Differential Entropy Estimation under Gaussian Noise

Ziv Goldfeld
MIT
zivg@mit.edu

Kristjan Greenewald
IBM Cambridge Research Center
Kristjan.H.Greenewald@ibm.com

Yury Polyanskiy
MIT
yp@mit.edu

Yihong Wu
Yale University
yihong.wu@yale.edu

***Abstract*—There is a recent growing interest in measuring mutual information between the data X and an internal representation T of a deep neural network (DNN). In particular, the evolution of $I(X;T)$ during training attracted much attention in the context of the Information Bottleneck theory. However, in deterministic networks with strictly monotone nonlinearities (e.g., tanh or sigmoid) $I(X;T)$ is either a constant independent of the network's parameters (discrete X) or infinite (continuous X), making the mutual information a vacuous quantity. A possible remedy for this issue is the recently proposed paradigm of *noisy* DNNs, where the outputs of the hidden activities are perturbed by (small) Gaussian noises, making the $X \mapsto T$ map a stochastic parameterized channel. This work focuses on the nonparametric differential entropy estimation problem that arises in this setup: the estimation of $h(S+Z)$, where S is the sampled variable while Z is an isotropic Gaussian with known parameters. Our main motivation it to provide estimation techniques and error bounds that are applicable in practice for real-life DNNs. We first show that the sample complexity of any good estimator must scale exponentially with dimension. Then, a natural estimator for $h(S+Z)$ is proposed which approximates it via a the entropy of a Gaussian mixture. A convergence rate of $O\left(\frac{(\log n)^{d/4}}{\sqrt{n}}\right)$ is derived for the absolute-error risk, with all constants explicit and the dependence on dimension and noise parameters made clear. We observe that (i) the inherent smoothness of the convolved distribution does not require any additional smoothness assumptions on the nonparametric class of distributions, and (ii) our explicit modeling of S and Z allows avoiding the undesirable $O\left(n^{-\frac{\alpha s}{\beta s+d}}\right)$ convergence rates that are typical under unstructured smoothness assumptions, with s being a smoothness parameter and $\alpha, \beta \in \mathbb{N}$. A Monte Carlo integration method for efficient computation of the estimator is proposed and theoretical guarantees on the accuracy of the computed values are provided. Finally, several simulations illustrate the superiority of our estimator over general-purpose differential entropy estimators for the considered model, including an experiment over a small noisy DNN.**

I. Introduction

Estimating the differential entropy of an unknown distribution P from independently and identically distributed (i.i.d.) samples from P is by now a well-studied statistical estimation problem. It is an instance of the general framework of nonparametric functional estimation, where one aims estimate a functional of the underlying distribution P (in our case, the entropy), rather than the distribution itself. Specifically, suppose P is supported on $\mathbb{R}^d$, absolutely continuous with respect to (w.r.t.) the Lebesgue measure and has density p. Letting $\{X_1, \ldots, X_n\}$ be n i.i.d. samples from p, the goal is to estimate the differential entropy

$$h(p) \triangleq -\int_{\mathbb{R}^d} p(x) \log p(x) dx$$

from the empirical observations $\{X_1, \ldots, X_n\}$. While estimating the entropy of discrete distributions P is well understood with known sharp bounds on the minimax risk [1], [2] and estimators attaining the minimax rates [2]–[4], differential entropy estimation remains a more challenging task.

There are two prevailing approaches for estimating the nonsmooth differential entropy functional: the first relying on kernel density estimators (KDEs) [5], and the other using k nearest neighbor (kNN) techniques (see, e.g., [6] for a comprehensive survey). Analyzing the performance of differential entropy estimators typically requires restricting attention to smooth nonparametric density classes and assuming the underlying densities are bounded away from zero. Various works apply only for densities that are uniformly bounded away from zero [5], [7], while others restrict the densities' closeness to zero on average [8], [9]. In practice, however, even the Gaussian distribution violates the boundedness from below assumptions, rendering this restriction highly unnatural. Even more so, these estimators often have the associated risk converging as $O\left(n^{-\frac{\alpha s}{\beta s+d}}\right)$, where d is the dimension, s is a smoothness parameter[1], and α, β are positive integers. This convergence rate quickly deteriorates with larger dimensions, becoming ineffective for bounding the error of implemented estimators in high-dimensional settings. Furthermore, the above results include implicit constants that depend on d (possibly exponentially) that may (when combined with the weak decay w.r.t. n) significantly increase the number of samples required to achieve a desired estimator accuracy.

Our work is motivated by the problem of mutual information estimation over deep neural networks (DNNs), where the dimension of the ambient space is inherently large. There has been a recent surge of interest in this estimation scenario [12]–[15] partially driven by the Information Bottleneck (IB) theory for DNNs [16]. An intriguing claim from [16] is that the mutual information $I(X;T)$, between the network's input X and a given hidden layer T, undergoes the so-called 'com-

This work was supported in part by the National Science Foundation CAREER award under grant agreement CCF-12-53205 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370

[1]Commonly, the values of s are assumed to be rather small; e.g., in the recent works [10], where the kNN-based estimator from [11] was analyzed without the boundedness from below assumption on the densities, the results hold from $s \in [0, 2)$.

pression' phase as the DNN's training progresses. Namely, after a short 'fitting' phase at the beginning of training, $I(X;T)$ exhibits a slow long-term decrease, which, according to [16], explains the excellent generalization performance of DNNs. The main caveat in the supporting empirical results provided in [16] (and the partially opposing results from the followup work [12]) 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 the data distribution P_X is continuous) or a constant (when P_X is discrete[2]). Thus, when the DNN is deterministic, $I(X;T)$ is not an informative quantity to consider. As explained in [17], the reason why [12], [16] miss this fact stems from an inadequate application of the binning-based mutual information estimator used in their plots of the evolution of $I(X;T)$ during training.

As a remedy for the constant/infinite mutual information issue, [17] proposed the framework of noisy DNNs, where 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 [17] was to ensure that $I(X;T)$ is a meaningful quantity, experimentally they find that DNN's performance is optimized at non-zero noise variance, thus providing a natural way for selecting this parameter. Adopting the noisy DNN framework from [17], our goal is to set the groundwork for estimating $I(X;T)$ in real-life DNNs while providing theoretical guarantees that are not vacuous when d is relatively large. We exploit the structure of the noisy DNN to alleviate the above described deficiencies of generic (KDE- or kNN-based) differential entropy estimators for high-dimensional data. Specifically, the structure allows removing any smoothness or boundedness assumptions on the sampled density[3], while attaining a significantly faster convergence rate as described below.

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. 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 boils down to the following new differential entropy estimation problem (see Section III of the full paper [18]): 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 $h(S + Z) = h(P * \varphi)$, where φ denotes the Gaussian probability density function (PDF).[4] To investigate the decision-theoretic fundamental limit, we consider the minimax absolute-error risk of differential entropy estimation:

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

where $\mathcal{F}_d$ is a nonparametric class of d-dimensional probability distributions and $\hat{h}$ is the estimator. The sample complexity $n_d^\star(\eta,\sigma)$ is the smallest number of samples (up to constant factors) for which estimation within an additive gap η is possible. The of this work is to (i) provide lower bounds on the sample complexity of this estimation problem in terms of d, σ and η, and (ii) design an estimator $\hat{h}$ that attains the minimax rates up to polylogarithmic factors.

We first show an unavoidable exponential dependence of the sample complexity on dimension. Specifically, it is established that $n_d^\star(\eta,\sigma) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma)$ is a positive, monotonic decreasing function of σ. Furthermore, it is known that the parametric estimation rate under the absolute-error loss cannot decay at a rate faster than $\frac{1}{\sqrt{n}}$ (see, e.g., [19, Proposition 1]). To achieve this lower bound up to polylogarithmic factors, we propose an estimator that approximates $h(P * \varphi)$ via the differential entropy of a Gaussian mixture with centers at the sample points $\{S_i\}_{i=1}^n$. We first construct 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, and then consider the estimator

$$\hat{h}_{\mathsf{SP}} \triangleq h\left(\hat{P}_{S^n} * \varphi\right). \quad (2)$$

The subscript SP stands for 'sample propagation', which stands for our technique for sampling S in DNN settings. When P belongs to a class of compactly supported distributions on $\mathbb{R}^d$ (corresponding to a tanh/sigmoid DNN), we show that the absolute error of this estimator is bounded by $C_{\sigma,d}\frac{(\log n)^{d/2}}{\sqrt{n}}$, with the constant $C_{\sigma,d}$ (that also depends on d exponentially) explicitly characterized. The full version of this work [18] contains an extension of this result to the class of d-dimensional distribution with subgaussian marginals (which accounts for ReLU DNNs with subgaussian inputs). The derived convergence rate coincides with the minimax lower bound up to polylogarithmic factors making it near minimax rate-optimal. More importantly, it significantly improves upon the $O\left(n^{-\frac{\alpha s}{\beta s+d}}\right)$ convergence guarantees of generic differential entropy estimators. This is, of course, expected since $\hat{h}_{\mathsf{SP}}$ is tailored for our particular estimation setup, while generic KDE- or kNN-based estimators are not designed to exploit the $T = S + Z$ structure nor the 'clean' samples S^n.

Finally, we provide the groundwork for practical implementations of the SP estimator. An efficient implementations of $\hat{h}_{\mathsf{SP}}$ based on Monte Carlo (MC) integration is proposed. Since $\hat{h}_{\mathsf{SP}}$ is simply the entropy of a known Gaussian mixture, MC integration using samples from this mixture allows a

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

[3]In fact, the boundedness from below assumption is not valid for noisy DNNs since the Gaussian density can get arbitrarily close to 0. Therefore, many of the above mentioned minimax results do not apply for our entropy estimation framework.

[4]See the notation section at the end of the introduction for a preciese definition of $P * \varphi$ when P is discrete / continuous / mixed.

simple computation of $\hat{h}_{\mathsf{SP}}$. We provide bounds on 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[5], n_{MC} is the number of MC samples, and $C^{(\mathsf{MC})}_{\sigma,d}$ is an explicit constant that depends linearly on the dimension. MSE bounds are provided both for compactly supported distributions p (tanh/sigmoid networks), as well as distributions with a bounded second moment (e.g., ReLU network with weight regularization). Several simulations (including an estimation experiment over a small DNN for classifying a spiral dataset) visualize the gain of the ad-hoc $\hat{h}_{\mathsf{SP}}$ estimator over its general-purpose counterparts, both in the rate of error decay and in its scalability with dimension.

Notations: Throughout this work logarithms are taken w.r.t. the natural base. For an integer $k \geq 1$, we set $[k] \triangleq \big\{i \in \mathbb{Z} \big| 1 \leq i \leq k\big\}$. For a real number $p \geq 1$, the L^p-norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|_p = \Big(\sum_{i=1}^d x^p(i)\Big)^{1/p}$, while $\|x\|_\infty = \max_{1 \leq i \leq d} |x(i)|$. Probability 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.

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

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

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 whose PDF is denoted by φ. 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 $P * \varphi$. By the latter we mean $(P * \varphi)(x) = \int_{\mathbb{R}^d} p(u)\varphi(x-u)du = (p * \varphi)(x)$, when P is continuous with density p. If P is discrete with PMF p, then $(P * \varphi)(x) = \sum_{u:\ p(u)>0} p(u)\varphi(x-u)$. For a mixed distribution P, Lebesgue's decomposition theorem allows to write $P * \varphi$ as the sum of two expressions as above. Henceforth, we typically overlook the exact structure of $P * \varphi$ only mentioning it when it is consequential.

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

II. Results for Differential Entropy Estimation under Gaussian Convolutions

A. Preliminary Definitions

Let $\mathcal{F}_d$ be the set of distributions P with $\mathrm{supp}(P) \subseteq [-1,1]^d$.[6] The minimax absolute-error risk over $\mathcal{F}_d$ is

$$\mathcal{R}^\star_d(n,\sigma) \triangleq \inf_{\hat{h}} \sup_{P \in \mathcal{F}_d} \mathbb{E}_{S^n} \Big| h(P * \varphi) - \hat{h}(S^n, \sigma) \Big|, \quad (3)$$

where $\hat{h}$ is an estimator of $h(P * \varphi)$ based on the empirical data $S^n = (S_1, \ldots, S_n)$ of i.i.d. samples from P and the noise parameter σ^2. The sample complexity $n^\star_d(\eta, \sigma)$ is defined as the smallest number of samples (up to constant factors) for which estimation within an additive gap η is possible. Namely,

$$n^\star_d(\eta,\sigma) \triangleq \min\big\{n \big| \mathcal{R}^\star_d(n,\sigma) \leq \eta\big\}. \quad (4)$$

In the full version of this work [18] we show that the sample complexity is exponential in d. The argument relates the estimation of $h(P * \varphi)$ to estimating the discrete entropy of a random variable distributed over a capacity achieving codebook for the peak-constrained additive white Gaussian noise (AWGN) channel. See Theorems 1 and 2 of [18].

B. Absolute-Error Risk Convergence Rates

We turn to analyze the performance of the SP estimator from (2). Recall that $\hat{h}_{\mathsf{SP}} \triangleq h\Big(\hat{P}_{S^n} * \varphi\Big)$, where $\hat{P}_{S^n} \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{S_i}$ is the empirical measure associated with S^n. The following theorem shows that the expected absolute error of $\hat{h}_{\mathsf{SP}}$ 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 aforementioned the sample complexity lower bounds.

Theorem 1 (Absolute-Error Risk for Bounded Support) *Fix $\sigma > 0$, $d \geq 1$ and any $\epsilon > 0$. The absolute-error risk of the SP estimator* (2) *over the class $\mathcal{F}_d$, for all n sufficiently large, is bounded as*

$$\begin{aligned}
&\sup_{P \in \mathcal{F}_d} \mathbb{E}_{S^n} \Big| h(P * \varphi) - \hat{h}_{\mathsf{SP}} \Big| \\
&\leq \log\left(\frac{n\Big(2 + 2\sigma\sqrt{(2+\epsilon)\log n}\Big)^d}{(\pi\sigma^2)^{\frac{d}{2}}} \right) \frac{\Big(2 + 2\sigma\sqrt{(2+\epsilon)\log n}\Big)^{\frac{d}{2}}}{2(4\pi\sigma^2)^{\frac{d}{4}}} \frac{1}{\sqrt{n}} \\
&+ \left(c^2_{\sigma,d} + \frac{2c_{\sigma,d} d(1+\sigma^2)}{\sigma^2} + \frac{8d(d + 2\sigma^4 + d\sigma^4)}{\sigma^4} \right) \frac{2}{n}, \quad (5)
\end{aligned}$$

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

$$\sup_{P \in \mathcal{F}_d} \mathbb{E}_{S^n} \Big| h(P * \varphi) - \hat{h}_{\mathsf{SP}} \Big| = O_{\sigma,d}\left(\frac{\mathsf{Polylog}(n)}{\sqrt{n}} \right), \quad (6)$$

and the right-hand sides (RHSs) of (5) *and* (6) *are, respectively, explicit and implicit upper bounds on the minimax absolute-error risk $\mathcal{R}^\star_d(n,\sigma)$.*

[6]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.

An outline of the proof is given in Section III; see Section V-D of [18] for the full derivation. Several things to note about the result are the following:

1) The theorem does not assume any smoothness conditions on the distributions in $\mathcal{F}_d$ due to the inherent smoothing introduced by the convolution with the Gaussian density. Another way to understand this is that while the differential entropy $h(q)$ is not a smooth functional of the underlying density q, our functional is $\mathsf{T}_\varphi(P) \triangleq h(P*\varphi)$, 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 $P*\varphi$ 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 or 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 calculus of variations techniques; the integral outside the region is controlled using tail bounds for the Chi-squared distribution.
3) In relation to general-purpose differential entropy estimators, one could always sample φ and add up these noise sample to S^n to obtain a sample set from $P*\varphi$. These samples can be used to get a proxy of $h(P*\varphi)$ via a kNN- or a KDE-based differential entropy estimator. However, as mentioned above, $P*\varphi$ violated the boundedness away from zero assumption that most of the convergence rate results in the literature rely on. The only result we are aware of that analyses a differential entropy estimator (namely, the kNN-based estimator from [11]) without assuming the density is bounded from below [10] relies on the density being supported inside $[0,1]^d$, satisfying periodic boundary conditions and having a Hölder smoothness parameter $s \in (0,2]$. The convolved density $P*\varphi$ satisfies neither of these three conditions.
4) Because the SP estimator is constructed to exploit the particular structure of our estimation setup it achieves a fast convergence rate of $\left(\frac{\mathsf{Polylog}(n)}{\sqrt{n}}\right)$. The risk associated with unstructured differential entropy estimators typically converges as the slower $O\left(n^{-\frac{\alpha s}{\beta s+d}}\right)$. This highlights the advantage of ad-hoc estimation as opposed to general-purpose estimation.

Remark 1 (Extension of Theorem 1) *Theorem 1 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 full paper [18] includes an extension of Theorem 1 the nonparametric class of d-dimensional distributions with subgaussian marginals (see Theorem 4 therein).*

Remark 2 (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 [19], which establishes this convergence rate as a lower bound for the parametric estimation problem given n i.i.d. samples. Consequently, the convergence rate of $O_{\sigma,d}\left(\frac{\mathsf{Polylog}(n)}{\sqrt{n}}\right)$ established in Theorem 1 for the SP estimator is near minimax rate-optimal (i.e., up to logarithmic factors).*

C. Computing the Samples Propagation Estimator

Evaluating $\hat{h}_{\mathsf{SP}}$ 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 [20]. To simplify the presentation, we present the method for an arbitrary Gaussian mixture without referring to the notation of the estimation setup.

Let $g(t) \triangleq \frac{1}{n}\sum_{i=1}^n \varphi(t-\mu_i)$ 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 $Z \sim \varphi$ and note that $V \triangleq C + Z \sim g$. First note that

$$h(g) = -\mathbb{E}\log g(V) = -\frac{1}{n}\sum_{i=1}^n \mathbb{E}\log g(\mu_i + Z), \qquad (7)$$

where the last 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. For each $i \in [n]$, we estimate the i-th summand on the RHS of (7) by

$$\hat{I}_{\mathsf{MC}}^{(i)} \triangleq \frac{1}{n_{\mathsf{MC}}}\sum_{j=1}^{n_{\mathsf{MC}}} \log g\left(\mu_i + Z_j^{(i)}\right), \qquad (8)$$

which produces $\hat{h}_{\mathsf{MC}} \triangleq \frac{1}{n}\sum_{i=1}^n \hat{I}_{\mathsf{MC}}^{(i)}$ as our estimate of $h(g)$. Define the mean squared error (MSE) of $\hat{h}_{\mathsf{MC}}$ as $\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \triangleq \mathbb{E}\left[\left(\hat{h}_{\mathsf{MC}} - h(g)\right)^2\right]$. 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 2 (MSE Bounds for MC Computation)

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

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

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

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

A full proof of Theorem 2 is found in [18, Section V-G]. The argument exploits the Gaussian Poincaré inequality to reduce the analysis to that of the log-mixture distribution gradient. The bounds on the MSE scale only linearly with

the dimension d. Experimentally, the dominating factor is the σ^2 in the denominator.

III. Proof Outline for Theorem 1

The analysis bounds the estimation error inside and outside a certain high probability region with respect to $q \triangleq P * \varphi$. 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 $\mathcal{R}_n$. Denoting $r_{S^n} = \hat{P}_{S^n} * \varphi$, we have

$$\begin{aligned} &\sup_{P \in \mathcal{F}_d} \mathbb{E}_{S^n} \left| h(q) - h(r_{S^n}) \right| \\ &\leq \sup_{P \in \mathcal{F}_d} \mathbb{E}_{S^n} \left| h_{\mathcal{R}_n}(q) - h_{\mathcal{R}_n}(r_{S^n}) \right| + 2 \sup_{P \in \mathcal{F}_d} \left| h_{\mathcal{R}_n^c}(q) \right|. \end{aligned} \tag{11}$$

Thus, we need to control the estimation inside $\mathcal{R}_n$ and show that $\left| h_{\mathcal{R}_n^c}(P * \varphi) \right|$ is small for any $P \in \mathcal{F}_d$. The former is is controlled using the following Lemma.

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

$$\begin{aligned} &\sup_{P \in \mathcal{F}_d} \mathbb{E}_{S^n} \left| h_{\mathcal{R}}(P * \varphi) - h_{\mathcal{R}}(\hat{P}_{S^n} * \varphi) \right| \\ &\qquad \leq \frac{1}{2(4\pi\sigma^2)^{\frac{d}{4}}} \log\left(\frac{n\lambda(\mathcal{R})}{(\pi\sigma^2)^{\frac{d}{2}}} \right) \sqrt{\frac{\lambda(\mathcal{R})}{n}}, \end{aligned} \tag{12}$$

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

The proof is omitted due to space limitations, but the derivation relies on the aforementioned $t \log\left(\frac{1}{t}\right)$ modulus of continuity for the function $x \mapsto x \log x$ and some calculus of variations arguments. The second summand on the RHS of (11) is handled using Lemma 2.

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 $(P * \varphi)(x) < 1$, for all $x \in \mathcal{R}^c$, and suppose $S \sim P$ satisfies $\mathbb{E}\|S\|_2^4 < \infty$. Then*

$$\begin{aligned} &\left| h_{\mathcal{R}^c}(P * \varphi) \right| \\ &\leq \left(\left(c'_{\sigma,d} + \mathbb{E}\|S\|_2^2 \right)^2 + \frac{2\left(c'_{\sigma,d} + \mathbb{E}\|S\|_2^2 \right)\left(\mathbb{E}\|S\|_2^2 + \sigma^2 d \right)}{\sigma^2} \right. \\ &\qquad \left. + \frac{8\left(\mathbb{E}\|S\|_2^4 + \sigma^4 d(2+d) \right)}{\sigma^4} \right) \mathbb{P}\left(T \notin \mathcal{R} \right), \end{aligned} \tag{13}$$

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

Provided these two auxiliary results, the proof of Theorem 1 follows by taking $\alpha_n = \sqrt{(2+\epsilon)\log n}$, with an arbitrarily small $\epsilon > 0$ is. In the result of Lemma 1 we insert $\lambda(\mathcal{R}_n) \leq \left(2 + 2\sigma\sqrt{(2+\epsilon)\log n} \right)^d$. The assumptions of Lemma 2 hold by noting that for sufficiently large n we have $(P * \varphi)(x) < 1$ for all $x \in \mathcal{R}_n^c$, uniformly in $P \in \mathcal{F}_d$. The moments of $\|S\|$ are bounded by using $\|S\|_2 \leq \sqrt{d}$. Finally, we show that $\mathbb{P}\left(T \notin \mathcal{R}_n \right) \leq \frac{1}{n}$ by leveraging the tail bound for the Chi-square distribution from [21, Equation (4.3)]. Combining the above described pieces establishes the result.

References

[1] G. Valiant and P. Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pages 2157–2165, 2013.

[2] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory*, 62(6):3702–3720, Jun. 2016.

[3] Y. Han, J. Jiao, and T. Weissman. Adaptive estimation of shannon entropy. In *IEEE International Symposium on Information Theory (ISIT-2016)*, pages 1372–1376, Hong Kong, China, Jun. 2015.

[4] J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory*, 61(5):2835–2885, May 2015.

[5] 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)*, pages 397–405, 2015.

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

[7] Sricharan K, R. Raich, and A. O. Hero. Estimation of nonlinear functionals of densities with confidence. *IEEE Trans. Inf. Theory*, 58(7):4135–4159, Jul. 2012.

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

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

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

[11] H. Stögbauer A. Kraskov and P. Grassberger. Estimating mutual information. *Phys. rev. E*, 69(6):066138, June 2004.

[12] 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.

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

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

[15] 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.

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

[17] Z. Goldfeld and B. Kingsbury I. Melnyk N. Nguyen Y. Polyanskiy E. van den Berg, K. Greenwald. Estimating information flow in neural networks. *arXiv preprint arXiv:*, 2018.

[18] Z. Goldfeld, K. Greenewald, Y. Polyanskiy, and Y. Wu. Differential entropy estimation under Gaussian convolutions. *arXiv preprint*, 2018.

[19] J. Chen. A general lower bound of minimax risk for absolute-error loss. *Canadian Journal of Statistics*, 25(4):545–558, Dec. 1997.

[20] Christian P Robert. *Monte Carlo Methods*. Wiley Online Library, 2004.

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