Optimality of the Plug-in Estimator for Differential Entropy Estimation under Gaussian Convolutions

Ziv Goldfeld
MIT
zivg@mit.edu

Kristjan Greenewald
IBM Research
kristjan.h.greenewald@ibm.com

Jonathan Weed
MIT
jweed@mit.edu

Yury Polyanskiy
MIT
yp@mit.edu

***Abstract*—This paper establishes the optimality of the plug-in estimator for the problem of differential entropy estimation under Gaussian convolutions. Specifically, we consider the estimation of the differential entropy $h(X+Z)$, where X and Z are independent d-dimensional random variables with $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. The distribution of X is unknown and belongs to some nonparametric class, but n independently and identically distributed samples from it are available. We first show that despite the regularizing effect of noise, any good estimator (within an additive gap) for this problem must have an exponential in d sample complexity. We then analyze the absolute-error risk of the plug-in estimator and show that it converges as $\frac{c^d}{\sqrt{n}}$, thus attaining the *parametric* estimation rate. This implies the optimality of the plug-in estimator for the considered problem. We provide numerical results comparing the performance of the plug-in estimator to general-purpose (unstructured) differential entropy estimators (based on kernel density estimation (KDE) or k nearest neighbors (kNN) techniques) applied to samples of $X+Z$. These results reveal a significant empirical superiority of the plug-in to state-of-the-art KDE- and kNN-based methods.**

I. Introduction

Consider the problem of estimating differential entropy under Gaussian convolutions that was recently introduced in [1]. Namely, let $X \sim P$ be an arbitrary 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 independently and identically distributed (i.i.d.) samples $X^n \triangleq (X_1, \ldots, X_n)$ from P and assuming σ is known[1], we aim to estimate $h(X+Z) = h(P * \mathcal{N}_\sigma)$, where $\mathcal{N}_\sigma$ is a centered isotropic Gaussian measure with parameter σ. To investigate the decision-theoretic fundamental limit, we consider the minimax absolute-error estimation risk

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

where $\mathcal{F}_d$ is a nonparametric class of distributions and $\hat{h}$ is the estimator. The sample complexity $n^\star(\eta, \sigma, \mathcal{F}_d)$ is the smallest number of samples for which estimation within an additive gap η is possible. This estimation setup was originally motivated by measuring the information flow in deep neural networks [2] for testing the Information Bottleneck compression conjecture of [3].

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 J. Weed was supported in part by the Josephine de Kármán fellowship.

[1]The extension to unknown σ is omitted for space reasons. Note that samples from P contain no information about σ. Hence for unknown σ, samples of both $X \sim P$ and Z would presumably be required. Under this alternative model, σ^2 can be estimated as the empirical variance of Z and then plugged into our estimator. It can be shown that this σ^2 estimate converges as $O\left((nd)^{-\frac{1}{2}}\right)$, which does not affect our estimator's overall convergence rate.

A. Contributions

The results herein establish the optimality of the plug-in estimator for the considered problem. Defining $\mathsf{T}_\sigma(P) \triangleq h(P * \mathcal{N}_\sigma)$ as the functional (of P) that we aim to estimate, the plug-in estimator is $\mathsf{T}_\sigma\big(\hat{P}_{X^n}\big) = h\big(\hat{P}_{X^n} * \mathcal{N}_\sigma\big)$, where $\hat{P}_{X^n} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure associated with the samples X^n and δ_{X_i} is the Dirac measure at X_i. Despite the suboptimality of plug-in techniques for vanilla discrete (Shannon) and differential entropy estimation (see [4] and [5], respectively), we show that $h\big(\hat{P}_{X^n} * \mathcal{N}_\sigma\big)$ attains the parametric estimation rate of $O_{\sigma,d}\left(\frac{1}{\sqrt{n}}\right)$ for the considered setup when P is a subgaussian measure. This establishes the plug-in estimator as minimax rate-optimal for differential entropy estimation under Gaussian convolutions.

The derivation of this optimal convergence rate first bounds the risk by a weighted total variation (TV) distance between the original measure $P * \mathcal{N}_\sigma$ and the empirical one $\hat{P}_{X^n} * \mathcal{N}_\sigma$. This bound is derived by linking the two measures via the maximal TV coupling, and reduces the analysis to controlling certain d-dimensional integrals. The subgaussianity of P is used to bound the integrals by a $\frac{c^d}{\sqrt{n}}$ term, with all constants explicitly characterized. It is then shown that the exponential dependence d is unavoidable. Specifically, we prove that any good estimator of $h(P * \mathcal{N}_\sigma)$, within an additive gap η, has a sample complexity $n^\star(\eta, \sigma, \mathcal{F}_d) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma)$ is positive and monotonically decreasing in σ. The proof relates the estimation of $h(P * \mathcal{N}_\sigma)$ to estimating the discrete entropy of a distribution supported on a capacity-achieving codebook for an additive white Gaussian noise (AWGN) channel.

B. Related Past Works and Comparison

General-purpose differential entropy estimators are applicable in the considered setup by accessing the noisy samples of $X+Z$. There are two prevailing approaches for estimating the nonsmooth differential entropy functional: the first relies on kernel density estimators (KDEs) [6], and the other uses k nearest neighbor (kNN) techniques (see [7] for a comprehensive survey). Many performance analyses of such estimators

restrict attention to smooth nonparametric density classes and assume these densities are bounded away from zero. Since the density associated with $P*\mathcal{N}_\sigma$ violates the boundedness from below assumption, any such result does not apply in our setup.

Two recent works weakened/dropped the boundedness from below assumption, providing general-purpose estimators whose risk bounds are valid in our setup. The first is [5], which proposed a KDE-based differential entropy estimator that also combines best polynomial approximation techniques. Assuming subgaussian densities with unbounded support, Theorem 2 of [5] bounded the estimation risk by[2] $O\left(n^{-\frac{s}{s+d}}\right)$, where s is a Lipschitz smoothness parameter assumed to satisfy $0 < s \leq 2$. While the result is applicable for our setup when P is compactly supported or subgaussian, its convergence rate quickly deteriorates with dimension d and is unable to exploit the smoothness of $P*\mathcal{N}_\sigma$ due to the $s \leq 2$ restriction.[3]

A second relevant work is [8], which studied a weighted-KL estimator (in the spirit of [9], [10]) for very smooth densities. Under certain assumptions on the densities' speed of decay to zero (which captures $P*\mathcal{N}_\sigma$ when, e.g., P is compactly supported) the proposed estimator was shown to attain $O\big(\frac{1}{\sqrt{n}}\big)$ risk. Despite the estimator's efficiency, empirically it is significantly outperformed by the plug-in estimator studied herein even in rather simple scenarios (see Section V). In fact, our simulations show that the vanilla (unweighted) kNN estimator of [11], which is also inferior to the plug-in, typically performs better than the weighted version from [8]. The poor empirical performance of the latter may originate from the dependence of the associated risk on d, which was overlooked in [8].

II. Preliminaries and Definitions

Logarithms are with respect to (w.r.t.) base e, $\|x\|$ is the Euclidean norm in $\mathbb{R}^d$, and I_d is the $d \times d$ identity matrix. We use $\mathbb{E}_P$ for an expectation w.r.t. a distribution P, omitting the subscript when P is clear. For a continuous $X \sim P$ with probability density function (PDF) p, we interchangeably use $h(X)$, $h(P)$ and $h(p)$ for its differential entropy. The n-fold product extension of P is denoted by $P^{\otimes n}$. The convolution of two distributions P and Q on $\mathbb{R}^d$ is $(P*Q)(\mathcal{A}) = \int\int \mathbb{1}_{\mathcal{A}}(x+y)\,\mathrm{d}P(x)\,\mathrm{d}Q(y)$, where $\mathbb{1}_{\mathcal{A}}$ is the indicator of the Borel set $\mathcal{A}$.

Let $\mathcal{F}_d$ be the set of distributions P with $\mathrm{supp}(P) \subseteq [-1,1]^d$.[4] We also consider the class of K-subgaussian distributions $\mathcal{F}_{d,K}^{(\mathsf{SG})}$ [12]. Namely, $P \in \mathcal{F}_{d,K}^{(\mathsf{SG})}$, for $K > 0$, if $X \sim P$ satisfies

$$\mathbb{E}_P\Big[\exp\big(\alpha^T(X-\mathbb{E}X)\big)\Big] \leq \exp\big(0.5K^2\|\alpha\|^2\big), \quad \forall \alpha \in \mathbb{R}^d. \tag{1}$$

In other words, every one-dimensional projection of X is subgaussian. Clearly, there exists a $K' > 0$ such that $\mathcal{F}_d \subseteq \mathcal{F}_{d,K'}^{(\mathsf{SG})}$. We therefore state our lower bound result (Theorem 1) for $\mathcal{F}_d$, while the upper bound (Theorem 2) is given for $\mathcal{F}_{d,K}^{(\mathsf{SG})}$.

[2]Multiplicative polylogarithmic factors are overlooked in this restatement

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

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

III. Exponential Sample Complexity

As claimed next, the sample complexity of any good estimator of $h(P*\mathcal{N}_\sigma)$ is exponential in d.

Theorem 1 (Exp. Sample Complexity): The following holds:

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

Part 1 of Theorem 1 is proven in Section VI-A. It relates the estimation of $h(P*\mathcal{N}_\sigma)$ to discrete entropy estimation of a distribution supported on a capacity-achieving codebook for a peak-constrained AWGN channel. Since the codebook size is exponential in d, discrete entropy estimation over the codebook within a small gap $\eta > 0$ is impossible with less than order of $\frac{2^{\gamma(\sigma)d}}{\eta d}$ samples [13]. The exponent $\gamma(\sigma)$ is monotonically decreasing in σ, implying that larger σ values are favorable for estimation. Part 2 of the theorem follows by similar arguments but for a d-dimensional AWGN channel with an input distributed on the vertices of the $[-1,1]^d$ hypercube; the proof is omitted (see [1, Section V-B2]).

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

IV. Optimality of Plug-in Estimator

We next establish the minimax-rate optimality of the plug-in estimator. Given a collection of samples $X^n \sim P^{\otimes n}$, the estimator is $h(\hat{P}_{X^n} * \mathcal{N}_\sigma)$, where $\hat{P}_{X^n} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

Theorem 2 (Plug-in Risk Bound): Fix $\sigma > 0, d \geq 1$. Then

$$\sup_{P \in \mathcal{F}_{d,\mu,K}^{(\mathsf{SG})}} \mathbb{E}_{P^{\otimes n}}\Big|h(P*\mathcal{N}_\sigma) - h(\hat{P}_{X^n}*\mathcal{N}_\sigma)\Big| \leq C_{\sigma,d,\mu,K} n^{-\frac{1}{2}}. \tag{2}$$

where $C_{\sigma,d,\mu,K} = O_{\sigma,\mu,K}(c^d)$ for a numerical constant c.

The proof of Theorem 2 is given in Section VI-B, where an explicit expression for $C_{\sigma,d,\mu,K}$ is stated in (12). The derivation exploits the maximal TV coupling to bound the right-hand side (RHS) of (2) by a weighted TV between $P*\mathcal{N}_\sigma$ and $\hat{P}_{X^n}*\mathcal{N}_\sigma$. Exploiting the Gaussian smoothing, we control this TV distance by a $c^d/\sqrt{n}$ term as desired.

Remark 2 (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 [14], which establishes this rate as a lower bound for the parametric estimation problem. Combined with Theorem 2, this establishes the plug-in estimator as minimax rate-optimal.

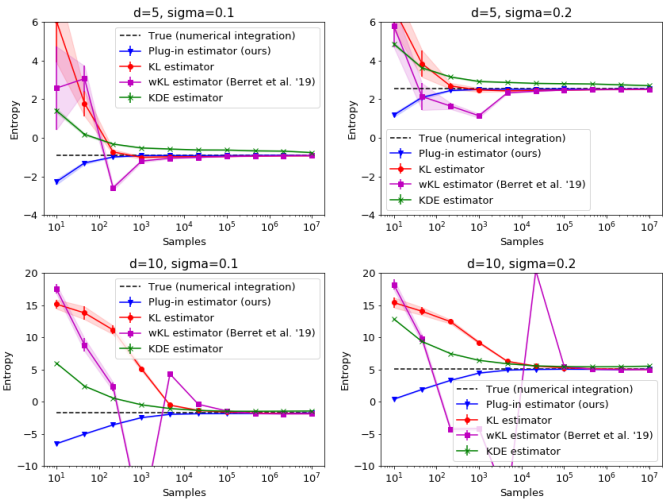


Fig. 1: Estimation results comparing the plug-in estimator to: (i) a KDE-based method [6]; (ii) the KL estimator [15]; and (iii) a weighted-KL estimator [8]. Here P is a truncated d-dimensional mixture of 2^d Gaussians and $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$. Error bars are one standard deviation over 20 random trials.

V. Experiments

We present empirical results illustrating the convergence of the plug-in estimator[5] compared to several competing methods: (i) the KDE-based estimator of [6]; (ii) and kNN Kozachenko-Leonenko (KL) estimator [15]; and (iii) the recently developed weighted-KL (wKL) estimator from [8].

***P* with Bounded Support:** Convergence rates in the bounded support regime are illustrated first. We set P as a mixture of Gaussians truncated to have support in $[-1,1]^d$. Before truncation, the mixture consists of 2^d Gaussian components with means at the 2^d corners of $[-1,1]^d$. Fig. 1 shows estimation results as a function of n, for $d = 5, 10$ and $\sigma = 0.1, 0.2$. The kernel width for the KDE estimate was chosen via cross-validation, varying with both d and n; the KL, wKL and plug-in estimators require no tuning parameters. We stress that the KDE estimate is highly unstable and, while not shown here, the estimated value is very sensitive to the chosen kernel width. The KDE, KL and wKL estimators converge slowly, at a rate that degrades with increased d, underperforming the plug-in estimator. Finally, we note that in accordance to the explicit risk bound from (12), the absolute error increases with larger d and smaller σ.

***P* with Unbounded Support:** In Fig. 2, we show the convergence rates in the unbounded support regime by considering the same setting with $d = 15$ but without truncating the 2^d-mode Gaussian mixture. The fast convergence of the plug-in estimator is preserved, outperforming the competing methods.

Reed-Muller Codes for AWGN Channels: We next consider data transmission over an AWGN channel using a binary

[5]Evaluating the plug-in estimator requires computing a d-dimensional integral, which has no closed form solution. Nonetheless, in [1] we propose an efficient Monte Carlo integration method to perform this computation. The method's accuracy is ensured via mean-squared error bounds [1, Theorem 5], and the computational complexity is shown to be on average $O(n \log n)$.

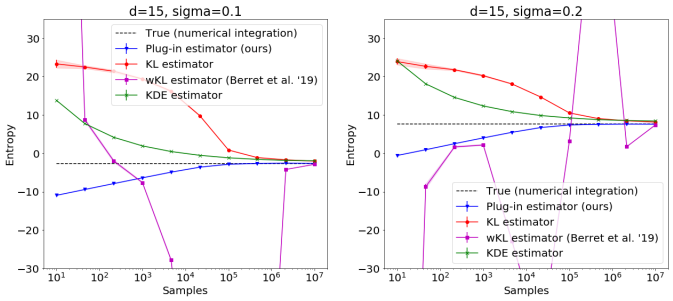
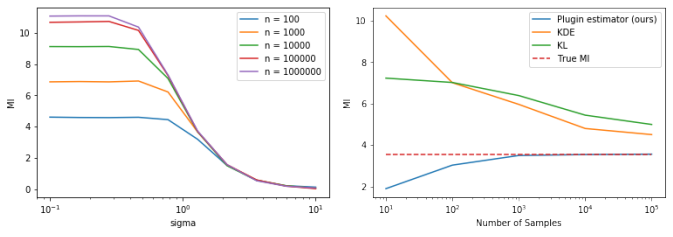


Fig. 2: Estimation results in the unbounded support regime, where P is a d-dimensional mixture of 2^d Gaussians. Error bars are one standard deviation over 20 random trials.

phase-shift keying (BPSK) modulation of a Reed-Muller code. A Reed-Muller code $\mathsf{RM}(r, m)$ of parameters $r, m \in \mathbb{N}$, where $0 \le r \le m$, encodes messages of length $k = \sum_{i=0}^{r} \binom{m}{i}$ into 2^m-lengthed binary codewords. Let $\mathcal{C}_{\mathsf{RM}(r,m)}$ be set of BPSK modulated sequences corresponding to $\mathsf{RM}(r, m)$ (with 0 and 1 mapped to -1 and 1, respectively). The number of bits reliably transmittable over the 2^m-dimensional AWGN channel with noise variance σ^2 is $I(X; X+Z) = h(X+Z) - \frac{d}{2}\log(2\pi e\sigma^2)$, where $X \sim \mathsf{Unif}(\mathcal{C}_{\mathsf{RM}(r,m)})$ and Z are independent. Despite $I(X; X+Z)$ being a well-behaved function of σ, an exact computation of this quantity is infeasible.

Our estimator readily estimates $I(X; X+Z)$ from samples of X. 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. 3 for various values of σ and n. Fig. 3(a) shows our estimate of $I(X; X+Z)$ for an $\mathsf{RM}(4,4)$ code as a function of σ, for different values of n. As expected, the plug-in estimator converges faster when σ is larger. Fig. 3(b) shows the estimated $I(X; X+Z)$ for $X \sim \mathsf{Unif}(\mathcal{C}_{\mathsf{RM}(5,5)})$ and $\sigma = 2$, with the KDE and KL estimates based on samples of $(X+Z)$ shown for comparison. Our method significantly outperforms the general-purpose estimators. The wKL estimator is omitted due to its instability in this high dimensional ($d = 32$) setting.

Remark 3: When $\mathrm{supp}(P)$ lies inside a ball of radius $\sqrt{d}$, the subgaussian constant K is proportional to d, and the bound from (2) scales as $\frac{d^d}{\sqrt{n}}$. This scenario corresponds to the popular setup of an AWGN channel with an input constraint.



(a) (b)

Fig. 3: Estimating $I(X; X+Z)$, where X comes from a BPSK modulated Reed-Muller and $Z \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$: (a) Estimated $I(X; X+Z)$ as a function of σ, for different n values, for the $\mathsf{RM}(4,4)$ code. (b) Plug-in, KDE and KL $I(X; X+Z)$ estimates for the $\mathsf{RM}(5,5)$ code and $\sigma = 2$ as a function of n.

VI. Proofs

A. Proof of Theorem 1

Let $Y = A + N$ be an AWGN channel with input peak constraint $A \in [-1,1]$ almost surely, and noise $N \sim \mathcal{N}(0,\sigma^2)$. The capacity $\mathsf{C}_{\mathsf{AWGN}}(\sigma) = \max_{A\sim P:\ \mathrm{supp}(P)\subseteq[-1,1]} I(A;Y)$ is positive for any $\sigma < \infty$. This positivity implies [16] that for any $\epsilon \in \big(0, \mathsf{C}_{\mathsf{AWGN}}(\sigma)\big)$ and large enough d, there exists a codebook $\mathcal{C}_d \subset [-1,1]^d$ of size $|\mathcal{C}_d| \doteq e^{d(\mathsf{C}_{\mathsf{AWGN}}(\sigma)-\epsilon)}$ and a decoder $\psi_d : \mathbb{R}^d \to [-1,1]^d$, such that

$$\mathbb{P}\Big(\psi_d(Y^d) = c \Big| A^d = c\Big) \geq 1 - e^{-\epsilon^2 d}, \quad \forall c \in \mathcal{C}_d, \tag{3}$$

where $A^d \triangleq (A_1, A_2, \ldots, A_d)$ and $Y^d \triangleq (Y_1, Y_2, \ldots, Y_d)$ are the channel input and output sequences, respectively.[6]

From (3) it follows that if $A^d \sim P$, for any P with $\mathrm{supp}(P) = \mathcal{C}_d$, and set $\hat{A}^d \triangleq \psi_d(Y^d)$, then

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

Invoking Fano's inequality, we further obtain

$$H\left(A^d \Big| \hat{A}^d\right) \leq H_b\left(e^{-\epsilon^2 d}\right) + e^{-\epsilon^2 d}\log|\mathcal{C}_d| \triangleq \delta^{(1)}_{\sigma,d}, \tag{4}$$

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

$$I\left(A^d; Y^d\right) \overset{(a)}{\geq} H\left(A^d\right) - H\left(A^d \Big| \hat{A}^d\right) \overset{(b)}{\geq} H\left(A^d\right) - \delta^{(1)}_{\sigma,d},$$

where (a) is since $H(A|B) \leq H\big(A\big|f(B)\big)$ for any random variables (A, B) and deterministic function f, while (b) uses (4).

Since we also have $I(A^d;Y^d) \leq H(A^d)$, it follows that

$$\Big|H(A^d) - I(A^d;Y^d)\Big| \leq \delta^{(1)}_{\sigma,d}, \tag{5}$$

which means that any good estimator of $H(A^d)$ over the class $\big\{P \,\big|\, \mathrm{supp}(P) = \mathcal{C}_d\big\} \subseteq \mathcal{F}_d)$ is also a good estimator of the mutual information. Using the well-known lower bound on discrete entropy estimation sample complexity (see, e.g., [17, Corollary 10]), we have that estimating $H(A^d)$ within a sufficiently small additive gap $\eta > 0$ requires at least $\Omega\left(\frac{|\mathcal{C}_d|}{\eta\log|\mathcal{C}_d|}\right) = \Omega\left(\frac{2^{\gamma(\sigma)d}}{\eta d}\right)$, where $\gamma(\sigma) \triangleq \mathsf{C}_{\mathsf{AWGN}}(\sigma) - \epsilon > 0$.

We relate the above back to the estimation of $h(X+Z)$ by noting that $I(A^d;Y^d) = h(A^d+N^d) - \frac{d}{2}\log_2(2\pi e\sigma^2)$. Letting $X \sim P$ and noting that $Z \overset{\mathcal{D}}{=} N^d$, where $\overset{\mathcal{D}}{=}$ denotes equality in distribution, we have $h(A^d+N^d) = h(X+Z)$. Assuming in contradiction that there exists an estimator of $h(X+Z)$ that uses $o\big(2^{\gamma(\sigma)d}/(\eta d)\big)$ samples and achieves an additive gap $\eta > 0$ over $\big\{P \,\big|\, \mathrm{supp}(P) = \mathcal{C}_d\big\}$, implies that $H(A^d)$ can be estimated from these samples within gap $\eta + \delta^{(1)}_{\sigma,d}$. This follows from (5) by taking the estimator of $h(X+Z)$ and subtracting the constant $\frac{d}{2}\log_2(2\pi e\sigma^2)$. We arrive at a contradiction.

[6] $a_k \doteq b_k$ denotes equality in the exponential scale: $\lim_{k\to\infty}\frac{1}{k}\log\frac{a_k}{b_k} = 0$.

B. Proof of Theorem 2

We start with two technical lemmata used for the proof.

Lemma 1: Let $U \sim P_U$ and $V \sim P_V$ be continuous random variables with densities p_U and p_V, respectively. If $\big|h(U)\big|, \big|h(V)\big| < \infty$, then

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

Proof: Recall the identity

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

Reversing the roles of U and V completes the proof. ∎

Lemma 2: Let $U \sim P_U$ and $V \sim P_V$ be continuous random variables with PDFs p_U and p_V, respectively. For any measurable function $g : \mathbb{R}^d \to \mathbb{R}$

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

Proof: We couple P_U and P_V via the TV maximal coupling[7]

$$\pi \triangleq (\mathrm{Id},\mathrm{Id})_\sharp(P_U \wedge P_V) + \frac{1}{\alpha}(P_U - P_V)_+ \otimes (P_U - P_V)_-, \tag{6}$$

where $(P_U - P_V)_+$ and $(P_U - P_V)_-$ are the positive and negative parts of the signed measure $(P_U - P_V)$; $(P_U \wedge P_V) \triangleq P_U - (P_U - P_V)_+$; $(\mathrm{Id},\mathrm{Id})_\sharp(P_U \wedge P_V)$ is the push-forward measure of $P_U \wedge P_V$ by the map $(\mathrm{Id},\mathrm{Id})$; $\otimes$ denotes a product measure; and $\alpha \triangleq \frac{1}{2}\int|p_U(x) - p_V(x)|dx$ satisfies $\int \mathrm{d}(P_U - P_V)_+ = \int \mathrm{d}(P_U - P_V)_- = \alpha$. Jensen's inequality implies $\big|\mathbb{E}g(U) - \mathbb{E}g(V)\big| \leq \mathbb{E}_\pi\big|g(U) - g(V)\big|$ and hence

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

∎

Fix any $P \in \mathcal{F}^{(\mathsf{SG})}_{d,K}$ and assume that $\mathbb{E}_P S = 0$. This assumption comes with no loss of generality since both the target functional $h(P*\mathcal{N}_\sigma)$ and the plug-in estimator are translation invariant. Note that $\big|h(P*\mathcal{N}_\sigma)\big|, \big|h(\hat{P}_{S^n}*\mathcal{N}_\sigma)\big| < \infty$. Combining Lemmas 1 and 2, we a.s. have

$$\begin{aligned}
&\big|h(P*\mathcal{N}_\sigma) - h(\hat{P}_{X^n}*\mathcal{N}_\sigma)\big| \\
&\qquad\leq \max\left\{\int\big|\log\tilde{r}_{X^n}(z)\big| \cdot \big|q(z) - r_{X^n}(z)\big|\,\mathrm{d}z,\right. \\
&\qquad\qquad\left.\int\big|\log\tilde{q}(z)\big| \cdot \big|q(z) - r_{X^n}(z)\big|\,\mathrm{d}z\right\},
\end{aligned} \tag{7}$$

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

[7] The maximal coupling attains maximal probability for the event $\{U = V\}$.

To control the above integrals we require one last lemma.

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

$$\mathbb{E}_{P^{\otimes n}}\big(\log \tilde{r}_{X^n}(z)\big)^2 \le \frac{1}{4\sigma^4}\mathbb{E}_P\|z-X\|^4 \tag{8a}$$

$$\big(\log \tilde{q}(z)\big)^2 \le \frac{1}{4\sigma^4}\mathbb{E}_P\|z-X\|^4\ . \tag{8b}$$

Proof: We prove (8a); the proof of (8b) is similar. The map $x \mapsto (\log x)^2$ is convex on $[0,1]$. For any fixed x^n, let $\hat{X} \sim \hat{P}_{x^n}$. Jensen's inequality gives

$$\big(\log \tilde{r}_{x^n}(z)\big)^2 = \left(\log \mathbb{E}_{\hat{P}_{x^n}} \exp\left(-\frac{\|z-\hat{X}\|^2}{2\sigma^2}\right)\right)^2 \le \mathbb{E}_{\hat{P}_{x^n}} \frac{\|z-\hat{X}\|^4}{4\sigma^4}.$$

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

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

■

Following (7), we bound $\mathbb{E}\int \big|\log \tilde{r}_{X^n}(z)\big|\big|p_U(z) – p_V(z)\big|\mathrm{d}z$. The bound for the other integral is identical and thus omitted.

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

$$\left(\mathbb{E}_{P^{\otimes n}} \int \big|\log \tilde{r}_{X^n}(z)\big|\big|q(z) - r_{X^n}(z)\big|\,\mathrm{d}z\right)^2 \tag{9}$$
$$\le \int \mathbb{E}_{P^{\otimes n}}\big(\log \tilde{r}_{X^n}(z)\big)^2 f_a(z)\mathrm{d}z \cdot \int \mathbb{E}_{P^{\otimes n}} \frac{\big(q(z)-r_{X^n}(z)\big)^2}{f_a(z)}\,\mathrm{d}z.$$

By virtue of Lemma 3, we bound the first integral as

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

where (a) follows from the triangle inequality, and (b) uses the K-subgaussianity of S [18, Lemma 5.5].

For the second integral, note that $r_{X^n}(z)$ is a sum of i.i.d. terms with expectation $q(z)$. This implies $\mathbb{E}_{P^{\otimes n}}\big(q(z) - r_{X^n}(z)\big)^2 \le \frac{c_1^2}{n}\mathbb{E}e^{-\frac{1}{\sigma^2}\|z-X\|^2}$ and further gives

$$\int \mathbb{E}_{P^{\otimes n}} \frac{\big(q(z)-r_{X^n}(z)\big)^2}{f_a(z)}\,\mathrm{d}z \le \frac{c_1}{n2^{d/2}}\mathbb{E}\frac{1}{f_a(X+Z/\sqrt{2})}, \tag{10}$$

where $Z \sim \mathcal{N}(0, \sigma^2\mathrm{I_d})$ and $X \sim P$ are independent.

Setting $c_2 \triangleq \left(\frac{\pi}{a}\right)^{\frac{d}{2}}$, we have $\big(f_a(z)\big)^{-1} = c_2 \exp\big(a\|z\|^2\big)$.

Since X is K-subgaussian and Z is σ-subgaussian, $X + Z/\sqrt{2}$ is $(K + \sigma/\sqrt{2})$-subgaussian. Following (10), for any $0 < a < \frac{1}{2(K+\sigma/\sqrt{2})^2}$, we have

$$\begin{aligned}
\frac{c_1}{n2^{d/2}}\mathbb{E}\frac{1}{f_a(X+Z/\sqrt{2})} &= \frac{c_1c_2}{n2^{d/2}}\mathbb{E}\exp\left(a\big\|X+Z/\sqrt{2}\big\|^2\right) \\
&\overset{(a)}{\le} \frac{c_1c_2}{n2^{d/2}}\exp\left(\big(K+\sigma/\sqrt{2}\big)^2 ad + \frac{(K+\sigma/\sqrt{2})^4a^2d}{1-2(K+\sigma/\sqrt{2})^2a}\right),
\end{aligned} \tag{11}$$

where (a) is by [12, Remark 2.3].

Setting $a = \frac{1}{4(K+\sigma/\sqrt{2})^2}$, we combine (7) and (9)-(11) to obtain the result (recalling that the second integral from (7) is bounded exactly as the first and using $\mathbb{E}[\max\{|X|,|Y|\}] \le \mathbb{E}|X| + \mathbb{E}|Y|$). For any $P \in \mathcal{F}_{d,K}^{(\mathsf{SG})}$ we have

$$\begin{aligned}
&\Big(\mathbb{E}_{P^{\otimes n}}\big|h(P*\mathcal{N}_\sigma) - h(\hat{P}_{X^n}*\mathcal{N}_\sigma)\big|\Big)^2 \\
&\qquad\le \frac{64\big(2d^2K^4 + d(d+2)(K+\sigma/\sqrt{2})^4\big)}{\sigma^4} \\
&\qquad\quad\times\left(\left(\frac{1}{\sqrt{2}} + \frac{K}{\sigma}\right)e^{\frac{3}{8}}\right)^d \frac{1}{n}.
\end{aligned} \tag{12}$$

ACKNOWLEDGEMENTS

We thank Yihong Wu for fruitful discussions on this topic.

REFERENCES

[1] Z. Goldfeld, K. Greenewald, and Y. Polyanskiy. Estimating differential entropy under Gaussian convolutions. *preprint arXiv:1810.11589*, 2018.

[2] Z. Goldfeld, E. van den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy. Estimating information flow in neural networks. *ArXiv preprint arXiv:1810.05728*, Nov. 2018.

[3] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *ArXiv preprint arXiv:1703.00810*, Mar. 2017.

[4] L. Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15:1191–1254, 2004.

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

[6] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and J. M. Robins. Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *NIPS*, pages 397–405, 2015.

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

[8] T. B. Berrett, R. J. Samworth, and M. Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *Annals of Statistics*, 47(1):288–318, 2019.

[9] K. Sricharan, D. Wei, and A. Hero. Ensemble estimators for multivariate entropy estimation. *IEEE Trans. Inf. Theory*, 59(7):4374–4388, 2013.

[10] K. Moon, K. Sricharan, K. Greenewald, and A. Hero. Ensemble estimation of information divergence. *Entropy*, 20(8), 2018.

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

[12] D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17, 2012.

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

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

[15] L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Prob. Pered. Inf.*, 23(2):9–16, 1987.

[16] R. Gallager. *Information theory and reliable communication*, volume 2. Springer, 1968.

[17] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. In *ECCC*, volume 17, page 9, Nov. 2010.

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