Y;T)$ and $I(X;T)$ both grow), $I(X;T)$ exhibits a slow long-term decrease, which, according to [7], explains the excellent generalization performance of DNNs. The main caveat in the supporting empirical results provided in [7] (and the partially opposing results from the followup work [1]) is that in a deterministic DNN the mapping $T = f(X)$ is almost always injective when the activation functions are strictly monotone. As a result, $I(X;T)$ is either infinite (when

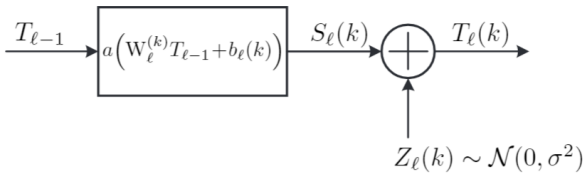


Fig. 1: k-th noisy neuron in a fully connected or a convolutional layer ℓ with activation function a; $\mathrm{W}_\ell^{(k)}$ and $b_\ell(k)$ are the k-th row and the k-th entry of the weight matrix and the bias vector, respectively.

the data distribution P_X is continuous) or a constant (when P_X is discrete[7]). Thus, when the DNN is deterministic, $I(X;T)$ is not an informative quantity to consider. As explained in [5], the reason [7] and [1] miss this fact stems from an inadequate application of a binning-based mutual information estimator for $I(X;T)$.

As a remedy for this constant/infinite mutual information issue, [5] 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. Interestingly, although the primary purpose of the noise injection in [5] was to ensure that $I(X;T)$ is a meaningful quantity, experimentally it was found that the DNN's performance is optimized at non-zero noise variance, thus providing a natural way for selecting this parameter. In the following, we first properly 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 [5]. Consider the learning problem of the feature-label pair $(X,Y) \sim P_{X,Y}$, where $P_{X,Y}$ is the (unknown) true distribution of (X,Y). The labeled dataset $\left\{(X_i,Y_i)\right\}_{i=1}^n$ comprises n i.i.d. samples from $P_{X,Y}$.

An $(L+1)$-layered noisy DNN for learning this model has layers $T_0, T_1, \ldots, T_L$, with input $T_0 = X$ and output $T_L = \hat{Y}$ (i.e., the output is 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.

[7]The mapping from a discrete set of X values 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.

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 each $\ell \in [L-1]$, the PDF of T_ℓ or any of its conditional versions is denoted by a lowercase p with the appropriate subscripts (e.g., p_{T_ℓ} is the PDF of T_ℓ, while $p_{T_\ell|X}$ is its conditional PDF given X). For any $\ell \in [L-1]$, consider the mutual information between the hidden layer and the input (see Remark 4 for an account of $I(Y;T_\ell)$):

$$I(X;T_\ell) = h(T_\ell) - h(T_\ell|X) = h(p_{T_\ell}) - \int dP_X(x) h(p_{T_\ell|X=x}). \tag{11}$$

Since $p_{T_\ell|X}$ has 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 (11), 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 $h(P * \varphi_\sigma)$ based on samples from P and knowledge of φ_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,

$$h(p_{T_\ell}) = h(P_{S_\ell} * \varphi_\sigma) \tag{12a}$$

and

$$h(p_{T_\ell|X=x_i}) = h(P_{S_\ell|X=x_i} * \varphi_\sigma), \tag{12b}$$

where φ_σ is the pdf $\mathcal{N}(0, \sigma^2 \mathrm{I}_{d_\ell})$. The DNN's generative model 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 $\boldsymbol{X}$:** 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 $S_\ell^n(X_i)$.[8]

The knowledge of φ_σ and these generated samples S_ℓ^n and $S_\ell^n(X_i)$ can be used to estimate the unconditional and the conditional entropies, from (12a) and (12b), 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 $h(P * \varphi_\sigma)$ for P supported inside

[8]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 (14) is modified by replacing the subtracted term with $h(Z)$.

$[-1,1]^d$ (i.e., a tanh / sigmoid network), based on i.i.d. samples $A^n = (A_1, \ldots, A_n)$ from P and knowledge of φ_σ. Assume that $\hat{h}$ attains

$$\sup_{P:\ \mathrm{supp}(P)\subseteq[-1,1]^d} \mathbb{E}_{A^n}\left|h(P*\varphi_\sigma) - \hat{h}(A^n,\sigma)\right| \le \Delta_{\sigma,d}(n). \tag{13}$$

An example of such an $\hat{h}$ is the estimator $\hat{h}_{n,\sigma}$ from (2); the corresponding $\Delta_{\sigma,d}(n)$ term is given in Theorem 2. Our estimator for the mutual information is

$$\hat{I}_{\mathsf{Input}}\left(X^n,\hat{h},\sigma\right) \triangleq \hat{h}(S^n,\sigma) - \frac{1}{n}\sum_{i=1}^{n}\hat{h}\big(S^n(X_i),\sigma\big). \tag{14}$$

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

Proposition 1 (Input-Hidden Layer Mutual Information Estimation Error) *For the above described estimation setting, we have*

$$\sup_{P_X}\mathbb{E}\left|I(X;T) - \hat{I}_{\mathsf{Input}}\left(X^n,\hat{h},\sigma\right)\right| \le 2\Delta_{\sigma,d}(n) + \frac{d\log\left(1+\frac{1}{\sigma^2}\right)}{4\sqrt{n}}. \tag{15}$$

Interestingly, the quantity $\frac{1}{\sigma^2}$ is the signal-to-noise ratio (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 for estimating $I(X;T)$. The bound depends only on the dimension of T (through $\Delta_{\sigma,d}$). This is because the additive noise resides in the T domain, limiting the possibility of encoding the rich structure of X into T in full. On a technical level, 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. In turn, this reduces the impact of X on the estimation of $I(X;T)$ to that of an empirical average converging to its expected value with rate $\frac{1}{\sqrt{n}}$.

Remark 3 (The sub-Gaussian Class $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ and Noisy ReLU DNNs) *We provide performance guarantees for the plugin estimator also over the more general class $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ of distributions with sub-Gaussian 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 sub-Gaussian. To see this recall that linear combinations of independent sub-Gaussian random variables are also sub-Gaussian. Furthermore, for any (scalar) random variable A, we have that $\left|\mathsf{ReLU}(A)\right| = \left|\max\{0,A\}\right| \le |A|$, 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 $\left\|S(i)\right\|_{\psi_2}$ by a constant, provided that the input X is coordinate-wise sub-Gaussian. This constant will depend on the network's weights and biases,*

the depth of the hidden layer, the sub-Gaussian norm of the input $\|X\|_{\psi_2}$ 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 [1], [5], [7]. Such a discrete distribution satisfies the required input sub-Gaussianity assumption.

Remark 4 (Mutual Information Between Hidden Layer and Label) *Another information-theoretic quantity of possible interest is the mutual information between the hidden layer and the true label (see, e.g., [7]). Let (X, Y) be a feature-label pair distributed according to $P_{X,Y}$. If T is a hidden layer in a noisy DNN with input X, the joint distribution of (X, Y, S, T) is $P_{X,Y}P_{S,T|X}$, under which $Y - X - (S, T)$ forms a Markov chain (in fact, the Markov chain is even $Y - X - S - T$ since $T = S + Z$ but this is inconsequential here). The mutual information of interest is then*

$$I(Y;T) = h(P_S * \varphi_\sigma) - \sum_{y\in\mathcal{Y}} P_Y(y) h(P_{S|Y=y} * \varphi_\sigma), \tag{16}$$

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 from (16)*, while approximating the expectation by an empirical average. There are several required modifications in 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 (a Binomial random variable). In appendix A, the process of estimating $I(Y;T)$ is described in detail and a bound on the estimation error is derived.*

This section, and, in particular, the result of Proposition 1 (see also Proposition 2 from Appendix A) show that the performance in estimating mutual information depends on our ability to estimate $h(P * \varphi_\sigma)$. In Section IV-B we present experimental results for $h(P * \varphi_\sigma)$, when P is induced by a DNN.

IV. Comparison to Past Works on Differential Entropy Estimation

In the considered estimation setup, one could always sample φ_σ and add the obtained noise samples to S^n, thus producing a sample set from $P * \varphi_\sigma$. This set can be used to estimate $h(P * \varphi_\sigma)$ via general-purpose differential entropy estimators, such as those based on kNN or KDE techniques. In the following we theoretically and empirically compare the performance of $\hat{h}_{n,\sigma}$ to state-of-the-art instances of these two methods.

A. Comparison of Theoretical Results

The main thing to note here is that convergence guarantees commonly found in the literature for KDE- and kNN-based differential estimation methods do not apply in the considered setup. Most past risk analyses [8], [9], [13], [14], [16], [21]–[25] rely on the distribution being bounded away from zero, an assumption that is violated by $P * \varphi_\sigma$. The only two works we are aware of that drop this assumption are [10], [18], the first for a KDE-based method and the second for the kNN-based KL estimator [11], assume that the density is supported inside $[0, 1]^d$, satisfies periodic boundary conditions and has a (Lipschitz or Hölder) smoothness parameter $s \in (0, 2]$. The convolved density $P * \varphi_\sigma$ does not satisfy the first two conditions. It is noteworthy that the analysis from [10] was

also extended to sub-Gaussian densities supported on the entire Euclidean space. This extension is applicable for estimation $h(P*\varphi_\sigma)$ based on samples from $P*\varphi_\sigma$, but as explained next, the obtained risk convergences slowly when d is large and is unable to exploit the smoothness of $P*\varphi_\sigma$ due to the $s\leq 2$ restriction.

Because $\hat{h}_{n,\sigma}$ is constructed to exploit the particular structure of our genie-aided estimation setup it achieves a fast convergence rate of $O\left(\frac{(\log n)^{\frac{d}{4}}}{\sqrt{n}}\right)$. The risk associated with unstructured differential entropy estimators typically converges as the slower[9] $O\left(n^{-\frac{\alpha}{\beta+d}}\right)$, where α,β are relatively small constants. In particular, the sub-Gaussian result from [10] upper bounds the risk by an $O\left((n\log n)^{-\frac{s}{s+d}}\cdot(\log n)^{\frac{d}{2}\left(1-\frac{d}{p(s+d)}\right)}+n^{-\frac{1}{2}}\right)$ term, where $0<s\leq 2$ is the Lipschitz smoothness and $2\leq p<\infty$ is a norm parameter. This rate (as well as the ones from previous works) is too slow to guarantee satisfactory estimation accuracy even for moderate dimensions, especially when taking into account the (possibly huge) multiplicative constants this asymptotic expression hides. This highlights the advantage of ad-hoc estimation as opposed to an unstructured approach.

B. Simulations

In the following we present empirical results illustrating the convergence of the $\hat{h}_{n,\sigma}$ estimator compared it to two such state-of-the-art methods: the KDE-based estimator of [8] and KL estimator from [11], [18].

1) Simulations for Differential Entropy Estimation: The 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 $[-1,1]^d$. 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 $h(P*\varphi_\sigma)=h(P*\varphi_\sigma)$ is estimated for various values of σ.

Estimation results as a function of the number of samples n are shown for dimensions $d=5$ and $d=10$ in Fig. 2, and for dimension $d=15$ in Fig. 3. The kernel width for the KDE estimate was chosen via cross-validation, varying with both d and n; the kNN estimator and our $\hat{h}_{n,\sigma}$ require no tuning parameters. Observe that the KDE estimate is rather unstable and, while not shown here, the estimated value is highly sensitive to the chosen kernel width (varying widely if the kernel width is perturbed from the cross-validated value). Note that both the kNN and the KDE estimators converge slowly, at a rate that degrades with increased d. This rate is significantly worse than that of our proposed estimator, which also lower bounds the true entropy (as according to our theory - see (61)). We also note that the difference between the performance of the KDE estimator and $\hat{h}_{n,\sigma}$ decreases for smaller σ. This is because for small enough σ the distribution of $S+Z$ and that of S become close, making the KDE estimator and our estimator (which bears some similarities to a KDE estimate on S directly) become more similar. However, when σ is larger, the KDE estimate does not coincide with the true entropy even for the maximal number of samples used in our simulations ($n=10^7$), for all considered dimensions. Finally, we note that in accordance to the upper bound from Theorem 2, the absolute estimation errors increase with larger d and smaller σ.

[9]This expression overlooking polylogarithmic factors that appear in many of the results

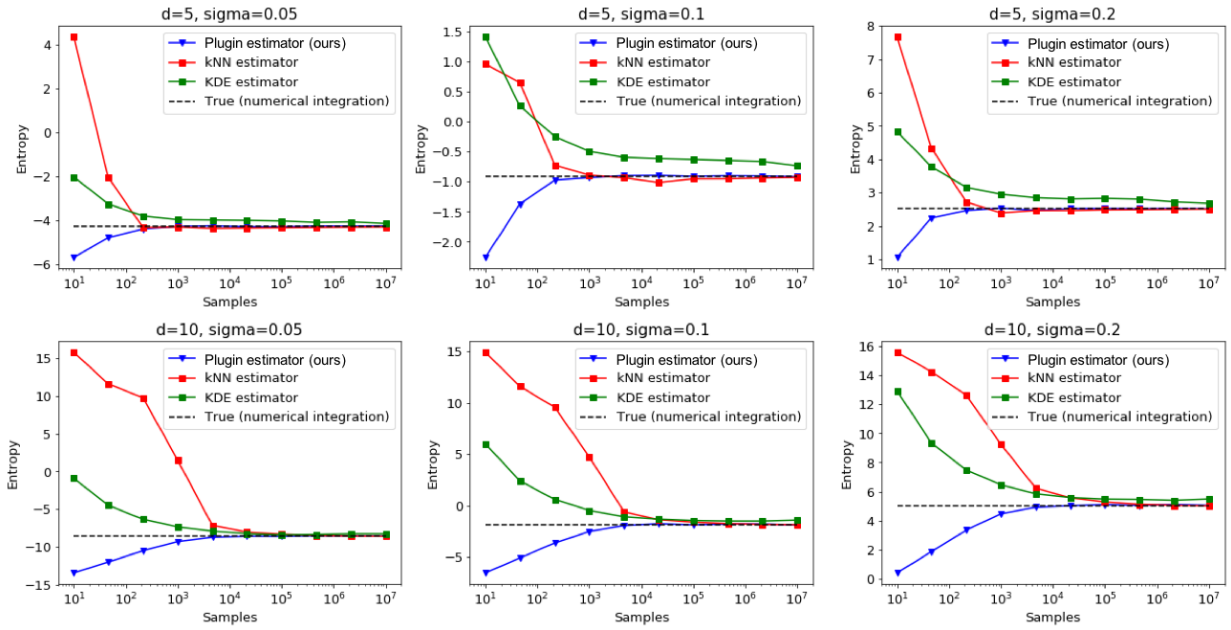


Fig. 2: Estimation results for $\hat{h}_{n,\sigma}$ compared to state-of-the-art kNN-based and KDE-based differential entropy estimators. The differential entropy of $S+Z$ is estimated, where S is a truncated d-dimensional mixture of 2^d Gaussians and $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. Results are shown as a function of n, for $d = 5, 10$ and $\sigma = 0.01, 0.1, 0.5$. The $\hat{h}_{n,\sigma}$ estimator presents faster convergence rates, improved stability and better scalability with dimension compared to the two competing methods.

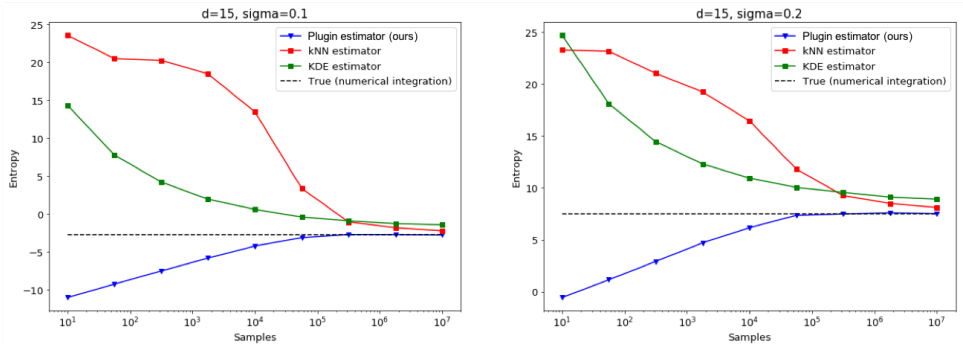


Fig. 3: Entropy estimation results for $\hat{h}_{n,\sigma}$ compared to state-of-the-art kNN-based and KDE-based differential entropy estimators. The setup is the same as Figure 2, with results shown as a function of n for $d = 15$ and $\sigma = 0.1, 0.2$.

In Fig. 4, we show the convergence rates in the unbounded support regime by considering the same setting but without truncating the 2^d-mode Gaussian mixture. Observe that a good convergence for $\hat{h}_{n,\sigma}$ is still attained, outperforming the competing methods.

2) Monte Carlo Integration: Fig. 5 illustrates the convergence of the MC integration method for computing $\hat{h}_{n,\sigma}$. 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 $\hat{P}_{S^n} * \varphi_\sigma$ whose entropy approximates $h(P * \varphi_\sigma)$), $d = 5, 10, 15$, and $\sigma = 0.01, 0.1$. Note the error decays approximately as $\frac{1}{\sqrt{n_{\mathsf{MC}}}}$

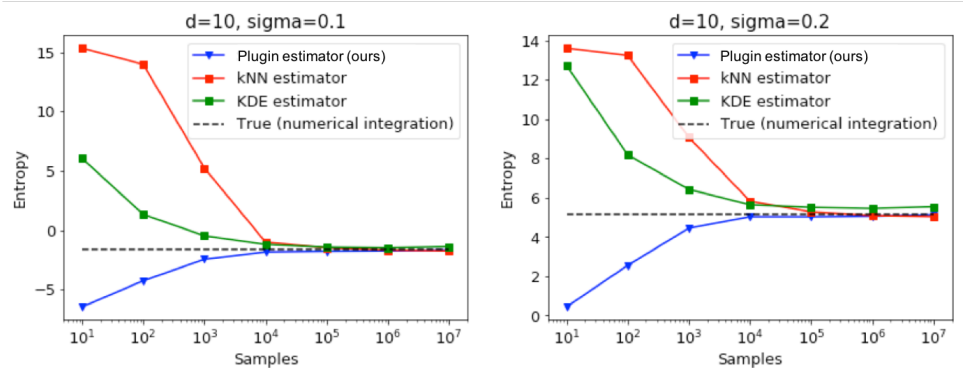


Fig. 4: Estimation results for the $\hat{h}_{n,\sigma}$ estimator, the kNN-based estimator and the KDE-based estimator in the unbounded support regime. Estimation of $h(S+Z)$ is considered, where S is a d-dimensional mixture of 2^d Gaussians and $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. Results for $d = 10$ and $\sigma = 0.1, 0.2$ are presented.

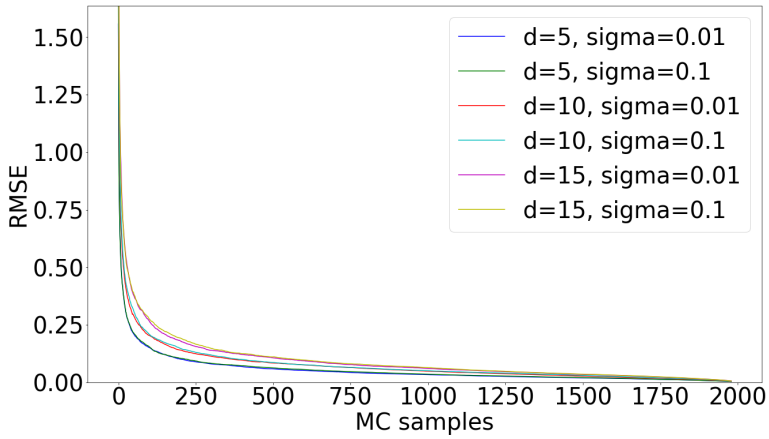
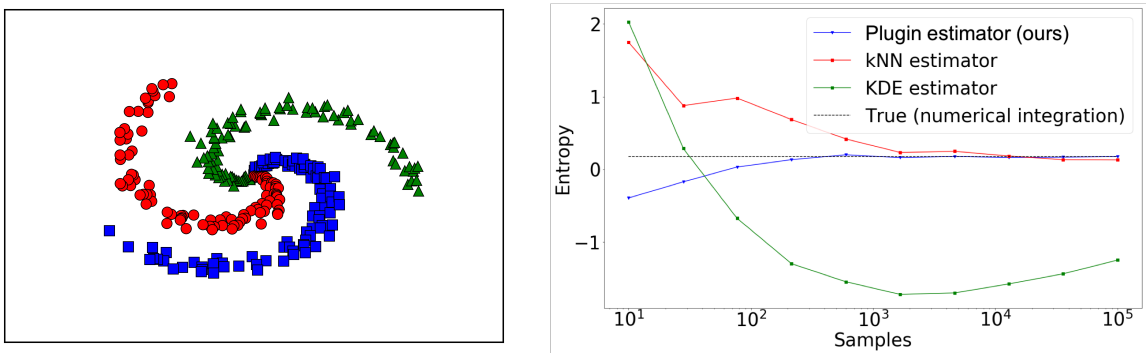


Fig. 5: Convergence of the Monte Carlo integrator computation of the proposed estimator. Shown is the decay of the RMSE as the number of Monte Carlo samples increases, for a variety of σ and d values. The MC integrator is computing the $\hat{h}_{n,\sigma}$ estimate of the entropy of $S+Z$ where S is 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 $\hat{h}_{n,\sigma}$ is 10^4.

in accordance with Theorem 5, and that the convergence does not vary excessively for different d and σ values.

3) 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. 6(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. 6(b), comparing our method to the kNN and KDE estimators. As before, our method converges faster than the competing methods illustrating its efficiency for entropy and mutual information estimation over noisy DNNs. Observe that the KDE estimate is particularly poor in this regime. Indeed, KDE is known to be not well-suited for high-dimensional problems and to underperform on distributions with widely-varying smoothness characteristics (as in these nonlinear-activation DNN hidden layer distributions). In our companion work [5], extensive additional examples of mutual information estimation in DNN

classifiers based on the proposed estimator are provided.



(a) (b)

Fig. 6: 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 $\hat{h}_{n,\sigma}$ compared to state-of-the-art kNN-based and KDE-based differential entropy estimators 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 $\sigma = 0.2$. Like in previous example, our estimator convergences faster and is more stable compared to the two competing methods.

4) Mutual Information of Reed-Mullar Codes over AWGN Channels: Consider data transmission over an AWGN channel via a binary phase-shift keying (BPSK) modulation of an error-correcting Reed-Muller code. Denote a Reed-Muller code of parameters $r, m \in \mathbb{N}$, where $0 \leq r \leq m$ by $\mathsf{RM}(r,m)$. An $\mathsf{RM}(r,m)$ code 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, e.g., 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

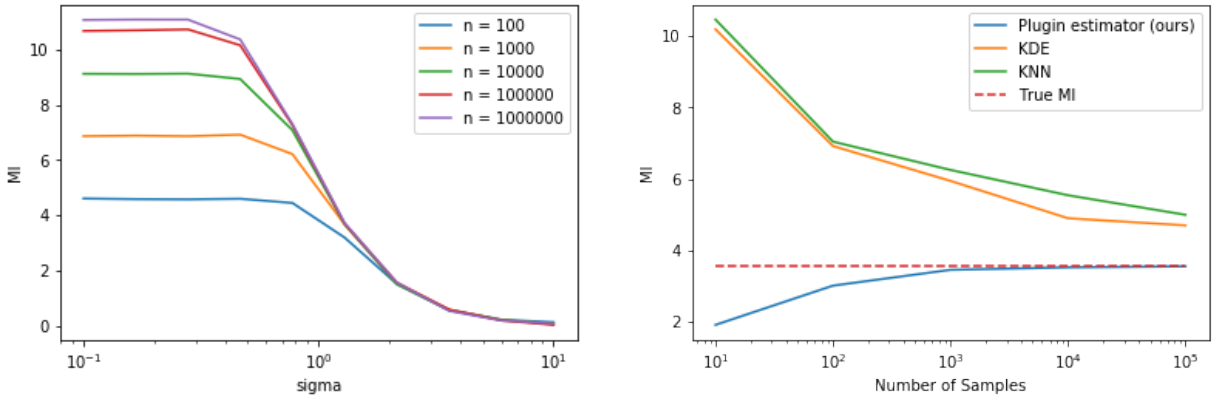
$$I(S; S+Z) = h(S+Z) - \frac{d}{2}\log(2\pi e\sigma^2), \tag{17}$$

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

Using our estimator for differential entropy under Gaussian convolutions, $I(S;S+Z)$ can be readily estimated based on samples of S. The estimation 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. 7 for various values of σ and samples n of S used for estimation. In Fig. 7(a) we plot our estimate of $I(S;S+Z)$ for an $\mathsf{RM}(4,4)$ code as a function of σ, for different values of n. This subfigure also shows that, as expected, $\hat{h}_{n,\sigma}$ converges faster for larger values of σ. Fig. 7(b) shows the estimated $I(S;S+Z)$ for $S \sim \mathsf{Unif}(\mathcal{C}_{\mathsf{RM}(5,5)})$ and $\sigma = 2$, with kNN- and KDE-based estimates based on samples of $(S+Z)$ shown for comparison. Our method significantly outperforms the general-purpose estimators, without requiring any tuning parameters. Of course that given more empirical samples the kNN and KDE estimates would have converged to the truth. Nonetheless, their relatively poor performance for the considered number of samples samples (up to $n = 10^5$) highlights the advantage our estimator enjoys being tailored for the AWGN scenario.

Remark 5 (Calculating the Ground Truth) *To compute the true value of $I(S;S+Z)$ in Fig. 7(b) (dashed red line) we used our MC integrator and the fact the Reed-Muller code was known to us (upon generating it). Specifically,*

the distribution of $S+Z$ is a Gaussian mixture, whose differential entropy we compute via the expression from (7). *Convergence of the computed value was ensure using Theorem 5.*



(a) (b)

Fig. 7: Estimation of the mutual information between a BPSK modulated Reed-Muller code and its output from an AWGN channel of noise parameter σ. (a) On the left we show the estimated mutual information as a function of noise σ and the number of samples n for the $\mathsf{RM}(4,4)$ code. (b) On the right, we show the $\hat{h}_{n,\sigma}$-based estimated mutual information for the 32-dimensional Reed Muller code $\mathsf{RM}(5,5)$ and $\sigma=2$ as a function of the number of samples 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).

V. Proofs

A. Proof of Proposition 1

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

$$I(X;T) = h(T) - h(T|X) = h(P_S * \varphi_\sigma) - \mathbb{E}g(X). \tag{18}$$

Applying the triangle inequality to (14) we obtain

$$\begin{aligned}
&\mathbb{E}\left|\hat{I}_{\mathsf{Input}}\left(X^n, \hat{h}, \sigma\right) - I(X;T)\right| \\
&\quad\le \mathbb{E}\left|\hat{h}(S^n,\sigma) - h(P_S * \varphi_\sigma)\right| + \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \hat{h}\big(S^n(X_i),\sigma\big) - \mathbb{E}g(X)\right|, \\
&\quad\le \underbrace{\mathbb{E}\left|\hat{h}(S^n,\sigma) - h(P_S * \varphi_\sigma)\right|}_{\text{(I)}} + \underbrace{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left|\hat{h}\big(S^n(X_i),\sigma\big) - g(X_i)\right|}_{\text{(II)}} + \underbrace{\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}g(X)\right|}_{\text{(III)}}
\end{aligned} \tag{19}$$

By assumption (13) and because $\mathrm{supp}(P_S) \subseteq [-1,1]^d$, we have

$$\mathbb{E}\left|\hat{h}(S^n,\sigma) - h(P_S * \varphi_\sigma)\right| \le \Delta_{\sigma,d}(n). \tag{20}$$

Similarly, for any fixed $X^n = x^n$, $\mathrm{supp}(P_{S|X=x_i}) \subseteq [-1,1]^d$ for all x_i, where $i \in [m]$, and hence

$$\mathbb{E}\left[\left|\hat{h}(S^n(X_i),\sigma) - g(X_i)\right| \middle| X^n = x^n\right] \overset{(a)}{=} \mathbb{E}\left|\hat{h}\big(S^n(x_i),\sigma\big) - h(P_{S|X=x_i} * \varphi_\sigma)\right| \le \Delta_{\sigma,d}(n), \tag{21}$$

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 21 and the law of total expectation we have

$$\text{(II)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left|\hat{h}(S^n(X_i)) - g(X_i)\right| \le \Delta_{\sigma,d}(n). \tag{22}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g(X) \tag{23}$$

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

$$\begin{aligned}
\text{(III)}^2 &= \left(\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g(X)\right|\right)^2 \\
&\le \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g(X)\right)^2\right] \\
&= \frac{1}{n}\mathsf{var}\big(g(X)\big) \\
&\le \frac{1}{4n}\left(\sup_x h(p_{T|X=x}) - \inf_x h(p_{T|X=x})\right)^2.
\end{aligned} \tag{24}$$

The last inequality follows since $\mathsf{var}(A) \le \frac{1}{4}(\sup A - \inf A)^2$ 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 $x \in \mathbb{R}^{d_0}$

$$h(p_{T|X=x}) = h(S+Z|X=x) \ge h(S+Z|S,X=x) = h(Z) = \frac{d}{2}\log(2\pi e\sigma^2), \tag{25}$$

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

$$h(p_{T|X=x}) = h(T|X=x) \le \sum_{k=1}^{d} h\big(T(k)\big|X=x\big), \tag{26}$$

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

$$\mathsf{var}\big(T(k)\big|X=x\big) \le \mathbb{E}\big[T^2(k)\big|X=x\big] \le 1+\sigma^2, \tag{27}$$

since $S(k) \in [-1,1]$ almost surely. For a fixed variance the Gaussian distribution maximizes differential entropy, and therefore

$$h(p_{T|X=x}) \le \frac{d}{2}\log\big(2\pi e(1+\sigma^2)\big). \tag{28}$$

for all $x \in \mathbb{R}^{d_0}$. Substituting the lower bound (25) and upper bound (28) into (24) gives

$$(\mathrm{III})^2 \leq \left(\frac{d \log\left(1+\frac{1}{\sigma^2}\right)}{4\sqrt{n}} \right)^2. \tag{29}$$

Inserting this along with (20) and (22) into the bound (19) bounds the expected estimation error as

$$\mathbb{E}\left| \hat{I}_{\mathsf{Input}}\left(X^n, \hat{h}, \sigma\right) - I(X;T) \right| \leq 2\Delta_n + \frac{d \log\left(1+\frac{1}{\sigma^2}\right)}{4\sqrt{n}}. \tag{30}$$

Taking the supremum over P_X concludes the proof.

B. Proof of Theorem 1

1) Proof of the 1st Part: 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

$$\mathsf{C}_{\mathsf{AWGN}}(\sigma) = \max_{X \sim P:\ \mathrm{supp}(P) \subseteq [-1,1]} I(X;Y), \tag{31}$$

which is positive for any $\sigma < \infty$. The positivity of capacity implies the following [29]: 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 $\epsilon \in \left(0, \mathsf{C}_{\mathsf{AWGN}}(\sigma)\right)$, there exists a codebook $\mathcal{C}_d \subset [-1,1]^d$ of size $|\mathcal{C}_d| \doteq e^{d(\mathsf{C}_{\mathsf{BI-AWGN}}(\sigma)-\epsilon)}$ and a decoding function $\psi_d : \mathbb{R}^d \to [-1,1]^n$ such that

$$\mathbb{P}\Big(\psi_d(Y^d) = c \Big| X^d = c\Big) \geq 1 - e^{-\epsilon^2 d}, \quad \forall c \in \mathcal{C}_d, \tag{32}$$

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 (32) ensures an exponentially decaying error probability for any $c \in \mathcal{C}_d$, we also have that the error probability 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 $\mathcal{C}_d$. Denoting $\hat{X}^d \triangleq \psi_d(Y^d)$, we have

$$\mathbb{P}\big(X^d \neq \hat{X}^d\big) = \sum_{c \in \mathcal{C}_d} P(c) \mathbb{P}\Big(\psi_d(c + N^d) \neq c \Big| X^d = c\Big) \leq e^{-\epsilon^2 d}. \tag{33}$$

Based on (33), Fano's inequality implies

$$H\left(X^d \middle| \hat{X}^d\right) \leq H_b\left(e^{-\epsilon^2 d}\right) + e^{-\epsilon^2 d} \log |\mathcal{C}_d| \triangleq \delta^{(1)}_{\sigma,d}, \tag{34}$$

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

$$I\left(X^d; Y^d\right) = H\left(X^d\right) - H\left(X^d \middle| Y^d\right) \overset{(a)}{\geq} H\left(X^d\right) - H\left(X^d \middle| \hat{X}^d\right) \overset{(b)}{\geq} H\left(X^d\right) + \delta^{(1)}_{\sigma,d}, \tag{35}$$

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 (34).

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:

$$\Big|H(X^d) - I(X^d;Y^d)\Big| \leq \delta^{(1)}_{\sigma,d}. \tag{36}$$

This means that any good estimator (within an additive gap) of $H(X^d)$ over the class of distributions $\big\{P\,\big|\,\mathrm{supp}(P) = \mathcal{C}_d\big\}$ (which is included in $\mathcal{F}_d$) is also a good estimator of the mutual information. Using the well-known lower bound on the sample complexity of discrete entropy estimation (see, e.g., [27, Proposition 3]), we have that estimating $H(X^d)$ within a sufficiently small additive gap $\eta > 0$ requires at least

$$\Omega\left(\frac{|\mathcal{C}_d|}{\eta \log|\mathcal{C}_d|}\right) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right), \tag{37}$$

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 as follows. Expanding the mutual information the other way around, we have

$$I(X^d;Y^d) = h(X^d + N^d) - h(N^d) = h(X^d + N^d) - \frac{d}{2}\log_2(2\pi e\sigma^2). \tag{38}$$

Letting $S \sim P$ and noting that $Z \overset{\mathcal{D}}{=} N^d$, where $\overset{\mathcal{D}}{=}$ stands for 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 a subexponential (in d) number of samples and achieves an additive gap $\eta > 0$ over the class $\big\{P\,\big|\,\mathrm{supp}(P) = \mathcal{C}_d\big\}$, would imply that $H(X^d)$ can be estimated from these samples within gap $\eta + \delta^{(1)}_{\sigma,d}$. This follows from (36) by taking the estimator of $h(S + Z)$ and subtracting the constant $\frac{d}{2}\log_2(2\pi e\sigma^2)$. This is a contradiction.

2) Proof of the 2nd Part: 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 $\mathrm{supp}(P) = \mathcal{C}$. 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

$$H(X|Y) \leq H\big(X\big|\psi_{\mathcal{C}}(Y)\big) \leq H_b\big(\mathsf{P}_{\mathsf{e}}(\mathcal{C})\big) + \mathsf{P}_{\mathsf{e}}(\mathcal{C}) \cdot \log|\mathcal{C}|, \tag{39}$$

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 a $y \in \mathbb{R}^d$ 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 quadrant that contains c. We have:

$$\mathsf{P}_{\mathsf{e}}(\mathcal{C}) = \sum_{c\in\mathcal{C}} P(c)\mathbb{P}\Big(\psi_{\mathcal{C}}(c + Z) \neq c\Big|X = c\Big) = 1 - \left(1 - Q\left(\frac{1}{\sigma}\right)\right)^d \triangleq \epsilon_{\sigma,d}, \tag{40}$$

where Q is the Q-function. Together, (39) and (40) give $H(X|Y) \leq H_b(\epsilon_{\sigma,d}) + \epsilon_{\sigma,d} d\log 2 \triangleq \delta^{(2)}_{\sigma,d}$. Note that for any $d \geq 1$, $\lim_{\sigma\to 0}\delta^{(2)}_{\sigma,d} = 0$ exponentially fast in $\frac{1}{\sigma^2}$ (this follows from the large x approximation of $Q(x)$). Similarly to (36), the above implies that

$$\Big|H(X) - I(X;Y)\Big| \leq \delta^{(2)}_{\sigma,d}. \tag{41}$$

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 the results on sample complexity of discrete entropy estimation for small additive gaps [27, Proposition 3]). 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 (41)) 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

$$\Omega\left(\frac{\mathrm{supp}(P)}{\left(\eta + \delta^{(2)}_{\sigma,d}\right)\log\left(\mathrm{supp}(P)\right)}\right) = \Omega\left(\frac{2^d}{\left(\eta + \delta^{(2)}_{\sigma,d}\right)d}\right) \tag{42}$$

samples. The result is extended to a class of continuous distributions which is a subset of $\bar{\mathcal{F}}_d$ via a Gaussian splitting argument like the one used for proving Theorem 1.

C. Proof of Theorem 2

The analysis bounds the estimation error inside and outside a certain high probability region with respect to $q \triangleq P * \varphi_\sigma$. Inside the high probability region we use the modulus of continuity $t\log\left(\frac{1}{t}\right)$ for the function $x \mapsto x\log x$ to dominate the difference between certain integrals. Outside the region, the estimation error is controlled via bounds on the tail probability of the Chi-squared distribution.

Define $\mathcal{R}_n \triangleq [-1,1]^d + \mathcal{B}(0,\alpha_n\sigma)$ as the Minkowski sum of the hypercube and a ball of radius $\alpha_n\sigma$, where $\alpha_n > 1$ will be specified later. For a PDF q we denote $h_{\mathcal{R}_n}(q) \triangleq -\int_{\mathcal{R}_n} q(x)\log q(x)dx$ and define $h_{\mathcal{R}_n^c}(q)$ analogously with respect to the complement of $\mathcal{R}_n$. We have

$$\begin{aligned}
&\sup_{P\in\mathcal{F}_d} \mathbb{E}_{S^n}\left|h(P*\varphi_\sigma) - h(\hat{P}_{S^n}*\varphi_\sigma)\right| \\
&\qquad \le \sup_{P\in\mathcal{F}_d} \mathbb{E}_{S^n}\left|h_{\mathcal{R}_n}(P*\varphi_\sigma) - h_{\mathcal{R}_n}(\hat{P}_{S^n}*\varphi_\sigma)\right| + 2\cdot\sup_{P\in\mathcal{F}_d}\left|h_{\mathcal{R}_n^c}(P*\varphi_\sigma)\right|.
\end{aligned} \tag{43}$$

Accordingly, we only need to control the estimation inside $\mathcal{R}_n$ and show that $\left|h_{\mathcal{R}_n^c}(P*\varphi_\sigma)\right|$ is small for any $P \in \mathcal{F}_d$ with a proper choice of α_n. The first term on the RHS of (43) is controlled using the following Lemma.

Lemma 1 (Entropy Restricted to Finite Volume Set) *Let $\mathcal{R} \subset \mathbb{R}^d$ be a set of finite Lebesgue measure. Then for all n sufficiently large, we have*

$$\sup_{P\in\mathcal{F}_d} \mathbb{E}_{S^n}|h_{\mathcal{R}}(P*\varphi_\sigma) - h_{\mathcal{R}}(\hat{P}_{S^n}*\varphi_\sigma)| \le \frac{1}{2(4\pi\sigma^2)^{\frac{d}{4}}}\log\left(\frac{n\lambda(\mathcal{R})}{(\pi\sigma^2)^{\frac{d}{2}}}\right)\sqrt{\frac{\lambda(\mathcal{R})}{n}}, \tag{44}$$

where λ is the Lebesgue measure on $\mathbb{R}^d$.

The lemma is proven in Appendix B based on the aforementioned $t\log\left(\frac{1}{t}\right)$ modulus of continuity for the function $x \mapsto x\log x$ and functional optimization arguments. Invoking the lemma for the region $\mathcal{R}_n$, the estimation error inside $\mathcal{R}_n$ is bounded as:

$$\sup_{P\in\mathcal{F}_d} \mathbb{E}_{S^n}\left|h_{\mathcal{R}_n}(q) - h_{\mathcal{R}_n}(r_{S^n})\right| \le \frac{1}{2(4\pi\sigma^2)^{\frac{d}{4}}}\log\left(\frac{n(2+2\sigma\alpha_n)^d}{(\pi\sigma^2)^{\frac{d}{2}}}\right)(2+2\sigma\alpha_n)^{\frac{d}{2}}\frac{1}{\sqrt{n}}, \tag{45}$$

which follows because $\mathcal{R}_n \subseteq [-1-\alpha_n\sigma, 1+\alpha_n\sigma]^d$.

The second summand on the RHS of (43) is handled using Lemma 2, whose proof is found in Appendix C.

Lemma 2 (Entropy Restricted to Complement Region) *Let P be a distribution on $\mathbb{R}^d$ and $\mathcal{R} \subset \mathbb{R}^d$ be a region of finite Lebesgue measure such that $(P * \varphi_\sigma)(x) < 1$, for all $x \in \mathcal{R}^c$. Suppose $S \sim P$ satisfies $\mathbb{E}\|S\|_2^4 < \infty$ and let $T \sim P * \varphi_\sigma$. Then*

$$|h_{\mathcal{R}^c}(P * \varphi_\sigma)| \le \left(\left(c'_{\sigma,d} + \frac{\mathbb{E}\|S\|_2^2}{\sigma^2}\right)\left(c'_{\sigma,d} + \frac{3\mathbb{E}\|S\|_2^2 + 2\sigma^2 d}{\sigma^2}\right) + \frac{8\left(\mathbb{E}\|S\|_2^4 + \sigma^4 d(2+d)\right)}{\sigma^4}\right)\mathbb{P}\big(T \notin \mathcal{R}\big), \quad (46)$$

where $c'_{\sigma,d} \triangleq \frac{d}{2}\log(2\pi\sigma^2)$.

To apply Lemma 2 on the second term from the RHS of (43), fix $P \in \mathcal{F}_d$, choose any $\epsilon > 0$ and let $\alpha_n = \sqrt{(2+\epsilon)\log n}$. Observe that for sufficiently large n we have $(P * \varphi_\sigma)(x) < 1$ for all $x \in \mathcal{R}_n^c$.[10] Thus, using (46) along with $\|S\|_2 \le \sqrt{d}$, we obtain

$$|h_{\mathcal{R}_n^c}(q)| \le \left(c_{\sigma,d}^2 + \frac{2c_{\sigma,d}d(1+\sigma^2)}{\sigma^2} + \frac{8d(d+2\sigma^4+d\sigma^4)}{\sigma^4}\right)\mathbb{P}\big(T \notin \mathcal{R}_n\big), \quad (47)$$

with $c_{\sigma,d} \triangleq \frac{d}{2}\log(2\pi\sigma^2) + \frac{d}{\sigma^2}$.

Finally, we bound the probability of T exceeding $\mathcal{R}_n$. Note that for any $s \in [-1,1]^d$, we have

$$\mathbb{P}\big(T \notin \mathcal{R}_n | S = s\big) \overset{(a)}{=} \mathbb{P}\big(s + Z \notin \mathcal{R}_n\big) \le \mathbb{P}\big(Z \notin \mathcal{B}_d(0, \alpha_n\sigma)\big) \overset{(b)}{=} \mathbb{P}\big(Q > \alpha_n^2\big), \quad (48)$$

where (a) uses the independence of S and Z, and in (b) we set $Q \sim \chi_d^2$ as a random variable distributed according to the Chi-squared distribution with d degrees of freedom. To bound the tail probability of Q we use Lemma 1 from [30], which states that (in particular, see [30, Equation (4.3)])

$$\mathbb{P}\big(Q \ge d + 2\sqrt{d\gamma} + 2\gamma\big) > e^{-\gamma}, \quad (49)$$

for any $\gamma > 0$. Recalling that $\alpha_n = \sqrt{(2+\epsilon)\log n}$ and letting n be large enough so that $\alpha_n^2 > d + 2\sqrt{d\log n} + 2\log n$, (49) gives $\mathbb{P}\big(Q > \alpha_n^2\big) \le \frac{1}{n}$. Consequently, we obtain,

$$\mathbb{P}\big(T \notin \mathcal{R}_n\big) = \mathbb{E}_S\mathbb{P}\big(S + Z \notin \mathcal{R}_n \big| S\big) \le \frac{1}{n}, \quad (50)$$

and together with (46) this gives

$$|h_{\mathcal{R}_n^c}(q)| \le \left(c_{\sigma,d}^2 + \frac{2c_{\sigma,d}d(1+\sigma^2)}{\sigma^2} + \frac{8d(d+2\sigma^4+d\sigma^4)}{\sigma^4}\right)\frac{1}{n}. \quad (51)$$

Taking a supremum over all $P \in \bar{\mathcal{F}}_d$ and plugging this along with (45) into (43) gives that for any $\epsilon > 0$ and n

[10] Such an n exists uniformly in $P \in \mathcal{F}_d$ since $\operatorname{supp} P = [-1,1]^d$. Consequently, $P * \varphi_\sigma$ has sub-Gaussian tails.

sufficiently large, we have

$$\sup_{P\in\mathcal{F}_d}\mathbb{E}_{S^n}\left|h(P*\varphi_\sigma)-\hat{h}_{n,\sigma}\right| \le \frac{1}{2(4\pi\sigma^2)^{\frac{d}{4}}}\log\left(\frac{n\Big(2+2\sigma\sqrt{(2+\epsilon)\log n}\Big)^d}{(\pi\sigma^2)^{\frac{d}{2}}}\right)\Big(2+2\sigma\sqrt{(2+\epsilon)\log n}\Big)^{\frac{d}{2}}\frac{1}{\sqrt{n}} + \left(c_{\sigma,d}^2+\frac{2c_{\sigma,d}d(1+\sigma^2)}{\sigma^2}+\frac{8d(d+2\sigma^4+d\sigma^4)}{\sigma^4}\right)\frac{2}{n}, \tag{52}$$

where $c_{\sigma,d}\triangleq\frac{d}{2}\log(2\pi\sigma^2)+\frac{d}{\sigma^2}$. This concludes the proof.

D. Proof of Theorem 3

Given Lemmas 1 and 2, to prove Theorem 3 it suffices to bound the probability of $T=S+Z\sim P*\varphi_\sigma$ exceeding a region $\mathcal{R}_n\subseteq\mathbb{R}^d$ whose diameter grows as $\sqrt{\log n}$ with an $O\left(\frac{1}{n}\right)$ term. To do so we exploit the coordinate-wise sub-Gaussianity of $S=\big(S(1),\ldots,S(d)\big)$. The class of sub-Gaussian random variables on a given probability space is a normed space (with norm $\|\cdot\|_{\psi_2}$), and any sub-Gaussian random variable X satisfies (see Lemma 5.5 from [31]):

1) $\mathbb{P}\big(|X|>t\big)\le\exp\left(1-\frac{ct^2}{\|X\|_{\psi_2}^2}\right)$, for all $t\ge 0$;
2) $(\mathbb{E}|X|^p)^{1/p}\le\|X\|_{\psi_2}\sqrt{p}$, for all $p\ge 1$,

where $c=\frac{e-1}{2}$ is an absolute constant (this value of c can be extracted from the proof of Lemma 5.5 from [31]).

Now, let $\mathcal{R}_n=\left[-c_1\sqrt{\log n},-c_1\sqrt{\log n}\right]$ be the d-dimensional hypercube of side $2c_1\sqrt{\log n}$ with the constant c_1 to be specified later. First note that

$$\{T\notin\mathcal{R}_n\}=\Big\{\|T\|_\infty>c_1\sqrt{\log n}\Big\}\subseteq\bigcup_{i=1}^d\Big\{\big|T(i)\big|>c_1\sqrt{\log n}\Big\}, \tag{53}$$

and therefore the union bound gives

$$\mathbb{P}\big(T\notin\mathcal{R}_n\big)\le\sum_{i=1}^d\mathbb{P}\left(\big|T(i)\big|>c_1\sqrt{\log n}\right). \tag{54}$$

By hypothesis, S has sub-Gaussian coordinates $S(i)$ with $\big\|S(i)\big\|_{\psi_2}\le K$, for all $i\in[d]$. For the Gaussian noise, we have $\big\|Z(i)\big\|_{\psi_2}\le\sigma$, for all $i\in[d]$. Since the sum of two independent sub-Gaussian random variable is also sub-Gaussian with norm that equals the sum of its components' norms, we thus obtain $\big\|T(i)\big\|_{\psi_2}\le K+\sigma$, for all $i\in[d]$. Inserting this into the sub-Gaussian tail bound from Item 1) above gives

$$\mathbb{P}\big(T\notin\mathcal{R}_n\big)\le de\cdot e^{-\frac{cc_1^2}{(K+\sigma)^2}\log n}=e\cdot\frac{d}{n}, \tag{55}$$

where the last equality follows by taking $c_1=\sqrt{\frac{2(K+\sigma)^2}{e-1}}$.

The proof is concluded by using the expansion from (43) and invoking Lemmas 1 and 2. First, we substitute $\lambda(\mathcal{R}_n)=\left(\frac{8(K+\sigma)^2}{e-1}\log n\right)^{\frac{d}{2}}$ into (44) and recast $\mathcal{F}_d$ from Lemma 1 as $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ to arrive at:

$$\sup_{P\in\mathcal{F}_{d,K}^{(\mathsf{SG})}}\mathbb{E}_{S^n}\big|h_{\mathcal{R}}(P*\varphi_\sigma)-h_{\mathcal{R}}(\hat{P}_{S^n}*\varphi_\sigma)\big|\le\log\left(\sqrt{n}\left(\frac{8(K+\sigma)^2}{(e-1)\pi\sigma^2}\log n\right)^{\frac{d}{4}}\right)\left(\frac{2(K+\sigma)^2}{(e-1)\pi\sigma^2}\log n\right)^{\frac{d}{4}}\frac{1}{\sqrt{n}}, \tag{56}$$

Then, to apply Lemma 2 we bound the second and fourth moments of $\|S\|_2$ based on the sub-Gaussianity of its coordinates. The obtained bounds are

$$\mathbb{E}\|S\|_2^2 \leq 2dK^2 \qquad ; \qquad \mathbb{E}\|S\|_2^4 \leq 16d^2K^4, \tag{57}$$

where the latter also uses the Cauchy-Schwartz inequality. Substituting these into (46) gives

$$\left|h_{\mathcal{R}_n^c}(q)\right| \leq + \left(\left(c'_{\sigma,d} + \frac{2dK^2}{\sigma^2}\right)\left(c'_{\sigma,d} + 2\frac{\sigma^2 d + 3dK^2}{\sigma^2}\right) + \frac{8\left(16d^2K^4 + \sigma^4 d(2+d)\right)}{\sigma^4}\right)\frac{ed}{n}, \tag{58}$$

with $c'_{\sigma,d} = \frac{d}{2}\log(2\pi\sigma^2)$. Taking the supremum of (58) over all $P \in \mathcal{F}_{d,K}^{(\mathsf{SG})}$ an inserting it along with (56) into (43) produces

$$\begin{aligned}\sup_{P\in\mathcal{F}_{d,K}^{(\mathsf{SG})}} \mathbb{E}_{S^n}\left|h(P*\varphi_\sigma) - \hat{h}_{n,\sigma}\right| \leq \log\left(\sqrt{n}\left(\frac{8(K+\sigma)^2}{(e-1)\pi\sigma^2}\log n\right)^{\frac{d}{4}}\right)\left(\frac{2(K+\sigma)^2}{(e-1)\pi\sigma^2}\log n\right)^{\frac{d}{4}}\frac{1}{\sqrt{n}} \\ + \left(\left(c'_{\sigma,d} + \frac{2dK^2}{\sigma^2}\right)\left(c'_{\sigma,d} + 2\frac{\sigma^2 d + 3dK^2}{\sigma^2}\right) + \frac{8\left(16d^2K^4 + \sigma^4 d(2+d)\right)}{\sigma^4}\right)\frac{ed}{n},\end{aligned} \tag{59}$$

where $c'_{\sigma,d} \triangleq \frac{d}{2}\log(2\pi\sigma^2)$.

Remark 6 (Relationship Between the Bounded Support and the sub-Gaussian Results) *As mentioned in Remark 3, the class $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ is rather general, and, in particular, includes $\mathcal{F}_d$ whenever $K \geq 1$. This means that Theorem 3 also provides an upper bound on the minimax risk under the setup of Theorem 2. Nonetheless, we chose to separately state Theorem 2 since the derivation under the bounded support assumption enables extracting slightly better constants (which can be important for the applications we have in mind). Furthermore, the simpler setup of distributions with bounded support is convenient for demonstrating our proof technique (Section V-C). Still, we highlight that the expressions from* (52) *and* (59) *with $K = 1$ not only have the same convergence rates, but their constants are also very close.*

Remark 7 (Near Minimax Rate-Optimality) *A convergence rate faster than $\frac{1}{\sqrt{n}}$ cannot be attained for parameter estimation under the absolute-error loss. This follows from, e.g., Proposition 1 of [32], which establishes this convergence rate as a lower bound for the parametric estimation problem. Consequently, the convergence rate of $O_{\sigma,d}\left(\frac{\mathsf{Polylog}(n)}{\sqrt{n}}\right)$ established in Theorems 2 and 3 for the $\hat{h}_{n,\sigma}$ estimator is near minimax rate-optimal (i.e., up to logarithmic factors).*

Remark 8 (Negligibility of $\frac{1}{n}$ Term) *Observe that the $\frac{1}{n}$ rate of the second term in* (59) *is arbitrary. By increasing the diameter of $\mathcal{R}_n$ by a constant factor we can improve the $\frac{1}{n}$ rate to $\frac{1}{n^\alpha}$ for any constant α we desire.*

Remark 9 (Cramér's Condition instead of sub-Gaussianity) *Our proof technique applies also for the class of distributions whose L^2-norm satisfies Cramér's condition (which is larger than $\mathcal{F}_{d,K}^{(\mathsf{SG})}$). Specifically, the L_2-norm of*

$S \sim P$ satisfies Cramér's condition if the moment generating function (MGF) of $\|S\|_2$ exists in a (possibly small) neighborhood of zero, i.e.,

$$\exists \lambda_c > 0, \quad \mathbb{E}e^{\lambda \|S\|_2} = \int e^{\lambda \|S\|_2} dP < \infty, \ \forall \lambda \in (-\lambda_c, \lambda_c). \tag{60}$$

Under this condition, the proof of Theorem 3 can be repeated almost verbatim with the main modification being that the high-probability region $\mathcal{R}_n$ needs to grow as $\log n$ rather then like $\sqrt{\log n}$. We chose to present the result under the sub-Gaussian assumption because practical scenarios of interest fall under this framework and since the obtained expressions are cleaner.

E. Proof of Theorem 4

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

$$\mathbb{E}_{S^n} \hat{h}_{n,\sigma} = \mathbb{E}_{S^n} h(\hat{P}_{S^n} * \varphi_\sigma) \leq h(P * \varphi_\sigma). \tag{61}$$

Now, let $W \sim \mathsf{Unif}([n])$ be independent of (S^n, Z) and define $Y = S_W + Z$. We have the following lemma, whose proof is found in Appendix D.

Lemma 3 *The following equality holds:*

$$h(P * \varphi_\sigma) - \mathbb{E}_{S^n} h(\hat{P}_{S^n} * \varphi_\sigma) = I(S^n; Y). \tag{62}$$

Using the lemma, we have

$$\left| \sup_{P \in \mathcal{F}_d} \mathbb{E}_{S^n} h(P * \varphi_\sigma) - \hat{h}_{n,\sigma} \right| = \sup_{P \in \mathcal{F}_d} I(S^n; Y), \tag{63}$$

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 exact 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

$$\begin{aligned} I(S^n; Y) &= I(S^n; Y, S_W) - I(S^n; S_W | Y) \\ &\overset{(a)}{\geq} I(S^n; S_W) - H(S_W | Y) \\ &= H(S_W) - H(S_W | S^n) - H(S_W | Y), \end{aligned} \tag{64}$$

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 (64)

we use Fano's inequality to obtain

$$H(S_W|Y) \le H\big(S_W\big|\psi_{\mathcal{C}}(Y)\big) \le H_b\big(\mathsf{P_e}(\mathcal{C})\big) + \mathsf{P_e}(\mathcal{C}) \cdot \log|\mathcal{C}|, \tag{65}$$

where $\psi_{\mathcal{C}} : \mathbb{R}^d \to \mathcal{C}$ is a function for decoding S_W from Y and $\mathsf{P_e}(\mathcal{C}) \triangleq \mathbb{P}\big(S_W \neq \psi_{\mathcal{C}}(Y)\big)$ 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 $\big\{y \in \mathbb{R}^d\big|\psi_{\mathcal{C}}(y) = c_i\big\}$ 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 $[-1,1]^d$. When $Y = S_W + Z$, $S_W \sim \mathsf{Unif}(\mathcal{C})$ and the probability of error for the decoder $\psi_{\mathcal{C}}$ is bounded as:

$$\begin{aligned}
\mathsf{P_e}(\mathcal{C}) &= \frac{1}{k^d}\sum_{i=1}^{k^d} \mathbb{P}\Big(\psi_{\mathcal{C}}(c_i + Z) \neq c_i \Big| S_W = c_i\Big) \\
&= \frac{1}{k^d}\sum_{i=1}^{k^d} \mathbb{P}\big(c_i + Z \notin \mathcal{D}_i\big) \\
&\overset{(a)}{\le} \mathbb{P}\left(\|Z\|_\infty > \frac{2/k}{2}\right) \\
&\overset{(b)}{=} 1 - \left(1 - 2Q\left(\frac{1}{k\sigma}\right)\right)^d,
\end{aligned} \tag{66}$$

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 (66) as given in the statement of the theorem gives the desired bound $\mathsf{P_e}(\mathcal{C}) \le \epsilon$. Collecting the pieces and inserting back to (64), we obtain

$$I(S^n; Y) \ge \log\left(\frac{k_\star^{d(1-\epsilon)}}{n}\right) - H_b(\epsilon). \tag{67}$$

Together with (63) this concludes the proof.

F. 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 is used to denote a 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:

$$\mathbb{E}\hat{h}_{\mathsf{MC}} = -\frac{1}{n \cdot n_{\mathsf{MC}}}\sum_{i=1}^{n}\sum_{j=1}^{n_{\mathsf{MC}}} \mathbb{E}\log g\left(\mu_i + Z_j^{(i)}\right) = h(g). \tag{68}$$

Therefore, the MSE expands as

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) = \frac{1}{n^2 \cdot n_{\mathsf{MC}}} \sum_{i=1}^{n} \mathsf{var}\Big(\log g(\mu_i + Z)\Big). \tag{69}$$

We next bound the variance of $\log g(\mu_i + Z)$ via Poincaré inequality for the Gaussian measure $\mathcal{N}(0, \sigma^2 \mathrm{I}_d)$ (with Poincaré constant σ^2). For each $i \in [n]$, we have

$$\mathsf{var}\Big(\log g(\mu_i + Z)\Big) \le \sigma^2 \mathbb{E}\Big[\big\|\nabla \log g(\mu_i + Z)\big\|_2^2\Big]. \tag{70}$$

We proceed with separate derivations of (9) and (10).

1) MSE Bound for Bounded Support: Since $\|C\|_2 \le \sqrt{d}$ almost surely, Proposition 3 from [33] implies

$$\big\|\nabla \log g(v)\big\|_2 \le \frac{\|v\|_2 + \sqrt{d}}{\sigma^2}. \tag{71}$$

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

$$\mathsf{var}\Big(\log g(\mu_i + Z)\Big) \le \frac{2d(4+\sigma^2)}{\sigma^2}, \tag{72}$$

for each $i \in [n]$. Together with (69), this concludes the proof of (9).

2) MSE Bound for Bounded Second Moment: To prove (10), we use Proposition 2 from [33] to obtain

$$\big\|\nabla \log g(v)\big\|_2 \le \frac{1}{\sigma^2}\big(3\|v\|_2 + 4\mathbb{E}\|C\|_2\big). \tag{73}$$

Via the Poincaré inequality from (70), the variance is bounded as

$$\begin{aligned}\mathsf{var}\Big(\log g(\mu_i + Z)\Big) &\le \frac{1}{\sigma^2}\mathbb{E}\Big[\big(3\|\mu_i + Z\|_2 + 4\mathbb{E}\|C\|\big)^2\Big] \\ &\le \frac{1}{\sigma^2}\left(9d\sigma^2 + 16M_C + 24\sigma\sqrt{dM_C} + 3\|\mu_i\|_2\left(3 + 9\sigma\sqrt{d} + 8\sigma\sqrt{dM_C}\right)\right),\end{aligned} \tag{74}$$

where the last step uses Hölder's inequality (namely, $\mathbb{E}\|C\|_2 \le \sqrt{\mathbb{E}\|C\|_2^2}$). The proof of (10) is concluded by plugging (74) into the MSE expression from (69) and noting that $\frac{1}{n}\sum_{i=1}^{n}\|\mu_i\|_2 \le \sqrt{M_C}$.

VI. Summary and Concluding Remarks

In this work we studied differential entropy estimation under Gaussian convolutions, aiming to estimate the functional $\mathsf{T}_{\varphi_\sigma}(P) = h(P * \varphi_\sigma)$ based on empirical i.i.d. samples from P and knowledge of the Gaussian noise distribution φ_σ. We are motivated to understand the decision-theoretic fundamental limits of this setup since it is the centerpiece in estimating mutual information estimation between layers of DNN [5].

We first showed that an exponential dependence of the sample complexity on the dimension is unavoidable. Then, the plugin estimator $\hat{h}_{n,\sigma} = h\big(\hat{P}_{S^n} * \varphi_\sigma\big)$ was proposed and its absolute-error risk was analyzed. We showed that its convergence rate over nonparametric distribution classes of interest is $O\left(\frac{(\log n)^{\frac{d}{4}}}{\sqrt{n}}\right)$, with all constants explicitly characterized. An ad hoc treatment of the considered estimation problem was crucial here because theoretical performance guarantees for general-purpose differential entropy estimators found in the literature (applicable by collecting samples from $P * \varphi_\sigma$) are not valid in our setup. This is since the convolved density $P * \varphi_\sigma$ fails to satisfy

many of the assumptions that such results rely on. Furthermore, the the explicit modeling of $P * \varphi_\sigma$ and the ability to sample P directly enabled: (i) establishing a faster convergence rate than the typical $O\left(n^{-\frac{\alpha}{\beta+d}}\right)$ rate attained by kNN- or KDE-based estimators; (ii) circumventing the need for smoothness or boundedness assumptions on the nonparametric class of distributions; and (iii) explicit derivation of all the constants. All three aspects are essential for implementing $\hat{h}_{n,\sigma}$ and using its error bounds to get concrete theoretical guarantees on the worst-case error of the estimated value. To facilitate the implementation of our estimator, an efficient Monte Carlo integration method was proposed for computing $\hat{h}_{n,\sigma}$ accompanied by an MSE bound on its performance. A condition on the number of samples that $\hat{h}_{n,\sigma}$ needs in order to avoid biased estimation was also provided as a guideline for choosing n. Finally, we performed a set of numerical experiments verifying the convergence and superiority of our estimator to existing approaches in the Gaussian convolution setting.

Several interesting future directions arise from this work. First, we note that our proof of the sample complexity being exponential in d applies only when either d is sufficiently large or σ is sufficiently small. An appealing goal is to extend this result to any d and σ, possibly using the generalized Le Cam's method from [10]. Furthermore, we aim to explore if the $c^d \frac{(\log n)^{\frac{d}{4}}}{\sqrt{n}}$ convergence rate established in Theorems 2 and 3 for $\hat{h}_{n,\sigma}$ can be improved to $c_1^d \frac{(\log n)^{c_2}}{\sqrt{n}}$, for some constants c_1 and c_2 that are independent of d. We conjecture that the latter is the best attainable convergence rate for the considered estimation setting. Proving this would, however, require a different method than the $\mathcal{R}_n$ integral-splitting technique we currently employ. Another possible improvement of our result is weakening the coordinate-wise sub-Gaussianity assumption on $S \sim P$ to a constraint on some moment of its Euclidean norm. A bounded fourth moment assumption may be sufficient to establish the result, but again, doing so would require a different proof. Beyond this particular work, we see considerable virtue in exploring additional ad-hoc estimation setups that have exploitable structure that would enable obtaining improved estimation results.

ACKNOWLEDGEMENT

The authors would like to thank Yihong Wu for helpful ideas concerning the proof of Theorem 2.

APPENDIX A
ESTIMATING THE MUTUAL INFORMATION BETWEEN THE LABEL AND A HIDDEN LAYER

We consider here 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 4). 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, however, that $\mathcal{Y} \triangleq \mathrm{supp}(P_Y)$ is finite and known (as is the case in any application of interest) and let $|\mathcal{Y}| = K$ be the cardinality of $\mathcal{Y}$, i.e. the number of distinct class labels. The labeled dataset $\left\{(X_i,Y_i)\right\}_{i=1}^n$ comprises n i.i.d. samples from $P_{X,Y}$. Let T be a hidden layer in a noisy DNN with input X and 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

$$I(Y;T) = h(P_S * \varphi_\sigma) - \sum_{y \in \mathcal{Y}} p_Y(y) h(P_{S|Y=y} * \varphi_\sigma), \tag{75}$$

based on a given estimator $\hat{h}$ of $h(P*\varphi_\sigma)$ that knows φ_σ and uses i.i.d. samples from P, where P is supported inside $[-1,1]^d$). In (75), p_Y is the PMF associated with P_Y.

We first describe the sampling procedure for estimating each of the differential entropies from (75). For the unconditional entropy, P_S is sampled in the same manner described in Section III-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 labeled dataset $\big\{(x_i,y_i)\big\}_{i=1}^n$ and consider the following. Define the set $\mathcal{I}_y \triangleq \big\{i \in [n] \big| y_i = y\big\}$ 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. It is readily verified that applying the appropriate deterministic function on each of these samples produces 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 III-B, suppose we are given an estimator $\hat{h}(A^m,\sigma)$ of $h(P*\varphi_\sigma)$, for P with $\mathrm{supp}(P) \subseteq [-1,1]^d$, based on m i.i.d. samples $A^m = (A_1,\ldots,A_m)$ from P. Assume that $\hat{h}$ attains

$$\sup_{P:\ \mathrm{supp}(P)\subseteq[-1,1]^d} \mathbb{E}_{A^m}\left|h(P*\varphi_\sigma) - \hat{h}(A^m,\sigma)\right| \le \Delta_{\sigma,d}(m). \tag{76}$$

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, the $\hat{h}$ estimator is bad to begin with 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

$$\hat{I}_{\mathsf{Label}}\Big(X^n,Y^n,\hat{h},\sigma\Big) \triangleq \hat{h}(S^n,\sigma) - \sum_{y\in\mathcal{Y}} \hat{p}_{Y^n}(y)\hat{h}\big(S^{n_y}(\mathcal{X}_y),\sigma\big), \tag{77}$$

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 of $\hat{I}_{\mathsf{Label}}\Big(X^n,Y^n,\hat{h},\sigma\Big)$; the proof is given after the statement.

Proposition 2 (Label-Hidden Layer Mutual Information Estimation Error) *For the above described estimation setting, we have*

$$\sup_{P_{X,Y}:\ |\mathcal{Y}|=K} \mathbb{E}\left|I(Y;T) - \hat{I}_{\mathsf{Label}}\Big(X^n,Y^n,\hat{h},\sigma\Big)\right| \le \Delta_{\sigma,d}(n) + c_{\sigma,d}^{(\mathsf{MI})}\sqrt{\frac{K-1}{n}} + K\left(\Delta_{\sigma,d}^\star \cdot e^{-\frac{np_l^2}{8p_u}} + \Delta_{\sigma,d}\left(\frac{np_l}{2}\right)\right), \tag{78}$$

where

$$c_{\sigma,d}^{(\mathsf{MI})} \triangleq \frac{d}{2}\max\Big\{-\log(2\pi e\sigma^2), \log\big(2\pi e(1+\sigma^2)\big)\Big\} \tag{79a}$$

$$p_l \triangleq \min_{y\in\mathcal{Y}} p_Y(y) \tag{79b}$$

$$p_u \triangleq \max_{y\in\mathcal{Y}} p_Y(y) \tag{79c}$$

$$\Delta_{\sigma,d}^\star \triangleq \max_{n\in\mathbb{N}} \Delta_{\sigma,d}(n). \tag{79d}$$

The proof is reminiscent of that of Proposition 1, 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

$$\mathbb{E}\Big|I(Y;T) - \hat{I}_{\mathsf{Label}}\left(X^n, Y^n, \hat{h}, \sigma\right)\Big| \leq \underbrace{\mathbb{E}\left|h(P_S * \varphi_\sigma) - \hat{h}(S^n, \sigma)\right|}_{\text{(I)}} + \underbrace{\sum_{y\in\mathcal{Y}} \left|h(P_{S|Y=y} * \varphi_\sigma)\right| \mathbb{E}\big|p_Y(y) - \hat{p}_{Y^n}(y)\big|}_{\text{(II)}} + \underbrace{\sum_{y\in\mathcal{Y}} \mathbb{E}\left|\hat{p}_{Y^n}(y)\Big(h(P_{S|Y=y} * \varphi_\sigma) - \hat{h}\big(S^{n_y}(\mathcal{X}_y), \sigma^2\big)\Big)\right|}_{\text{(III)}}, \tag{80}$$

where we have added and subtracted $\sum_{y\in Y} \hat{p}_{Y^n}(y) h(P_{S|Y=y} * \varphi_\sigma)$ 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

$$h(P_{S|Y=y} * \varphi_\sigma) = h(S + Z|Y = y) \geq h(S + Z|S, Y = y) = \frac{d}{2}\log(2\pi e\sigma^2), \tag{81}$$

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

$$h(P_{S|Y=y} * \varphi_\sigma) \leq \sum_{k=1}^{d} h\big(S(k) + Z(k)\big|Y = y\big) \leq \frac{d}{2}\log\big(2\pi e(1 + \sigma^2)\big), \tag{82}$$

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 (81) and (82) we obtain

$$\big|h(P_{S|Y=y} * \varphi_\sigma)\big| \leq c^{(\mathsf{MI})}_{\sigma,d} \triangleq \frac{d}{2}\max\Big\{-\log(2\pi e\sigma^2), \log\big(2\pi e(1 + \sigma^2)\big)\Big\}. \tag{83}$$

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

$$\mathbb{E}\big|p_Y(y) - \hat{p}_{Y^n}(y)\big| \leq \sqrt{\mathsf{var}\big(p_{Y^n}(y)\big)} = \sqrt{\frac{1}{n}\mathsf{var}\left(\mathbb{1}_{\{Y=y\}}\right)} = \sqrt{\frac{p_Y(y)\big(1 - p_Y(y)\big)}{n}}. \tag{84}$$

Using (83) and (84) we bound Term (II) as follows:

$$\text{(II)} \leq \frac{c^{(\mathsf{MI})}_{\sigma,d}}{\sqrt{n}} \sum_{y\in\mathcal{Y}} \sqrt{p_Y(y)\big(1 - p_Y(y)\big)} \leq c^{(\mathsf{MI})}_{\sigma,d}\sqrt{\frac{K-1}{n}}, \tag{85}$$

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 2. Using a Chernoff bound for the Binomial distribution we have that for any $k \leq np_Y(y)$,

$$\mathbb{P}\Big(n_y \leq k\Big) \leq \exp\left(-\frac{1}{2p_Y(y)} \cdot \frac{\big(np_Y(y) - k\big)^2}{n}\right) \leq \exp\left(-\frac{1}{2p_u} \cdot \frac{\big(np_Y(y) - k\big)^2}{n}\right). \tag{86}$$

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

$$\mathbb{P}\Big(n_y \leq k_y^\star\Big) \leq \exp\left(-\frac{np_l^2}{8p_u}\right). \tag{87}$$

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

$$\begin{aligned}
\text{(III)} &\leq \sum_{y\in\mathcal{Y}} \mathbb{E}\Big|h(P_{S|Y=y} * \varphi_\sigma) - \hat{h}\big(S^{n_y}(\mathcal{X}_y), \sigma^2\big)\Big| \\
&\overset{(a)}{=} \sum_{y\in\mathcal{Y}} \mathbb{E}_{n_y}\left[\mathbb{E}\left[\left|\hat{p}_{Y^n}(y)\Big(h(P_{S|Y=y} * \varphi_\sigma) - \hat{h}\big(S^{n_y}(\mathcal{X}_y), \sigma^2\big)\Big|\right| n_y\right]\right] \\
&\overset{(b)}{=} \sum_{y\in\mathcal{Y}} \mathbb{E}_{n_y}\Delta_{\sigma,d}(n_y) \\
&\overset{(c)}{=} \sum_{y\in\mathcal{Y}} \mathbb{P}\Big(n_y \leq k_y^\star\Big)\mathbb{E}\big[\Delta_{\sigma,d}(n_y)\big|n_y \leq k_y^\star\big] + \mathbb{P}\Big(n_y > k_y^\star\Big)\mathbb{E}\big[\Delta_{\sigma,d}(n_y)\big|n_y > k_y^\star\big] \\
&\overset{(d)}{\leq} K\left(\Delta_{\sigma,d}^\star \cdot e^{-\frac{np_l^2}{8p_u}} + \Delta_{\sigma,d}\left(\frac{np_l}{2}\right)\right),
\end{aligned} \tag{88}$$

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 (87), the definition of $\Delta_{\sigma,d}^\star$ and the fact that $\Delta_{\sigma,d}(n)$ is monotonically decreasing with n along with $k_y^\star \geq \frac{np_l}{2}$, for all $y \in \mathcal{Y}$. Inserting (I)$\leq \Delta_{\sigma,d}(n)$ together with the bounds from (85) and (88) back into (80) and taking the supremum over all $P_{X,Y}$ with $|\mathcal{Y}| = K$ concludes the proof.

■

APPENDIX B
PROOF OF LEMMA 1

For any $P \in \mathcal{F}_d$, define the shorthand $q \triangleq P * \varphi_\sigma$ and $r_{S^n} = \hat{P}_{S^n} * \varphi_\sigma$. Note that $\|q\|_\infty, \|r_{S^n}\|_\infty \leq c_1 \triangleq (2\pi\sigma^2)^{-\frac{d}{2}}$ ($P^{\otimes n}$-almost surely for the latter), since φ_σ is an isotropic Gaussian with variance σ^2. Defining $\tilde{q} = \frac{q}{c_1}$ and $\tilde{r}_{S^n} = \frac{r_{S^n}}{c_1}$, for any $P \in \mathcal{F}_d$ we have

$$\begin{aligned}
\mathbb{E}_{S^n}\big|h_{\mathcal{R}}(q) - h_{\mathcal{R}}(r_{S^n})\big| &= \mathbb{E}_{S^n}\left|\int_{\mathcal{R}} \big(q(x)\log q(x) - r_{S^n}(x)\log r_{S^n}(x)\big)dx\right| \\
&\leq \mathbb{E}_{S^n}\int_{\mathcal{R}} \big|q(x)\log\tilde{q}(x) - r_{S^n}(x)\log\tilde{r}_{S^n}(x)\big|dx + |\log c_1| \cdot \mathbb{E}_{S^n}\int_{\mathcal{R}} |q(x) - r_{S^n}(x)|dx \\
&\overset{(a)}{=} c_1 \cdot \mathbb{E}_{S^n}\int_{\mathcal{R}} \big|\tilde{q}(x)\log\tilde{q}(x) - \tilde{r}_{S^n}(x)\log\tilde{r}_{S^n}(x)\big|dx \\
&\overset{(b)}{\leq} c_1 \cdot \mathbb{E}_{S^n}\int_{\mathcal{R}} |\tilde{q}(x) - \tilde{r}_{S^n}(x)|\log\frac{1}{|\tilde{q}(x) - \tilde{r}_{S^n}(x)|}dx,
\end{aligned} \tag{89}$$

where (a) is because $\mathbb{E}_{S^n} r_{S^n}(x) = q(x)$ for all $x \in \mathbb{R}^d$, while (b) follows because $g(t) \triangleq t\log\left(\frac{1}{t}\right)$ is a modulus of continuity for the map $x \mapsto x\log x$, when $x \in [0,1]$ (see, e.g., Equation (17.27) in [34]).

Taking the supremum over all $P \in \mathcal{F}_d$ of (89) and using the concavity of g along with Jensen's inequality, we

further obtain

$$\sup_{P\in\mathcal{F}_d}\mathbb{E}_{S^n}\left|h_{\mathcal{R}}(q)-h_{\mathcal{R}}(r_{S^n})\right|\le c_1\cdot\sup_{P\in\mathcal{F}_d}\int_{\mathcal{R}}g\left(\frac{1}{c_1}\mathbb{E}_{S^n}\left|q(x)-r_{S^n}(x)\right|\right)dx, \tag{90}$$

and focus on bounding $\mathbb{E}_{S^n}\left|q(x)-r_{S^n}(x)\right|$. To do so, consider the variance of $r_{S^n}(x)$:

$$\mathsf{var}\big(r_{S^n}(x)\big)\overset{(a)}{\le}\frac{1}{n}\mathbb{E}\varphi_\sigma^2(x-S)\overset{(b)}{\le}\frac{1}{n}\frac{1}{(4\pi\sigma^2)^{\frac{d}{2}}}\mathbb{E}\varphi_{\frac{\sigma}{\sqrt{2}}}(x-S)\overset{(c)}{\le}\frac{1}{(2\pi\sigma^2)^d}\frac{1}{n}, \tag{91}$$

where (a) is because $r_{S^n}(x)=\frac{1}{n}\sum_{i=1}^n\varphi_\sigma(x-S_i)$ with i.i.d. $S_i\sim P$, in (b) we use $\varphi_{\frac{\sigma}{\sqrt{2}}}$ for the PDF of $\mathcal{N}\left(0,\frac{\sigma^2}{2}\mathrm{I}_d\right)$, while (c) follows because $\mathbb{E}\varphi_{\frac{\sigma}{\sqrt{2}}}(x-S)=(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)\le\left\|P*\varphi_{\frac{\sigma}{\sqrt{2}}}\right\|_\infty\le(\pi\sigma^2)^{-\frac{d}{2}}$.

By Hölder's inequality, we thus obtain

$$\mathbb{E}_{S^n}\left|q(x)-r_{S^n}(x)\right|\le\sqrt{\mathsf{var}\big(r_{S^n}(x)\big)}\le\frac{1}{(4\pi\sigma^2)^{\frac{d}{4}}\sqrt{n}}\sqrt{(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)}=c_n'\sqrt{(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)}, \tag{92}$$

for all $x\in\mathcal{R}$, where we have used the shorthand $c_n'\triangleq\frac{1}{(4\pi\sigma^2)^{\frac{d}{4}}\sqrt{n}}$. Assume that n is large enough so that $\frac{c_n'}{c_1}\sqrt{(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)}\in\left(0,\frac{1}{e}\right)$, for all $x\in\mathcal{R}$, which is where g is monotonically increasing.[11] Inserting back into (90) while setting $c_n\triangleq\frac{c_n'}{c_1}$, we have

$$\sup_{P\in\mathcal{F}_d}\mathbb{E}_{S^n}\left|h_{\mathcal{R}}(q)-h_{\mathcal{R}}(r_{S^n})\right|\le c_1\cdot\sup_{P\in\mathcal{F}_d}\int_{\mathcal{R}}g\Big(c_n\sqrt{(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)}\Big)dx. \tag{93}$$

We proceed by bounding the RHS above by the supremum of the integral over all densities supported on $\mathcal{R}$, which we then show is attained by the uniform density. Let $\mathcal{G}_d(\mathcal{R})$ be the set of all densities supported on $\mathcal{R}$. Now, observe that

$$\begin{aligned}\sup_{P\in\mathcal{F}_d}\int_{\mathcal{R}}g\Big(c_n\sqrt{(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)}\Big)dx&=\sup_{P\in\mathcal{F}_d}\int_{\mathcal{R}}g\Big(c_n\sqrt{(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)\cdot\mathbb{1}_{\{x\in\mathcal{R}\}}}\Big)dx\\&\overset{(a)}{\le}\sup_{P\in\mathcal{F}_d}\int_{\mathcal{R}}g\left(c_n\sqrt{\frac{\big((P*\varphi_{\frac{\sigma}{\sqrt{2}}})\cdot\mathbb{1}_{\mathcal{R}}\big)(x)}{\int_{\mathcal{R}}(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)dx}}\right)dx\\&\overset{(b)}{\le}\sup_{\gamma\in\mathcal{G}_d(\mathcal{R})}\int_{\mathcal{R}}g\Big(c_n\sqrt{\gamma(x)}\Big)dx,\end{aligned} \tag{94}$$

where (a) is since $0<\int_{\mathcal{R}}(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)dx\le1$ and g is monotonically increasing, while (b) is because for any $P\in\mathcal{F}_d$ we have $\frac{(P*\varphi_{\frac{\sigma}{\sqrt{2}}})\cdot\mathbb{1}_{\mathcal{R}}}{\int_{\mathcal{R}}(P*\varphi_{\frac{\sigma}{\sqrt{2}}})(x)dx}\in\mathcal{G}_d(\mathcal{R})$.

Note that the mapping $x\mapsto\sqrt{x}\log\left(\frac{1}{x}\right)$ is concave for $x\in[0,1]$ and that for sufficiently large n, $c_n^2\gamma(x)\in[0,1]$ for all $x\in\mathcal{R}$. Consequently,

$$\begin{aligned}\int_{\mathcal{R}}g\left(c_n\sqrt{\gamma(x)}\right)dx&=\lambda(\mathcal{R})\int_{\mathcal{R}}\frac{1}{\lambda(\mathcal{R})}g\left(c_n\sqrt{\gamma}(x)\right)dx\\&=\frac{\lambda(\mathcal{R})}{2}\mathbb{E}_{X\sim\mathsf{Unif}}\left[\sqrt{c_n^2\gamma(X)}\log\left(\frac{1}{c_n^2\gamma(X)}\right)\right]\\&\overset{(a)}{\le}\frac{c_n\sqrt{\lambda(\mathcal{R})}}{2}\log\left(\frac{\lambda(\mathcal{R})}{c_n^2}\right)\end{aligned} \tag{95}$$

[11] Clearly, this can be attained uniformly in $x\in\mathcal{R}$ due the uniform upper bound on $\mathbb{E}_{S^n}|q(x)-r_{S^n}(x)|$ from the RHS of (91).

where (a) follows by Jensen's inequality and because $\mathbb{E}_{X\sim\mathsf{Unif}}\gamma(X) = \frac{1}{\lambda(\mathcal{R})}$. The RHS of (95) is thus an upper bound on $\sup_{\gamma\in\mathcal{G}_d(\mathcal{R})}\int_{\mathcal{R}} g\Big(c_n\sqrt{\gamma(x)}\Big)dx$ (which is also achievable by $\gamma^\star = \frac{1}{\lambda(\mathcal{R})}$, i.e., the uniform distribution over the domain). Inserting this back into (93), the result follows:

$$\sup_{P\in\mathcal{F}_d}\mathbb{E}_{S^n}\big|h_{\mathcal{R}}(q) - h_{\mathcal{R}}(r_{S^n})\big| \overset{(a)}{\leq} \frac{c_n'}{2}\log\left(\frac{\lambda(\mathcal{R})}{c_n^2}\right)\sqrt{\lambda(\mathcal{R})} = \frac{1}{2(4\pi\sigma^2)^{\frac{d}{4}}}\log\left(\frac{n\lambda(\mathcal{R})}{(\pi\sigma^2)^{\frac{d}{2}}}\right)\sqrt{\frac{\lambda(\mathcal{R})}{n}}, \tag{96}$$

where we have used the fact that $c_n = \frac{c_n'}{c_1} = \frac{(\pi\sigma^2)^{\frac{d}{4}}}{\sqrt{n}}$.

Appendix C
Proof of Lemma 2

Since $(P*\varphi_\sigma)(x) < 1$ for all $x\in\mathcal{R}^c$, we have

$$0 < \log\frac{1}{q(x)} = -\log\mathbb{E}\varphi_\sigma(x-S) \overset{(a)}{\leq} \frac{d}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}\|x-S\|_2^2 \overset{(b)}{\leq} c_{\sigma,d}' + \frac{\mathbb{E}\|S\|_2^2 + \|x\|_2^2}{\sigma^2}, \tag{97}$$

where (a) is Jensen's inequality, while (b) uses $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$ and defines $c_{\sigma,d}' \triangleq \frac{d}{2}\log(2\pi\sigma^2)$. Letting $T\sim q = P*\varphi_\sigma$ and denoting $A \triangleq c_{\sigma,d}' + \frac{\mathbb{E}\|S\|_2^2}{\sigma^2}$, we obtain

$$\begin{aligned}
|h_{\mathcal{R}^c}(q)| &= \int_{\mathcal{R}^c} q(x)\log\frac{1}{q(x)}dx \\
&= \mathbb{E}\left[\log\frac{1}{q(T)}\mathbb{1}_{\{T\notin\mathcal{R}\}}\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\left(\log\frac{1}{q(T)}\right)^2\right]\mathbb{P}\big(T\notin\mathcal{R}\big) \\
&\overset{(b)}{\leq} \mathbb{E}\left[A^2 + 2A\frac{\|T\|_2^2}{\sigma^2} + \frac{1}{\sigma^4}\|T\|_2^4\right]\mathbb{P}\big(T\notin\mathcal{R}\big) \\
&\overset{(c)}{\leq} \left(A^2 + 2A\frac{\mathbb{E}\|S\|_2^2 + \sigma^2 d}{\sigma^2} + \frac{8\left(\mathbb{E}\|S\|_2^4 + \sigma^4 d(2+d)\right)}{\sigma^4}\right)\mathbb{P}\big(T\notin\mathcal{R}\big),
\end{aligned} \tag{98}$$

where (a) follows from the Cauchy-Schwarz inequality, (b) uses (97), while (c) is because $\mathbb{E}\|T\|_2^2 = \mathbb{E}\|S\|_2^2 + \sigma^2 d$, $\|a+b\|_2^4 \leq 8\|a\|_2^4 + 8\|b\|_2^4$ and $\mathbb{E}\|Z\|_2^4 = 2\sigma^4 d + \sigma^4 d^2$.

Appendix D
Proof of Lemma 3

We expand $I(S^n;Y) = h(Y) - h(Y|S^n)$ and denote by F_A the cumulative distribution function (CDF) of a random variable A. Let $T = S + Z \sim P*\varphi_\sigma$ and first note that

$$F_Y(y) = \mathbb{P}\big(S_W + Z \leq y\big) = \frac{1}{n}\sum_{i=1}^n \mathbb{P}(S_i + Z \leq y) = F_T(y). \tag{99}$$

Thus, $h(Y) = h(P*\varphi_\sigma)$.

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

$$F_{Y|S^n}(y|s^n) = \mathbb{P}\big(S_W + Z \leq y\big|S^n = s^n\big) = \frac{1}{n}\mathbb{P}\big(s_i + Z \leq y\big), \tag{100}$$

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

REFERENCES

[1] 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.

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

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

[4] 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.

[5] Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, "Estimating information flow in neural networks," Submitted to *the International Conference on Learning Representations (ICLR-2019)*, vol. New Orleans, Louisiana, US, May 2019, arxiv link: https://arxiv.org/abs/1810.05728.

[6] 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.

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

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

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

[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] T. B. Berrett, R. J. Samworth, and M. Yuan, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *arXiv preprint arXiv:1606.00304*, 2016.

[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] 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.

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

[21] P. Hall, "Limit theorems for sums of general functions of m-spacings," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 96, no. 3, pp. 517–532, Nov. 1984.

[22] H. Joe, "Estimation of entropy and other functionals of a multivariate density," *Annals of the Institute of Statistical Mathematics*, vol. 41, no. 4, pp. 683–697, Dec. 1989.

[23] B. Y. Levit, "Asymptotically efficient estimation of nonlinear functionals," *Problemy Peredachi Informatsii*, vol. 14, no. 3, pp. 65–72, 1978.

[24] P. Hall and S. C. Morton, "On the estimation of entropy," *Annals of the Institute of Statistical Mathematics*, vol. 45, no. 1, pp. 69–88, Mar. 1993.

[25] H. F. E. Haje and Y. Golubev, "On entropy estimation by m-spacing method," *Journal of Mathematical Sciences*, vol. 163, no. 3, pp. 290–309, Dec. 2009.

[26] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[27] 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.

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

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

[30] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Stat.*, pp. 1302–1338, Oct. 2000.

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

[32] 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.

[33] 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.

[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991.