Supplement to Estimating Information Flow in Deep Neural Networks

Ziv Goldfeld [1 2] **Ewout van den Berg** [3 2] **Kristjan Greenewald** [3 2] **Igor Melnyk** [3 2] **Nam Nguyen** [3 2] **Brian Kingsbury** [3 2] **Yury Polyanskiy** [1 2]

NOTE: All references from the main text to this supplementary document can be replaced at publication time by references to a preprint on arXiv, per ICML guidelines.

7. Two-Neuron Leaky-ReLU Network Example

To expand upon Section **??**, we provide here a second example to illustrate the relation between clustering and compression of mutual information. In particular, this example also shows that as opposed to the claim from (Saxe et al., 2018), non-saturating nonlinearities can achieve compression. Consider the non-saturating Leaky-ReLU nonlinearity $R(x) \triangleq \max(x, x/10)$. Let $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_{1/4}$, with $\mathcal{X}_0 = \{1, 2, 3, 4\}$ and $\mathcal{X}_{1/4} = \{5, 6, 7, 8\}$, and labels 0 and $1/4$, respectively. We train the network via GD with learning rate 0.001 and mean squared loss. Initialization (shown in Fig. 9(a)) was chosen to best illustrate the connection between the Gaussians' motion and mutual information. The network converges to a solution where $w_1 < 0$ and b_1 is such that the elements in $\mathcal{X}_{1/4}$ cluster. The output of the first layer is then negated using $w_2 < 0$ and the bias ensures that the elements in $\mathcal{X}_0$ are clustered without spreading out the elements in $\mathcal{X}_{1/4}$. Figs. 9(b) show the Gaussian motion at the output of the first layer and the resulting clustering. For the second layer (Fig. 9(c)), the clustered bundle $\mathcal{X}_{1/4}$ is gradually raised by growing b_2, such that its elements successively split as they cross the origin; further tightening of the bundle is due to shrinking $|w_2|$. Fig. 9(d) shows the mutual information of the first (blue) and second (red) layers. The merging of the elements in $\mathcal{X}_{1/4}$ after their initial divergence is clearly reflected in the mutual information. Likewise, the spreading of the bundle, and successive splitting and coalescing of the elements in $\mathcal{X}_{1/4}$ are visible in the spikes in the red mutual information curve. The figure also shows how the bounds on $I\big(X;T(k)\big)$ precisely track its evolution.

[*]Equal contribution [1]Massachusetts Institute of Technology [2]MIT-IBM Watson AI Lab [3]IBM Research. Correspondence to: Ziv Goldfeld <zivg@mit.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

8. Experimental Details

8.1. SZT Model

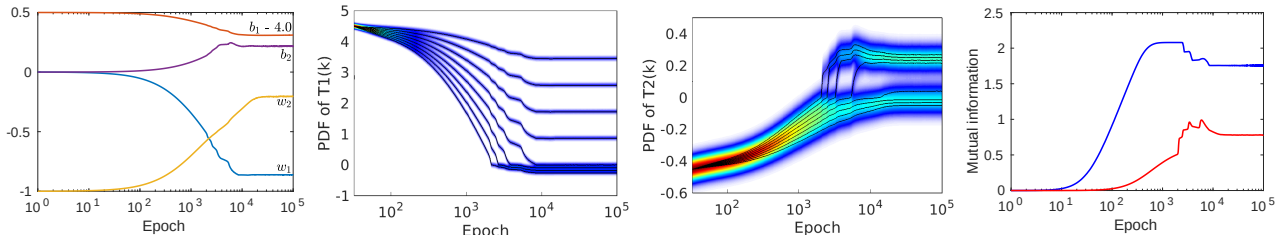
In this section we provide additional experimental details and results for the SZT model discussed in Section **??** of the main paper.

To regularize the network weights, we followed (Cisse et al., 2017) and adopted their approach for enforcing an orthonormality constraint. Specifically, we first update the weights $\{\mathrm{W}_\ell\}_{\ell\in[L]}$ using the standard gradient descent step, and then perform a secondary update to set

$$\mathrm{W}_\ell \leftarrow \mathrm{W}_\ell - \alpha\left(\mathrm{W}_\ell\mathrm{W}_\ell^T - \mathrm{I}_{d_\ell}\right)\mathrm{W}_\ell,$$

where the regularization parameter α controls the strength of the orthonormality constraint. The value of α was was selected from the set $\{1.0\times10^{-5}, 2.0\times10^{-5}, 3.0\times10^{-5}, 4.0\times10^{-5}, 5.0\times10^{-5}, 6.0\times10^{-5}, 7.0\times10^{-5}\}$ and the optimal value was found to be equal to 5.0×10^{-5} for both the tanh and ReLU.

In Fig. 10 we present additional experimental results that provide further insight into the clustering and compression phenomena for both tanh and ReLU nonlinearities. Fig. 10(a) shows what happens when the additive noise has a high variance. In this case, although saturation still occurs (see the histograms on top of Fig. 10(a)) and the Gaussians still cluster together (see the scatter plots on the right for the epoch 54 and epoch 8990), compression overall is very mild. The effect of increasing the noise parameter was explained in Section **??** of the main text (see, in particular, Fig. **??**(d) therein). Comparing Fig. 10(a) to Fig. **??**(a) of the main text, for which $\beta = 0.005$ was used and compression was observed, further highlights the effect of large β. Recall that smaller β values correspond to narrow Gaussians, while larger β values correspond to wider Gaussians. When β is small, even Gaussians that belong to the same cluster are distinguishable so long as they are not too close. When clusters tighten, the in-class movement brings these Gaussians closer together, effectively merging them, and causing a reduction



(a) (b) (c) (d)

Figure 9. Two-layer leaky ReLU network: (a) network parameters as a function of epoch, (b,c) the corresponding PDFs $p_{T_1(k)}$ and $p_{T_2(k)}$, and (d) the mutual information for both layers.

in mutual information (compression). One the other hand, for large β, the in-class movement is blurred at the outset (before clusters tighten). Thus, the only effect on mutual information is the separation between the clusters: as these blobs move away from each other, mutual information rises.

Based on the above observation, we can conclude that while the two notions of "clustering Gaussians" and "compression/decrease in mutual information" are strongly related in the low-beta regime, once the noise becomes large, these phenomena decouple, i.e., the network may cluster inputs and neurons may saturate, but this will not be reflected in a decrease of mutual information.

Finally, we present results for ReLU activation without weight normalization (Fig. 10(b)) and with orthonormal weight regularization (Fig. 10(c)). We see that both these networks exhibit almost no compression. For Fig. 10(c), the lack of compression is attributed to regularization of the weight matrices, as explained in Section **??** of the main text. For Fig. 10(b), the reduction in compression can be explained by the fact that although ReLU forces saturation of the neurons at the origin (which promotes clustering), since the positive axes remain unconstrained, the Gaussians can move off towards infinity without bound. This is visible from the histograms in the top row of Fig. 10(b), where, for example, in layer 5 the neurons can take arbitrarily large positive values (note that the bin corresponding to the value 5 accumulates all the values from 5 to infinity). Therefore, the clustering at the origin and the potential drop in mutual information is counterbalanced by the spread of Gaussians along the positive axes and the potential increase of mutual information it causes. Eventually, this leads to the approximately constant profile of the mutual information plot in Fig. 10(b).

The behavior of the weight-normalized ReLU in Fig. 10(c) is similar to Fig. 10(b), although now the growth of the network weights is bounded and the saturation around origin is reduced. For example, for layers 4 and 5 we can see an upward trend in the mutual information, which is then flattened at the end of training. This occurs since more Gaussians are moving away from the origin, although their motion remains bounded (see the histograms on the top and the scatter plots on the right), thus decreasing the clustering density, leading to the rise in the mutual information profile. Once the Gaussians are prevented from moving any further along the positive axes, a slight compression occurs and the mutual information flattens.

8.2. Spiral Model

In this section we present results for another synthetic example. We generated data in the form of spiral as in Fig. 11. The network architecture was similar to SZT model, except that the size of each layer was set to 3.

Fig. 12 shows MI estimates $I(X;T_\ell)$ computed using SP estimator and the discrete entropy estimates $H\big(\mathsf{Bin}(T_\ell)\big)$ for weight un-normalized Fig. 12 (a) and normalized models Fig. 12 (b) and using additive noise $\beta = 0.005$. Similar as in the main paper, the results in the figure illustrate a connection between clustering and compression.

Finally, in Fig. 13 we also show an estimate of $H\big(\mathsf{Bin}(T_\ell)\big)$ for the case of deterministic DNN trained on spiral data. For the particular choice of the bin size, the result of the estimated entropy reveal a certain level of clustering granularity.

8.3. MNIST CNN

In this section, we describe in detail the architecture of the MNIST CNN models used in Sections **??** and **??** in the main paper.

The MNIST CNNs were trained using PyTorch (Paszke et al., 2017) version `0.3.0.post4`. The CNNs use the following fairly standard architecture with two convolutional layers, two fully connected layers, and batch normalization.

1. 2-d convolutional layer with 1 input channel, 16 output channels, 5x5 kernels, and input padding of 2 pixels
2. Batch normalization
3. Tanh() activation function

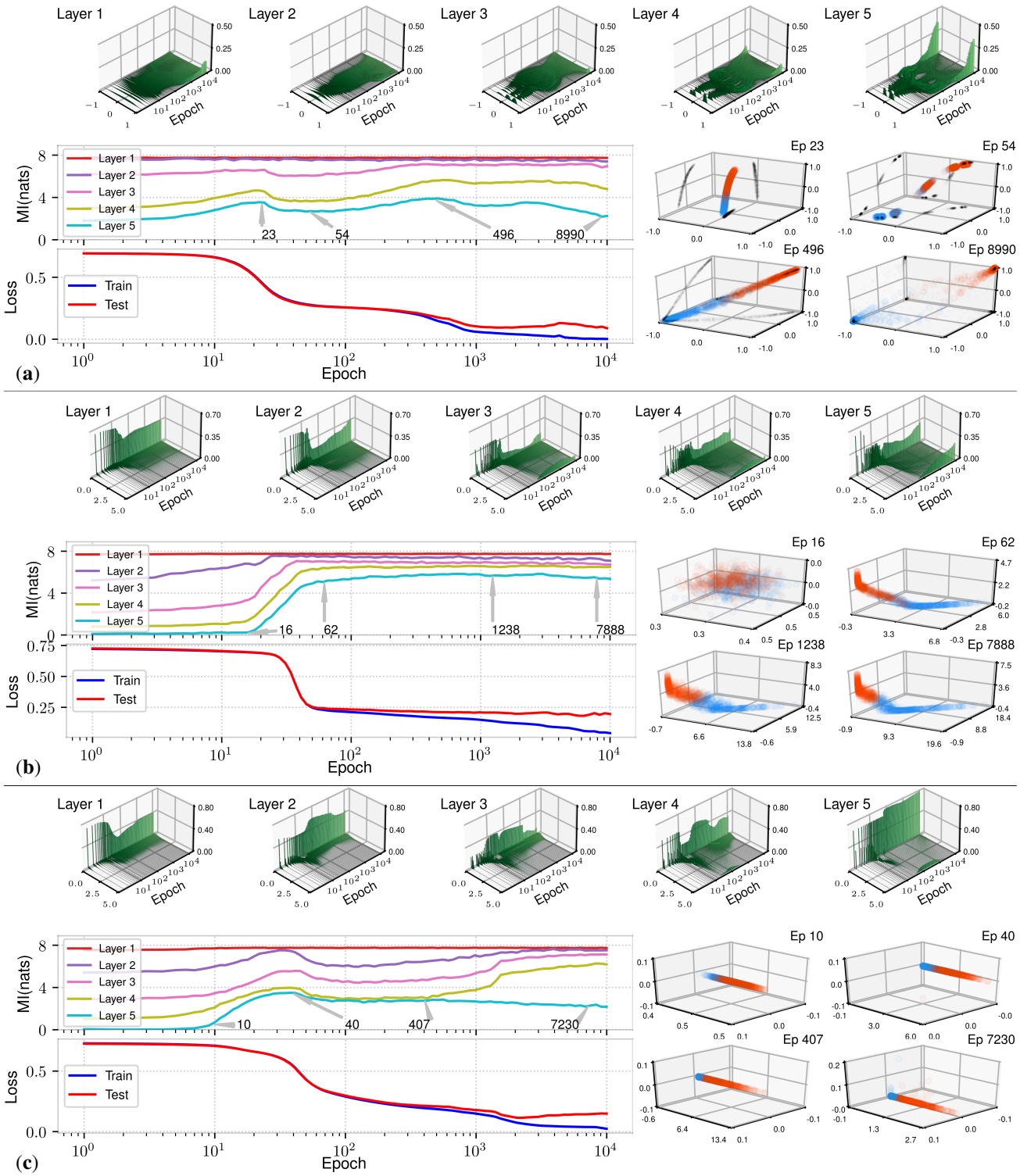


Figure 10. SZT model with (a) tanh nonlinearity and additive noise $\beta = 0.01$ without weight normalization, (b) ReLU nonlinearity and $\beta = 0.01$ without weight normalization, (c) ReLU nonlinearity and $\beta = 0.01$ with weight normalization. Test classification accuracy is 97%, 96%, and 97%, respectively.

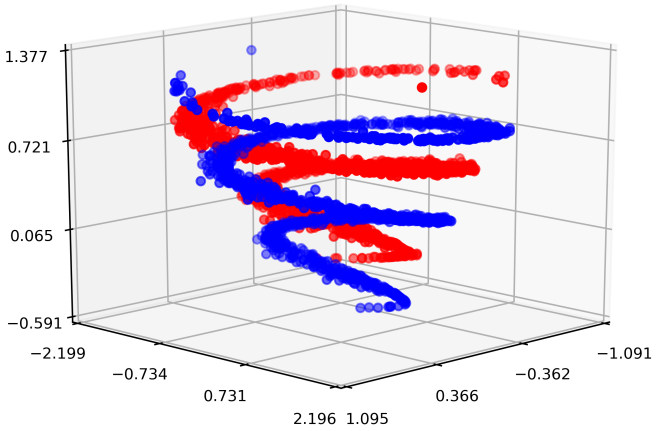


Figure 11. Generated spiral data for binary classification problem.

4. Zero-mean additive Gaussian noise with variance β^2 or dropout with a dropout probability of 0.2
5. 2x2 max-pooling
6. 2-d convolutional layer with 16 input channels, 32 output channels, 5x5 kernels, and input padding of 2 pixels
7. Batch normalization
8. Tanh() activation function
9. Zero-mean additive Gaussian noise with variance β^2 or dropout with a dropout probability of 0.2
10. 2x2 max-pooling
11. Fully connected layer with 1586 (32x7x7) inputs and 128 outputs
12. Batch normalization
13. Tanh() activation function
14. Zero-mean additive Gaussian noise with variance β^2 or dropout with a dropout probability of 0.2
15. Fully connected layer with 128 inputs and 10 outputs

All convolutional and fully connected layers have weights and biases, and the weights are initialized using the default initialization, which draws weights from $\mathsf{Unif}[-1/\sqrt{m}, 1/\sqrt{m}]$, with m the fan-in to a neuron in the layer. Training uses cross-entropy loss, and is performed using stochastic gradient descent with no momentum, 128 training epochs, and 32-sample minibatches. The initial learning rate is 5×10^{-3}, and it is reduced following a geometric schedule such that the learning rate in the final epoch is 5×10^{-4}. To improve the test set performance of our models, we applied data augmentation to the training set by translating, rotating, and shear-transforming each training example each time it was selected. Translations in the x- and y-directions were drawn uniformly from $\{-2, -1, 0, 1, 2\}$, rotations were drawn from $\mathsf{Unif}(-10°, 10°)$, and shear transforms were drawn from $\mathsf{Unif}(-10°, 10°)$.

To obtain more reliable performance results, we train eight different models and report the mean number of errors and standard deviation of the number of errors on the MNIST validation set. To ensure that the internal representations of different models are comparable, which is necessary for the use of the cosine similarity measure between internal representations, for each noise condition (deterministic, noisy with $\beta = 0.05$, noisy with $\beta = 0.1$, noisy with $\beta = 0.2$, noisy with $\beta = 0.5$, and dropout with $p = 0.2$), we use a common random seed (different for the eight replications, of course) so the models have the same initial weights and access the training data in the same order (use the same minibatches).

At test time, all models are fully deterministic: the additive noise blocks and dropout layers are replaced by identities. Thus, in the figures and text in the main paper, "Layer 1" is the output of step 5 (2x2 max-pooling), "Layer 2" is the output of step 10 (2x2 max-pooling), "Layer 3" is the output of step 13 (Tanh() activation function), and "Layer 4" is the output of step 15 (fully connected layer with 10 outputs).

9. Sample Propagation Estimator - Theoretic Guarantees

In this section we state performance guarantees for the SP estimator. We cite several foundational theorems from our work (Goldfeld et al., 2019), where this estimation problem is thoroughly studied. An anonymized copy of that paper is found at the end of the supplement and cited when needed. Proofs of all other results are relegated to Supplement 10.

9.1. Preliminary Definitions

Consider the estimation of the differential entropy $h(S + Z) = h(P * \varphi_\beta)$ based on n i.i.d. samples of $S \sim P$, where P is unknown and belongs to some nonparametric class, and φ_β (a PDF of an isotropic Gaussian with parameter β) is known. The minimax absolute-error risk over a given nonparametric class of distributions $\mathcal{F}$ is

$$\mathcal{R}^\star(n, \beta, \mathcal{F}) \triangleq \inf_{\hat{h}} \sup_{P \in \mathcal{F}} \mathbb{E}\left|h(P * \varphi_\beta) - \hat{h}(S^n, \beta)\right|, \quad (5)$$

where $\hat{h}$ is the estimator and $S^n \triangleq (S_i)_{i \in [n]}$ are the samples from P. In (5), by $P * \varphi_\beta$ we mean either: (i) $(P * \varphi_\beta)(x) = \int p(u)\varphi_\beta(x - u)du = (p * \varphi_\beta)(x)$, when P is continuous with density p; or (ii) $(P * \varphi_\beta)(x) =$

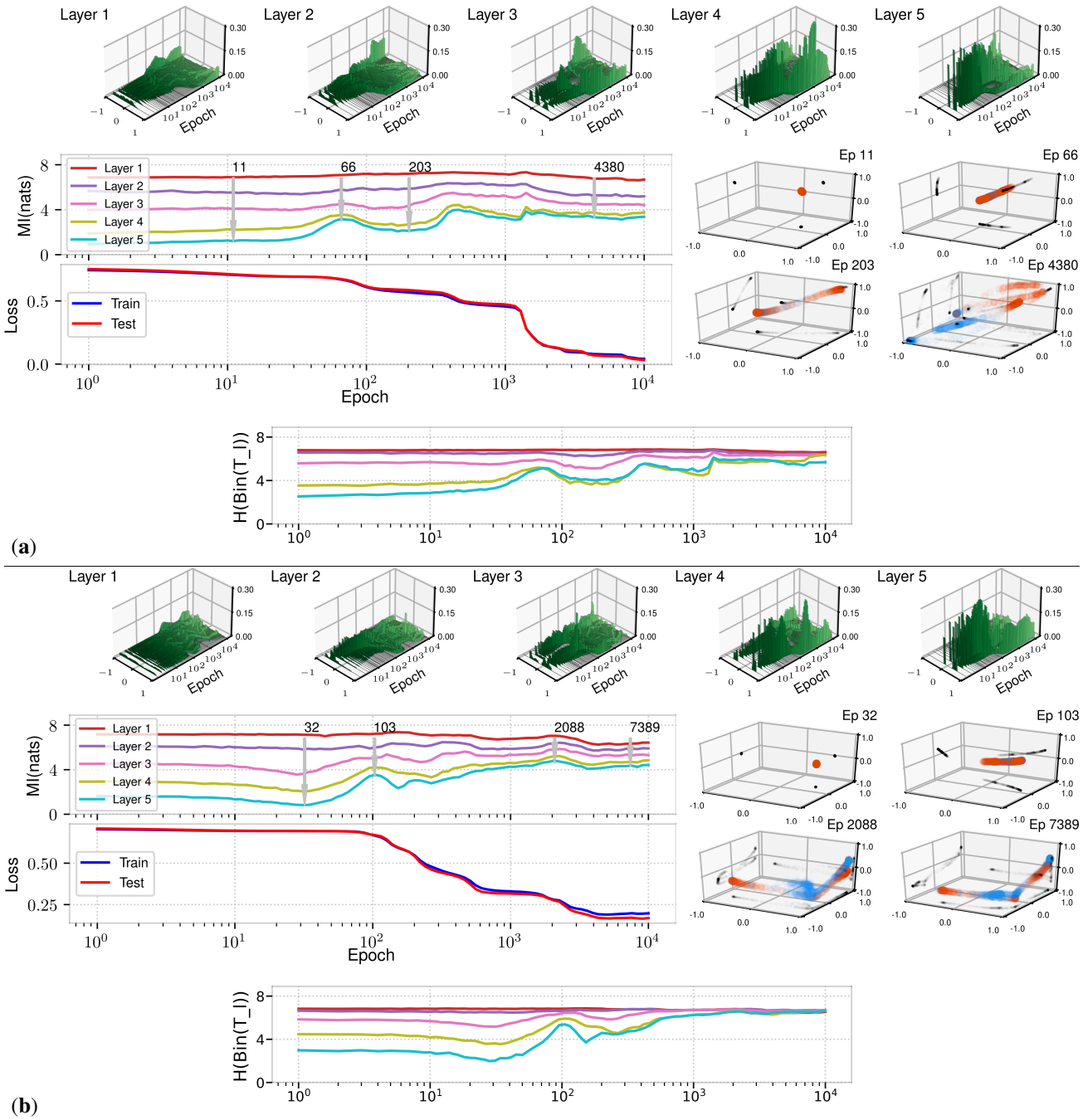


Figure 12. (a) Evolution of $I(X;T_\ell)$ and training/test losses across training epochs for Spiral dataset with $\beta = 0.005$ and tanh nonlinearities. The scatter plots on the right are the values of Layer 5 ($d_5 = 3$) at the arrow-marked epochs on the mutual information plot. The bottom plot shows the entropy estimate $H\big(\mathsf{Bin}(T_\ell)\big)$ across epochs for bin size $B = 10\beta$. (b) Same setup as in (a) but with a regularization that encourages orthonormal weight matrices.

$\sum_{u:\ p(u)>0} p(u)\varphi_\beta(x-u)$, if P is discrete with PMF p. This convolved distribution can be defined generally in a way that the two instances above as special cases using measure-theoretic concepts (see (Goldfeld et al., 2019)). Regardless of the nature of P, however, we stress that $P*\varphi_\beta$ is always a continuous distribution since it corresponds to the random variable $S+Z$, where Z is an isotropic Gaussian vector. The sample complexity $n^\star(\eta,\beta,\mathcal{F})$ is defined as the smallest number of samples n required to achieve a risk value less than or equal to a specified constant η in (5).

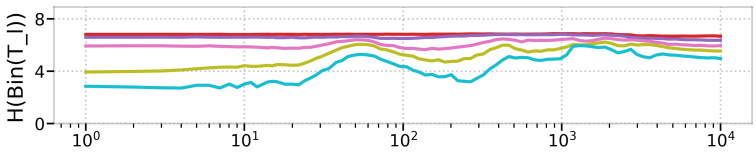


Figure 13. $H\big(\mathsf{Bin}(T_\ell)\big)$ estimate for deterministic net using spiral data. Bin size was set to $B = 0.001$.

Let $\mathcal{F}_d$ be the set of distributions P with $\mathrm{supp}(P) \subseteq [-1,1]^d$.[1] Furthermore, let $\mathcal{F}^{(\mathsf{SG})}_{d,\mu,K}$ be the class of K-subgaussian distributions, where we adopt the subgaussianity definition from (Hsu et al., 2012). Namely, $P \in \mathcal{F}^{(\mathsf{SG})}_{d,\mu,K}$, for $\mu \geq 0$ and $K > 0$, if $X \sim P$ satisfies $\|\mathbb{E}X\| \leq \mu$ and

$$\mathbb{E}\Big[\exp\big(\alpha^T(X - \mathbb{E}X)\big)\Big] \leq \exp\big(0.5K^2\|\alpha\|^2\big), \forall \alpha \in \mathbb{R}^d, \tag{6}$$

i.e., every one-dimensional projection of X is subgaussian. Clearly, there exists a $K' > 0$ such that $\mathcal{F}_d \subseteq \mathcal{F}^{(\mathsf{SG})}_{d,0,K'}$. We therefore state our lower bound results (Theorem 2) for $\mathcal{F}_d$, while the upper bound (Theorem 3) is given for $\mathcal{F}^{(\mathsf{SG})}_{d,\mu,K}$. The class $\mathcal{F}_d$ corresponds to hidden layers with bounded nonlinearities (such as tanh or sigmoid), while $\mathcal{F}^{(\mathsf{SG})}_{d,\mu,K}$ accounts for ReLU nonlinearities (when, for example, the input X is itself subgaussian).

9.2. Sample Complexity is Exponential in Dimension

We start with Theorem 1 from (Goldfeld et al., 2019), which states that the sample complexity of any good estimator of $h(P * \varphi_\beta)$ (to within an additive gap η) is exponential in d.

Theorem 2 (Theorem 1 from (Goldfeld et al., 2019))**.** *The following holds:*

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

2. *Fix $d \in \mathbb{N}$. There exist $\beta_0(d), \eta_0(d) > 0$, such that for all $\beta < \beta_0(d)$ and $\eta < \eta_0(d)$ we have sample complexity $n^\star(\eta, \beta, \mathcal{F}_d) \geq \Omega\left(\frac{2^d}{\eta d}\right)$.*

The exponent $\gamma(\beta)$ being monotonically decreasing in β suggests that larger values of β are favorable for estimation. Part 1 of the theorem states that an exponential sample complexity is inevitable when d is large. As a complementary result, the second part gives a sample complexity lower bound valid in any dimension for a small noise parameter. Nonetheless, the result accounts for orders of β considered in this work.

Remark 1 (Critical β Values)**.** *Theorem 2 is stated in asymptotic form for simplicity. We note that, for any d, the critical $\beta_0(d)$ value from the second part can be extracted by following the constants through the proof (which relies on Proposition 3 from (Wu & Yang, 2016)). These critical values are not unreasonably small. For example for $d = 1$, a careful analysis gives that Theorem 2 holds for all $\beta < 0.08$, which is satisfied by most of the experiments in this paper. This threshold on β changes very slowly when increasing d due to the rapid decay of the PDF of the normal distribution.*

9.3. Estimation Risk Bounds

We next focus on analyzing the performance of the SP mutual information estimator. We start by citing Theorem 2 of (Goldfeld et al., 2019), where the risk of the entropy estimation problem is bounded. Recall that the estimator of $h(P * \varphi_\beta)$ is $h(\hat{P}_{S^n} * \varphi_\beta)$, where $S^n = (S_i)_{i=1}^n$ is an i.i.d. sample set from P and $\hat{P}_{S^n}$ is their empirical distribution. The following theorem shows that the expected absolute error of this estimator decays at a rate of estimation $O\left(\frac{c^d}{\sqrt{n}}\right)$, for a numerical constant c and all dimensions d. A better rate of convergence with n cannot be attained due to the parametric estimation lower bound (see, e.g., Proposition 1 of (Chen, 1997)). The exponential dependence in d is also necessary as established by Theorem 2.

Theorem 3 (Theorem 2 from (Goldfeld et al., 2019))**.** *Fix $\beta > 0, d \geq 1$. Then*

$$\begin{aligned}
&\sup_{P \in \mathcal{F}^{(\mathsf{SG})}_{d,\mu,K}} \mathbb{E}\left|h(P * \varphi_\beta) - h(\hat{P}_{S_n} * \varphi_\beta)\right| \\
&\leq \left(\frac{1}{\sqrt{2}} + \frac{K}{\beta}\right)^{\frac{d}{2}} \\
&\quad \times \left(\frac{8\big(2\mu^4 + 32d^2K^4 + d(d+2)(K + \beta/\sqrt{2})^4\big)}{\beta^4}\right)^{\frac{1}{2}} \\
&\quad \times \exp\left(\frac{3d}{16} + \frac{\mu^2}{4(K + \beta/\sqrt{2})^2}\right) \frac{1}{\sqrt{n}}.
\end{aligned} \tag{7}$$

Remark 2 (Improved Constant for Bounded Support)**.** *Theorem 3 also applies to the narrower nonparametric class $\mathcal{F}_d$*

[1] 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.

in place of $\mathcal{F}^{(\mathsf{SG})}_{d,\mu,K}$. By directly analyzing this bounded support scenario[2] ($P \in \mathcal{F}_d$) one may improve the constant factor in Theorem 3 to give a bound of $\max\{1, \beta^{-d}\}2^{d+2}\sqrt{\frac{d}{n}}$.

Remark 3 (Comparison to Generic Estimators)**.** *Note that one could always sample φ_β and add the obtained noise samples to S^n to obtain a sample set from $P * \varphi_\beta$. These samples can be used to get a proxy of $h(P * \varphi_\beta)$ via a kNN- or a KDE-based differential entropy estimator. However, $P * \varphi_\beta$ violates the boundedness away from zero assumption that most of the convergence rate results in the literature rely on (Levit, 1978; Hall, 1984; Joe, 1989; Hall & Morton, 1993; Tsybakov & Van der Meulen, 1996; Haje & Golubev, 2009; Sricharan et al., 2012; Singh & Póczos, 2016; Kandasamy et al., 2015). Two recent works that weakened/dropped the boundedness from below assumption, providing general-purpose estimators whose risk bounds are valid in our setup, are (Han et al., 2017) and (Berrett et al., 2019). However, the analysis of the KDE-based estimator proposed in (Han et al., 2017) holds only for Lipschitz smoothness parameters up to $s \leq 2$ and attains the slow rate (overlooking multiplicative polylogarithmic factors) of $O\left(n^{-\frac{s}{s+d}}\right)$. The second work (Berrett et al., 2019) studies a weighted-kNN estimator in the high smoothness regime and proved its asymptotic efficiency. However, no explicit risk bounds were derived in that work and empirically the estimator is significantly outperformed by $h(\hat{P}_{S^n} * \varphi_\beta)$ (see Section V of (Goldfeld et al., 2019)).*

We now show how the theoretical guarantee on the accuracy of the differential entropy estimator (Theorem 3) translates to mutual information estimation via the SP estimator from (**??**). To formulate the claim, recall that $T_\ell = S_\ell + Z_\ell$, where $S_\ell \sim P_{S_\ell} = P_{f_\ell(T_{\ell-1})}$ and $Z_\ell \sim \mathcal{N}(0, \beta^2 \mathrm{I}_{d_\ell})$ are independent. Thus,

$$h(T_\ell) = h(P_{S_\ell} * \varphi_\beta) \tag{8a}$$

$$h(T_\ell|X = x) = h(P_{S_\ell|X=x_i} * \varphi_\beta). \tag{8b}$$

Provided n i.i.d. samples $\mathcal{X} = \{X_i\}_{i\in[n]}$ from P_X, 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 $S^n_\ell \triangleq \{S_{\ell,1}, S_{\ell,2}, \ldots, S_{\ell,n}\}$, which is a set of n i.i.d. samples from P_{S_ℓ}.

2. **Conditional Sampling Given X:** To generate i.i.d. samples from $P_{S_\ell|X=x_i}$, for $i \in [n]$, we feed X_i into the DNN n times, collect outputs from $T_{\ell-1}$ corresponding to different noise realizations, and apply f_ℓ on each. Denote the obtained samples by $S^n_\ell(X_i)$.[3]

The knowledge of φ_β and the generated samples S^n_ℓ and $S^n_\ell(X_i)$ can be used to estimate the unconditional and the conditional entropies, from (8a) and (8b), respectively.

For notational simplicity, the layer index ℓ is dropped for the remainder of this subsection. With the above sampling procedure we construct an estimator $\hat{I}_{\mathsf{SP}}\big(X^n, \hat{h}\big)$ of $I(X;T)$ based on a given estimator $\hat{h}(A^n, \beta)$ of $h(P * \varphi_\beta)$ for $P \in \mathcal{F}_d$ that uses i.i.d. samples $A^n = (A_1, \ldots, A_n)$ from P and knowledge of φ_β. Assume that $\hat{h}$ attains

$$\sup_{P\in\mathcal{F}_d} \mathbb{E}\left|h(P * \varphi_\beta) - \hat{h}(A^n, \beta)\right| \leq \Delta_{\beta,d}(n). \tag{9}$$

An example of such an $\hat{h}$ is the estimator $h(\hat{P}_{S^n} * \varphi_\beta)$ from Theorem 3; the corresponding $\Delta_{\beta,d}(n)$ term is the RHS of (7). Our SP mutual information estimator is (see (**??**))

$$\hat{I}_{\mathsf{SP}}\left(X^n, \hat{h}, \beta\right) \triangleq \hat{h}(S^n, \beta) - \frac{1}{n}\sum_{i=1}^{n} \hat{h}\big(S^n(X_i), \beta\big). \tag{10}$$

The following theorem bounds the expected absolute error of $\hat{I}_{\mathsf{SP}}\left(X^n, \hat{h}, \beta\right)$. The proof is given in Supplement 10.1.

Theorem 4. *For the above described setup, we have*

$$\sup_{P_X} \mathbb{E}\left|I(X;T) - \hat{I}_{\mathsf{SP}}\left(X^n, \hat{h}, \beta\right)\right| \leq 2\Delta_{\beta,d}(n) + \frac{d\log\left(1+\frac{1}{\beta^2}\right)}{4\sqrt{n}}. \tag{11}$$

Theorem **??** of the main text is an immediate consequence of Theorems 3 and 4. Interestingly, the quantity $\frac{1}{\beta^2}$ is the signal-to-noise ratio (SNR) between S and Z. The larger β 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 $\Delta_{\beta,d}$). 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

[2] e.g., by employing Proposition 5 from (Polyanskiy & Wu, 2016) to control the entropy difference via a Wasserstein 1 distance and them using Theorem 6.15 from (Villani, 2006) to bound the latter by an expression that lands itself for an elementary analysis.

[3] 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 $h(T_1|X) = h(Z) = \frac{d}{2}\log(2\pi e\beta^2)$, we see that no estimation of the conditional entropy is needed. The mutual information estimator given in (10) is modified by replacing the subtracted term with $h(Z)$.

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 $\frac{1}{\sqrt{n}}$.

9.4. Sample Propagation Estimator Bias

The results of the previous subsection are of a minimax flavor. That is, they state worst-case convergence rates of $h(P * \varphi_\beta)$ estimation over a nonparametric class of distributions. In practice, the true distribution may not be one that attains these worst-case rates, and convergence may be faster. However, while variance of $h(\hat{P}_{S^n} * \varphi_\beta)$ can be empirically evaluated using bootstrapping, there is no empirical test for the bias. Specifically, even if multiple estimations of $h(P * \varphi_\beta)$ via $h(\hat{P}_{S^n} * \varphi_\beta)$ consistently produce similar values, this does not necessarily suggest that these values are close to the true $h(P * \varphi_\beta)$. To have a guideline to the least number of samples needed to avoid biased estimation, we present the following lower bound on the estimation bias.

Theorem 5. *Fix $\beta > 0$, $d \geq 1$, and let $\epsilon \in \left(1 - \left(1 - 2Q\left(\frac{1}{2\beta}\right)\right)^d, 1\right]$, where Q is the Q-function.*[4] *Set $k_\star \triangleq \left\lfloor \frac{1}{\beta Q^{-1}\left(\frac{1}{2}\left(1-(1-\epsilon)^{\frac{1}{d}}\right)\right)} \right\rfloor$, where Q^{-1} is the inverse of the Q-function. By the choice of ϵ, clearly $k_\star \geq 2$, and the bias of the SP estimator over the class $\mathcal{F}_d$ is bounded as*

$$\sup_{P \in \mathcal{F}_d} \left| h(P * \varphi_\beta) - \mathbb{E} h(\hat{P}_{S^n} * \varphi_\beta) \right| \geq \log\left(\frac{k_\star^{d(1-\epsilon)}}{n}\right) - H_b(\epsilon). \tag{12}$$

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

Theorem 5 is proved in Supplement 10.2. Since $H_b(\epsilon)$ shrinks with ϵ, for sufficiently small ϵ values the lower bound from (12) shows that the SP estimator will not have negligible bias unless $n > k_\star^{d(1-\epsilon)}$ is satisfied. The condition $\epsilon > 1 - \left(1 - 2Q\left(\frac{1}{2\beta}\right)\right)^d$ is non-restrictive in any relevant regime of β and d. For instance, for typical β values we work with - around 0.1 - this lower bound is at most 0.0057 for all dimensions up to at least $d = 10^4$. Setting, e.g., $\epsilon = 0.01$ (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, in order to have negligible bias the number of estimation samples n should be at least $2^{0.99d}$, for any conceivably relevant dimension d.

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

9.5. Computing the Sample Propagation Estimator

Evaluating the SP mutual information estimator 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 MCI (Robert, 2004). To simplify the presentation, we present the method for an arbitrary Gaussian mixture without referring to the notation of the estimation setup.

Let $g(t) \triangleq \frac{1}{n} \sum_{i \in [n]} \varphi_\beta(t - \mu_i)$ be a d-dimensional n-mode Gaussian mixture, with $\{\mu_i\}_{i \in [n]} \subset \mathbb{R}^d$ and φ_β as the PDF of $\mathcal{N}(0, \beta^2 \mathrm{I}_d)$. Let $C \sim \mathsf{Unif}\{\mu_i\}_{i \in [n]}$ be independent of $Z \sim \mathcal{N}(0, \beta^2 \mathrm{I}_d)$ and note that $V \triangleq C + Z \sim g$.

We use MCI (Robert, 2004) to compute $h(g)$. First note that

$$\begin{aligned} h(g) &= -\mathbb{E} \log g(V) \\ &= -\frac{1}{n} \sum_{i \in [n]} \mathbb{E}\Big[\log g(\mu_i + Z) \Big| C = \mu_i \Big] \\ &= -\frac{1}{n} \sum_{i \in [n]} \mathbb{E} \log g(\mu_i + Z), \end{aligned} \tag{13}$$

where the last step follows by 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_β. For each $i \in [n]$, we estimate the i-th summand on the RHS of (13) by

$$\hat{L}_{\mathsf{MC}}^{(i)} \triangleq \frac{1}{n_{\mathsf{MC}}} \sum_{j \in [n_{\mathsf{MC}}]} \log g\left(\mu_i + Z_j^{(i)}\right), \tag{14a}$$

which produces

$$\hat{h}_{\mathsf{MC}} \triangleq \frac{1}{n} \sum_{i \in [n]} \hat{L}_{\mathsf{MC}}^{(i)} \tag{14b}$$

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

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \triangleq \mathbb{E}\left[\left(\hat{h}_{\mathsf{MC}} - h(g) \right)^2 \right]. \tag{15}$$

We have the following bounds on the MSE for tanh and ReLU networks.

Theorem 6 (MSE Bounds for MC Estimator)**.** *The following holds:*

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

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \leq \frac{2d(2 + \beta^2)}{\beta^2} \frac{1}{n \cdot n_{\mathsf{MC}}}. \tag{16}$$

2. *Assume $M_C \triangleq \mathbb{E}\|C\|_2^2 < \infty$ (e.g., ReLU network with bounded second moments), then*

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) \leq \frac{9d\beta^2 + 8(2+\beta\sqrt{d})M_C + 3(11\beta\sqrt{d}+1)\sqrt{M_C}}{\beta^2} \times \frac{1}{n \cdot n_{\mathsf{MC}}}. \quad (17)$$

The proof of Theorem 6 is found in Supplement 10.3. The MSE bounds scale only linearly with the dimension d, making β^2 in the denominator often the dominating factor experimentally.

10. Proofs

10.1. Proof of Theorem 4

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

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

Applying the triangle inequality to (10) we obtain

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

By assumption (9) and because $\mathrm{supp}(P_S) \subseteq [-1,1]^d$, we have

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

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

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{h}(S^n(X_i),\beta) - g(X_i)\right| \middle| X^n = x^n\right] \\
&\qquad\overset{(a)}{=} \mathbb{E}\left|\hat{h}\big(S^n(x_i),\beta\big) - h(P_{S|X=x_i} * \varphi_\beta)\right| \\
&\qquad\leq \Delta_{\beta,d}(n), \quad (21)
\end{aligned}$$

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

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

Turning to term (III), observe that $\left\{g(X_i)\right\}_{i\in[n]}$ are i.i.d random variables. Hence

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

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

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

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

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

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

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 $T \triangleq \big(T(k)\big)_{k\in[d]}$, we can obtain an upper bound as

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

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

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

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

$$h(p_{T|X=x}) \leq \frac{d}{2}\log\left(2\pi e(1+\beta^2)\right). \tag{28}$$

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

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

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

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

Taking the supremum over P_X concludes the proof.

10.2. Proof of Theorem 5

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

$$\mathbb{E}h(\hat{P}_{S^n} * \varphi_\beta) \leq h(P * \varphi_\beta). \tag{31}$$

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

Lemma 1. *The following equality holds:*

$$h(P*\varphi_\beta) - \mathbb{E}h(\hat{P}_{S^n}*\varphi_\beta) = I(S^n;Y). \tag{32}$$

Proof. 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 $T = S + Z \sim P * \varphi_\beta$ and first note that

$$F_Y(y) = \mathbb{P}\big(S_W+Z \leq y\big) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(S_i+Z\leq y) = F_T(y). \tag{33}$$

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

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

$$F_{Y|S^n}(y|s^n) = \mathbb{P}\big(S_W+Z\leq y\big|S^n=s^n\big) = \frac{1}{n}\mathbb{P}\big(s_i+Z\leq y\big), \tag{34}$$

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

□

Using the lemma, we have

$$\left|\sup_{P\in\mathcal{F}_d}\mathbb{E}h(P*\varphi_\beta) - h(\hat{P}_{S^n}*\varphi_\beta)\right| = \sup_{P\in\mathcal{F}_d} I(S^n;Y), \tag{35}$$

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

$$\begin{aligned} I(S^n;Y) &= I(S^n;Y,S_W) - I(S^n;S_W|Y) \\ &\overset{(a)}{\geq} I(S^n;S_W) - H(S_W|Y) \\ &= H(S_W) - H(S_W|S^n) - H(S_W|Y), \end{aligned} \tag{36}$$

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 (36) we use Fano's inequality to obtain

$$\begin{aligned} H(S_W|Y) &\leq H\big(S_W\big|\psi_{\mathcal{C}}(Y)\big) \\ &\leq H_b\big(\mathsf{P_e}(\mathcal{C})\big) + \mathsf{P_e}(\mathcal{C})\cdot\log|\mathcal{C}|, \end{aligned} \tag{37}$$

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 $\left\{y \in \mathbb{R}^d \middle| \psi_{\mathcal{C}}(y) = c_i\right\}$ that $\psi_{\mathcal{C}}$ maps to c_i. Note that $\mathcal{D}_i = \mathsf{C}_i$ for all $i \in [k^d]$ for which C_i doesn't intersect with the boundary of $[-1,1]^d$. 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:

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

$$\overset{(b)}{=} 1 - \left(1 - 2Q\left(\frac{1}{k\beta}\right)\right)^d, \tag{38}$$

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 (38) as given in the statement of the theorem gives the desired bound $\mathsf{P_e}(\mathcal{C}) \leq \epsilon$. Collecting the pieces and inserting back to (36), we obtain

$$I(S^n;Y) \geq \log\left(\frac{k_\star^{d(1-\epsilon)}}{n}\right) - H_b(\epsilon). \tag{39}$$

Together with (35) this concludes the proof.

10.3. Proof of Theorem 6

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 $p_Z = \varphi_\beta$ and $p_V = g$ are the PDFs of Z and V, respectively.

First observe that the estimator is unbiased:

$$\mathbb{E}\hat{h}_{\mathsf{MC}} = -\frac{1}{n \cdot n_{\mathsf{MC}}} \sum_{i=1}^{n} \sum_{j=1}^{n_{\mathsf{MC}}} \mathbb{E} \log g\left(\mu_i + Z_j^{(i)}\right) = h(g). \tag{40}$$

Therefore, the MSE expands as

$$\mathsf{MSE}\left(\hat{h}_{\mathsf{MC}}\right) = \frac{1}{n^2 \cdot n_{\mathsf{MC}}} \sum_{i=1}^{n} \mathsf{var}\Big(\log g(\mu_i + Z)\Big). \tag{41}$$

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

$$\mathsf{var}\Big(\log g(\mu_i + Z)\Big) \leq \beta^2 \mathbb{E}\Big[\big\|\nabla \log g(\mu_i + Z)\big\|_2^2\Big]. \tag{42}$$

We proceed with separate derivations of (16) and (17).

10.3.1. MSE Bound for Bounded Support

Since $\|C\|_2 \leq \sqrt{d}$ almost surely, Proposition 3 from (Polyanskiy & Wu, 2016) implies

$$\big\|\nabla \log g(v)\big\|_2 \leq \frac{\|v\|_2 + \sqrt{d}}{\beta^2}. \tag{43}$$

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

$$\mathsf{var}\Big(\log g(\mu_i + Z)\Big) \leq \frac{2d(4+\beta^2)}{\beta^2}, \tag{44}$$

for each $i \in [n]$. Together with (41), this concludes the proof of (16).

10.3.2. MSE Bound for Bounded Second Moment

To prove (17), we use Proposition 2 from (Polyanskiy & Wu, 2016) to obtain

$$\big\|\nabla \log g(v)\big\|_2 \leq \frac{1}{\beta^2}\big(3\|v\|_2 + 4\mathbb{E}\|C\|_2\big). \tag{45}$$

Via the Poincaré inequality from (42), the variance is bounded as

$$\begin{aligned}
&\mathsf{var}\Big(\log g(\mu_i + Z)\Big) \\
&\leq \frac{1}{\beta^2}\mathbb{E}\Big[(3\|\mu_i + Z\|_2 + 4\mathbb{E}\|C\|)^2\Big] \\
&\leq \frac{1}{\beta^2}\Big(9d\beta^2 + 16M_C + 24\beta\sqrt{dM_C} \\
&\qquad + 3\|\mu_i\|_2\left(3 + 9\beta\sqrt{d} + 8\beta\sqrt{dM_C}\right)\Big),
\end{aligned} \tag{46}$$

where the last step uses Hölder's inequality (namely, $\mathbb{E}\|C\|_2 \leq \sqrt{\mathbb{E}\|C\|_2^2}$). The proof of (17) is concluded by plugging (46) into the MSE expression from (41) and noting that $\frac{1}{n}\sum_{i=1}^n \|\mu_i\|_2 \leq \sqrt{M_C}$.

References

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

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

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Goldfeld, Z., Greenewald, K., Weed, J., and Polyanskiy, Y. Optimality of the plug-in estimator for differential entropy estimation under Gaussian convolutions. Paris, France, July 2019.

Haje, H. F. E. and Golubev, Y. On entropy estimation by m-spacing method. *Journal of Mathematical Sciences*, 163(3):290–309, Dec. 2009.

Hall, P. Limit theorems for sums of general functions of m-spacings. *Mathematical Proceedings of the Cambridge Philosophical Society*, 96(3):517–532, Nov. 1984.

Hall, P. and Morton, S. C. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45(1): 69–88, Mar. 1993.

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

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

Joe, H. Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 41(4):683–697, Dec. 1989.

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

Levit, B. Y. Asymptotically efficient estimation of nonlinear functionals. *Problemy Peredachi Informatsii*, 14(3):65–72, 1978.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

Polyanskiy, Y. and Wu, Y. Wasserstein continuity of entropy and outer bounds for interference channels. *IEEE Transactions on Information Theory*, 62(7):3992–4002, Jul. 2016.

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

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

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

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

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

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2006.

Wu, Y. and Yang, P. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, June 2016.