Formal Limitations on the Measurement of Mutual Information

David McAllester
Toyota Technological Institute at Chicago

Karl Stratos
Rutgers University

Abstract

Measuring mutual information from finite data is difficult. Recent work has considered variational methods maximizing a lower bound. In this paper, we prove that serious statistical limitations are inherent to any method of measuring mutual information. More specifically, we show that any distribution-free high-confidence lower bound on mutual information estimated from N samples cannot be larger than $O(\ln N)$.

1 INTRODUCTION

Mutual information has important applications to unsupervised learning. It underlies classical representation learning methods such as Brown clustering (Brown *et al.*, 1992), INFOMAX (Bell and Sejnowski, 1995), and the information bottleneck method (Tishby *et al.*, 1999). Maximizing mutual information is also central to recent works on unsupervised representation learning with neural networks (McAllester, 2018; Belghazi *et al.*, 2018; Oord *et al.*, 2018; Stratos, 2019; Hjelm *et al.*, 2019).

Unfortunately, measuring mutual information from finite data is a notoriously difficult estimation problem. This difficulty has motivated researchers to consider more computationally amenable measurement methods that maximize a parameterized lower bound on mutual information. The hope is that the optimized value of the bound is close to the true value of mutual information and can serve as a good enough approximation. For instance, this is the approach advocated by Mutual Information Neural Estimator (MINE) (Belghazi *et al.*, 2018) and contrastive predictive coding (CPC) (Oord *et al.*, 2018).

Here we prove that serious statistical limitations are inherent to all methods of measuring mutual information. More specifically, we show that any distribution-free high-confidence lower bound on mutual information cannot be larger than $O(\ln N)$ where N is the number of samples. Thus a meaningful high-confidence lower bound is infeasible when underlying mutual information is large (e.g., hundreds of bits).

Our result corrects, generalizes, and unifies past work on measuring mutual information. Unlike previous intractability results that are tied to specific estimators (Gao *et al.*, 2015; Oord *et al.*, 2018), our result is universal to all estimators. Hence we show that without making strong assumptions on the population distribution, such as the small support assumption in Valiant and Valiant (2011) and minimax bounds (Jiao *et al.*, 2015), it is generally impossible to guarantee an accurate estimate of mutual information. Our results contradict a theorem in Belghazi *et al.* (2018) claiming a polynomial sample size convergence rate for a mutual information estimator and we point out an error in their proof.

While it is infeasible to give meaningful high-confidence lower bound guarantees for large mutual information, estimators lacking formal guarantees might still be useful in practice. To this end, we propose expressing mutual information as a difference of entropies and estimating the entropy terms by cross-entropy upper bounds. This difference-of-entropies (DoE) estimator gives neither an upper bound nor a lower bound guarantee on mutual information. Nevertheless, we give theoretical and empirical evidence that DoE can meaningfully estimate large mutual information from feasible samples.

1.1 Overview

We state our main result below. We write (X, Y) to denote a pair of random variables with ranges $(\mathcal{X}, \mathcal{Y})$. Unless otherwise specified, they can be discrete or continuous. For simplicity, all distributions are assumed to have full support.

Theorem 1.1. *Let B be any mapping from N samples of (X, Y) to a real number with the following property: for any distribution p_{XY} over (X, Y), given N iid samples $(x_1, y_1) \ldots (x_N, y_N) \sim p_{XY}$, with probability at least 0.99*

$$I(X, Y; p_{XY}) \geq B((x_1, y_1) \ldots (x_N, y_N))$$

where $I(X, Y; p_{XY})$ is the mutual information between (X, Y) under p_{XY}. Now pick any distribution q_{XY}

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

over (X, Y). *Then given* $N \geq 50$ *iid samples* $(x'_1, y'_1) \dots (x'_N, y'_N) \sim q_{XY}$*, with probability at least 0.96*

$$B((x'_1, y'_1) \dots (x'_N, y'_N)) \leq 2 \ln N + 5$$

In the rest of the paper, we build toward this result by proving statistical limitations on measuring lower bounds on Kullback-Leibler (KL) divergence and entropy. We derive several intermediate results, specifically:

- We first consider the Donsker-Varadhan lower bound on KL divergence which underlies the approach in MINE. We give an intuitive explanation on why estimating the bound from samples is problematic. We show that the polynomial sample complexity result given by Belghazi *et al.* (2018) is incorrect.
- We formally prove that any lower bound on KL divergence cannot be larger than $O(\ln N)$. This subsumes the Donsker-Varadhan bound as a special case.
- We formally prove that any lower bound on entropy cannot be larger than $O(\ln N)$. Theorem 1.1 is a special case of this result.
- We motivate measuring mutual information as a difference of entropies, each of which measured by minimizing cross entropy. We empirically show that it outperforms existing variational lower bounds in synthetic experiments and produce realistic estimates of mutual information in real-world datasets.

2 ISSUES WITH THE DONSKER-VARADHAN LOWER BOUND

The Donsker-Varadhan (DV) lower bound on KL divergence is stated below. For completeness, a simple proof is given in the supplementary material.

Theorem 2.1 (Donsker and Varadhan, 1983)**.** *Let* p_X *and* q_X *be distributions over* X *with finite KL divergence. Then for all bounded functions* $f : \mathcal{X} \to \mathbb{R}$

$$D_{\mathrm{KL}}(p_X || q_X) \geq \mathrm{DV}_f(p_X || q_X) \tag{1}$$

where

$$\mathrm{DV}_f(p_X || q_X) := \mathop{\mathbb{E}}_{x \sim p_X}[f(x)] - \ln \mathop{\mathbb{E}}_{x \sim q_X}\left[e^{f(x)}\right] \tag{2}$$

Moreover, (1) *holds with equality for some* f *with range* $[0, F_{\max}]$.

The DV bound (2) computes the difference between the expected value of $f(x)$ under p_X and the log of the expected value of the exponential of $f(x)$ under q_X. It can be easily estimated by sampling: given $x_1 \dots x_N \sim p_X$ and $x'_1 \dots x'_N \sim q_X$, we can compute the empirical estimate

$$\widehat{\mathrm{DV}}_f^N(p_X || q_X) := \frac{1}{N}\sum_{i=1}^{N} f(x_i) - \ln\left(\frac{1}{N}\sum_{i=1}^{N} e^{f(x'_i)}\right) \tag{3}$$

An application of the DV bound is measuring KL divergence. Specifically, we estimate KL divergence by estimating the associated DV bound from samples:

$$\begin{aligned} D_{\mathrm{KL}}(p_X || q_X) &\geq \mathrm{DV}_f(p_X || q_X) \\ &\gtrsim \widehat{\mathrm{DV}}_f^N(p_X || q_X) \end{aligned}$$

The first inequality holds for all choices of f by Theorem 2.1. We require that the second inequality holds with high probability (with respect to random sampling). We now show that this approach to measuring KL divergence is problematic.

2.1 Statistical Limitations on Measuring the DV Bound

A crucial observation is that the second term in the DV bound (2) involves an expectation of the exponential

$$\mathop{\mathbb{E}}_{x \sim q_X}\left[e^{f(x)}\right]$$

This expression has the same form as the moment generating function used in analyzing large deviation probabilities. The utility of expectations of exponentials in large deviation theory is that such expressions can be dominated by extremely rare events (large deviations). The rare events dominating the expectation will never be observed by sampling from q_X.

To quantitatively analyze the risk of unseen outlier events, we will make use of the following simple lemma where we write $p_X(\Phi[x])$ for the probability over drawing x from p_X that the statement $\Phi[x]$ holds.

Lemma 2.2 (Outlier risk lemma)**.** *Given* $N \geq 2$ *samples from* p_X *and a property* $\Phi[x]$ *such that* $p_X(\Phi[x]) \leq 1/N$*, the probability that no sample* x *satisfies* $\Phi[x]$ *is at least 1/4.*

Proof. The probability that $\Phi[x]$ is unseen in the sample is at least $(1 - 1/N)^N$ which is at least $1/4$ for $N \geq 2$ and where $\lim_{N \to \infty}(1 - 1/N)^N = 1/e$. □

We can use the outlier risk lemma to perform a quantitative risk analysis of the DV bound. Assume without loss of generality that $[0, F_{\max}]$ is the range of f taken on $\mathcal{X}$. Let us consider the best case scenario in which the empirical estimate (3) is the largest. It is easy to see that the largest value is $F_{\max}$ and attained when

$$\begin{aligned} f(x_i) &= F_{\max} && \forall i = 1 \dots N \\ f(x'_i) &= 0 && \forall i = 1 \dots N \end{aligned}$$

where $x_1 \dots x_N$ are samples from p_X and $x'_1 \dots x'_N$ are samples from q_X. But by the outlier risk lemma, there is still at least a 1/4 probability that

$$\mathop{\mathbb{E}}_{x\sim q_X}\left[e^{f(x)}\right] \geq \frac{1}{N}e^{F_{\max}}$$

Since we require that (3) is a high-confidence lower bound on the DV bound, it must account for the unseen outlier risk.[1] In particular, we must have

$$\widehat{\mathrm{DV}}_f^N(p_X||q_X) \leq F_{\max} - \ln\frac{e^{F_{\max}}}{N} = \ln N$$

2.2 Discussion of MINE

Our investigation of measuring the DV bound from samples is motivated by Belghazi *et al.* (2018) who propose measuring and maximizing mutual information by formulating it as KL divergence and considering the DV lower bound. More specifically, they introduce a function $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ parameterized by a neural network and perform gradient descent to optimize

$$\sup_f \frac{1}{N}\sum_{i=1}^N f(x_i, y_i) - \ln\frac{1}{N}\sum_{i=1}^N e^{f(x'_i,y'_i)} \tag{4}$$

where $(x_1, y_1)\dots(x_N, y_N)$ are drawn from p_{XY} and $(x'_1, y'_1)\dots(x'_N, y'_N)$ are drawn from $p_X \times p_Y$. This is an empirical estimate of the DV bound on

$$D_{\mathrm{KL}}(p_{XY}||p_X \times p_Y) = I(X, Y; p_{XY}) \tag{5}$$

They claim that (4) yields a high-confidence accurate measurement of underlying mutual information with polynomial sample complexity under mild assumptions (Theorem 3 in their paper), which apparently contradicts our previous observation.

Upon inspection, we have found that their claim is wrong. In the appendix of the arXiv version v4, they make an incorrect application of the Hoeffding inequality in equation (46) from which they incorrectly derive equation (49). Hoeffding depends on the bounded range of the random variable: (49) is bounding the exponential of the variable. Thus their bound stated in Theorem 3 has an exponential dependence on the variable M.

3 STATISTICAL LIMITATIONS ON MEASURING LOWER BOUNDS ON KL DIVERGENCE

The analysis given in Section 2.1 is specific to the DV bound. One might hope that there exists a different lower bound on KL divergence (e.g., $D_{\mathrm{KL}}(p_X||q_X) = \sup_{f>0} \mathop{\mathbb{E}}_{x\sim p_X}[\ln f(x)] - \mathop{\mathbb{E}}_{x\sim q_X}[f(x)] + 1$ in Nguyen *et al.* (2010) which does not involve an expectation of the exponential) that overcomes the limitation.

We now strengthen the result by formally proving that no lower bound on KL divergence estimated from samples can be large. We consider a more challenging setting in which we have perfect knowledge of p_X (i.e., we can compute probabilities under the distribution) and only sample from q_X to estimate a lower bound on $D_{\mathrm{KL}}(p_X||q_X)$. Even in this setting, we have the following negative result.

Theorem 3.1. *Let B be any distribution-free high-confidence lower bound on $D_{\mathrm{KL}}(p_X||q_X)$ computed with complete knowledge of p_X but only samples from q_X.*

More specifically, let $B(p_X, S, \delta)$ be any real-valued function of a distribution p_X, a multiset S, and a confidence parameter δ such that, for any p_X, q_X and δ, with probability at least $1-\delta$ over a draw of S from q_X^N we have

$$D_{\mathrm{KL}}(p_X||q_X) \geq B(p_X, S, \delta)$$

For any such bound, and for $N \geq 2$, for any q_X, with probability at least $1-4\delta$ over the draw of S from q_X^N we have

$$B(p_X, S, \delta) \leq \ln N$$

Proof. Consider distributions p_X and q_X and $N \geq 2$. Define $\tilde{q}_X$ by

$$\tilde{q}_X(x) = \left(1 - \frac{1}{N}\right)q_X(x) + \frac{1}{N}p_X(x)$$

We now have $D_{\mathrm{KL}}(p_X||\tilde{q}_X) \leq \ln N$. We will prove that from samples $S \sim q_X^N$ we cannot reliably distinguish between q_X and $\tilde{q}_X$.

The premise is that B yields a high-confidence lower bound for any distribution. Thus we have

$$\Pr_{S\sim\tilde{q}_X^N}(\mathrm{Small}(S)) \geq 1-\delta \tag{6}$$

where $\mathrm{Small}(S)$ represent the event that $B(p_X, S, \delta) \leq \ln N$. The distribution $\tilde{q}_X$ equals the marginal on x of a distribution on pairs (s, x) where s is the value of Bernoulli variable with bias $1/N$ such that if $s = 1$ then x is drawn from p_X and otherwise x is drawn from q_X. By Lemma 2.2 the probability that all coins are zero is at least 1/4. Conditioned on all coins being zero the distributions $\tilde{q}_X^N$ and q_X^N are the same. Let $\mathrm{Pure}(S)$ represent the event that all coins

[1] We intentionally keep this argument informal to give intuition since the result on the DV bound will be subsumed by the general result in Theorem 3.1.

are 0. We now have

$$\begin{aligned}
&\Pr_{S\sim q_X^N}(\text{Small(S)}) \\
&= \Pr_{S\sim \tilde{q}_X^N}(\text{Small}(S)|\text{Pure}(S)) \\
&= \frac{\Pr_{S\sim \tilde{q}_X^N}(\text{Pure}(S)\wedge \text{Small(S)})}{\Pr_{S\sim \tilde{q}_X^N}(\text{Pure}(S))} \\
&\geq \frac{\Pr_{S\sim \tilde{q}_X^N}(\text{Pure}(S)) - \Pr_{S\sim \tilde{q}_X^N}(\neg\text{Small}(S))}{\Pr_{S\sim \tilde{q}_X^N}(\text{Pure}(S))} \\
&\geq \frac{\Pr_{S\sim \tilde{q}_X^N}(\text{Pure}(S)) - \delta}{\Pr_{S\sim \tilde{q}_X^N}(\text{Pure}(S))} \\
&= 1 - \frac{\delta}{\Pr_{S\sim \tilde{q}_X^N}(\text{Pure}(S))} \\
&\geq 1 - 4\delta
\end{aligned}$$

□

Note the importance of the fact that the lower bound B is distribution-free (the DV bound is distribution-free): this allows us to construct an adversarial distribution $\tilde{q}_X$ with small KL divergence and turn the premise on its head to upper bound B (6). It may be possible to construct distribution-specific lower bounds that do not suffer the same limitations. Theorem 3.1 proves that without making such additional assumptions, it is not possible to guarantee a large lower bound on KL divergence.

This result has immediate implications on measuring mutual information which is a special case of KL divergence (5). A direct application of Theorem 3.1 gives the following: in the setting in which we have perfect knowledge of p_{XY} but can only sample from p_X and p_Y (i.e., we cannot compute marginals) and measure a lower bound on $I(X,Y;p_{XY})$ from N samples, we cannot guarantee that the bound is larger than $\ln N$. However, this setting is arguably awkward. In the next section, we make a more natural argument by considering entropy.

4 STATISTICAL LIMITATIONS ON MEASURING LOWER BOUNDS ON ENTROPY

Recall that mutual information can be formulated as a difference of entropies

$$I(X,Y;p_{XY}) = H(X;p_X) - H(X|Y;p_{XY}) \quad (7)$$

where $H(X;p_X)$ is the entropy of X under p_X and $H(X|Y;p_{XY})$ is the conditional entropy of X given Y under p_{XY}. Entropy is nonnegative for discrete variables: in this case we have

$$I(X,Y;p_{XY}) \leq H(X;p_X)$$

It states that the mutual information between X and Y cannot be larger than information content of X alone. Thus a lower bound on mutual information implies a lower bound on entropy. We will show that any distribution-free high-confidence lower bound on entropy requires a sample size exponential in the size of the bound.

The above argument seems problematic for the case of continuous densities as differential entropy can be negative. However, for the continuous case we have

$$I(X,Y;p_{XY}) = \sup_{C,C'} I(C(X),C'(Y);p_{XY}) \quad (8)$$

where C and C' range over all maps from the underlying continuous space to discrete sets (all binnings of the continuous space). A proof is given in the supplementary material. Hence an $O(\ln N)$ upper bound on the measurement of mutual information for the discrete case applies to the continuous case as well. We assume the discrete case in this section without loss of generality.

We use the following definition.

Definition 4.1. *The type of a multiset S, denoted $\mathcal{T}(S)$, is a function on positive integers such that*

$$\mathcal{T}(S)(i) := \left|\left\{x \in S : \sum_{x'\in S:\ x'=x} 1 = i\right\}\right|$$

That is, $\mathcal{T}(S)(i)$ is the number of elements of S that occur i times in S.

The type $\mathcal{T}(S)$ contains all information relevant to estimating the actual probability of the items of a given count and of estimating the entropy of the underlying distribution. The problem of estimating distributions and entropies from sample types has been investigated by various authors (McAllester and Schapire, 2000; Orlitsky *et al.*, 2003; Orlitsky and Suresh, 2015; Arora *et al.*, 2018). Here we give the following negative result on lower bounding the entropy of a distribution by sampling.

Theorem 4.1. *Let B be any distribution-free high-confidence lower bound on $H(X;p_X)$ computed from a type $\mathcal{T}(S)$ with $S\sim p_X^N$.*

More specifically, let $B(\mathcal{T},\delta)$ be any real-valued function of a type $\mathcal{T}$ and a confidence parameter δ such that for any p_X, with probability at least $1-\delta$ over a draw of S from p_X^N, we have

$$H(X;p_X) \geq B(\mathcal{T}(S),\delta)$$

For any such bound, and for $N \geq 50$ and $k \geq 2$, for any p_X, with probability at least $1-\delta-1.01/k$ over the draw of S from p_X^N we have

$$B(\mathcal{T}(S),\delta) \leq \ln 2kN^2$$

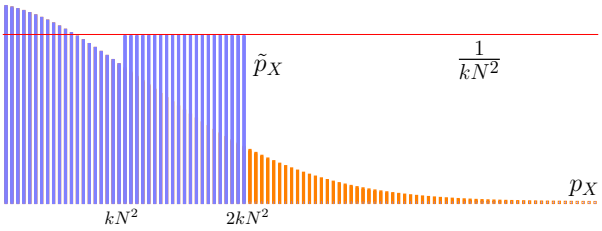

Figure 1: Construction of an adversarial distribution $\tilde{p}_X$ from p_X with entropy $H(X;\tilde{p}_X) \leq \ln 2kN^2$.

Proof. Consider a distribution p_X and $N \geq 100$. If the support of p_X has fewer than $2kN^2$ elements then $H(X;p_X) < \ln 2kN^2$ and by the premise of the theorem we have that, with probability at least $1-\delta$ over the draw of S, $B(\mathcal{T}(S),\delta) \leq H(X;p_X)$ and the theorem follows. If the support of p_X has at least $2kN^2$ elements then we sort the support of p_X into a (possibly infinite) sequence $x_1, x_2, \ldots$ so that $p_X(x_i) \geq p_X(x_{i+1})$. We then define a distribution $\tilde{p}_X$ on the elements $x_1 \ldots x_{2kN^2}$ by

$$\tilde{p}_X(x_i) = \begin{pmatrix} p_X(x_i) & \text{for } i \leq kN^2 \\ \\ \frac{\mu}{kN^2} & \text{for } kN^2 < i \leq 2kN^2 \end{pmatrix}$$

where $\mu := \sum_{j>kN^2} p_X(x_j)$. See Figure 1 for illustration. We will let $\text{Small}(S)$ denote the event that $B(\mathcal{T}(S),\delta) \leq \ln 2kN^2$ and let $\text{Pure}(S)$ abbreviate the event that no element x_i for $i > kN^2$ occurs twice in the sample. Since $\tilde{p}_X$ has a support of size $2kN^2$ we have $H(X;\tilde{p}_X) \leq \ln 2kN^2$. Applying the premise of the lemma to $\tilde{p}_X$ gives

$$\Pr_{S\sim\tilde{p}_X^N}(\text{Small}(S)) \geq 1-\delta \tag{9}$$

For a type $\mathcal{T}$ let $\Pr_{S\sim P^N}(\mathcal{T})$ denote the probability over drawing $S \sim P^N$ that $\mathcal{T}(S) = \mathcal{T}$. We now have

$$\Pr_{S\sim p_X^N}(\mathcal{T}(S)|\text{Pure}(S)) = \Pr_{S\sim\tilde{p}_X^N}(\mathcal{T}(S)|\text{Pure}(S))$$

This gives the following.

$$\begin{aligned}
&\Pr_{S\sim p_X^N}(\text{Small}(S)) \\
&\geq \Pr_{S\sim p_X^N}(\text{Pure}(S) \wedge \text{Small}(S)) \\
&= \Pr_{S\sim p_X^N}(\text{Pure}(S)) \Pr_{S\sim p_X^N}(\text{Small}(S) \mid \text{Pure}(S)) \\
&= \Pr_{S\sim p_X^N}(\text{Pure}(S)) \Pr_{S\sim \tilde{p}_X^N}(\text{Small}(S) \mid \text{Pure}(S)) \\
&\geq \Pr_{S\sim p_X^N}(\text{Pure}(S)) \Pr_{S\sim \tilde{p}_X^N}(\text{Pure}(S) \wedge \text{Small}(S)) \qquad (10)
\end{aligned}$$

For $i > kN^2$ we have $\tilde{p}_X(x_i) \leq 1/(kN^2)$ which gives

$$\Pr_{S\sim\tilde{p}_X^N}(\text{Pure}(S)) \geq \prod_{j=1}^{N-1}\left(1-\frac{j}{kN^2}\right)$$

Using $1-z \geq e^{-1.01z}$ for $z \leq 1/100$ we have the following birthday paradox calculation.

$$\begin{aligned}
\ln \Pr_{S\sim\tilde{p}_X^N}(\text{Pure}(S)) &\geq -\frac{1.01}{kN^2}\sum_{j=1}^{N-1} j \\
&= -\frac{1.01}{kN^2}\frac{(N-1)N}{2} \\
&\geq -\frac{.505}{k}
\end{aligned}$$

Therefore

$$\Pr_{S\sim\tilde{p}_X^N}(\text{Pure}(S)) \geq e^{-.505/k} \geq 1-\frac{.505}{k} \tag{11}$$

Applying the union bound to (9) and (11) gives

$$\Pr_{S\sim\tilde{p}_X^N}(\text{Pure}(S) \wedge \text{Small}(S)) \geq 1-\delta-\frac{.505}{k} \tag{12}$$

By a derivation similar to that of (11) we get

$$\Pr_{S\sim p_X^N}(\text{Pure}(S)) \geq 1-\frac{.505}{k} \tag{13}$$

Combining (10), (12) and (13) gives

$$\Pr_{S\sim p_X^N}(\text{Small}(S)) \geq 1-\delta-\frac{1.01}{k}$$

□

Again, we note the importance of the fact that the entropy lower bound B is distribution-free: this allows us to construct an adversarial distribution $\tilde{p}_X$ with small entropy and apply the premise to upper bound B (9). Theorem 1.1 is derived from Theorem 4.1 by choosing appropriate hyperparameter values δ, k and using the fact that a lower bound on mutual information implies a lower bound on entropy.

5 TOWARD ACCURATE MEASUREMENT OF MUTUAL INFORMATION

Our main contribution is a set of fundamental statistical limitations on measuring a lower bound on mutual information implied by the difficulty of measuring a lower bound on KL divergence (Theorem 3.1) or entropy (Theorem 4.1). But it is natural to ask: then how can we achieve an accurate measurement of mutual information? As a complementary piece of contribution, in this section we explore a new estimator for mutual information that, while lacking formal guarantees, does not suffer from the same limitations.

5.1 Mutual Information as a Difference of Entropies

Since mutual information can be expressed as a difference of entropies (7), the problem of measuring mutual information can be reduced to the problem of measuring entropies. More specifically, we write mutual information as

$$\begin{aligned} &I(X, Y; p_{XY}) \\ &= \inf_{q_X} H(p_X, q_X) - \inf_{q_{X|Y}} H(p_{X|Y}, q_{X|Y}) \end{aligned} \tag{14}$$

where we use the fact that the entropy $H(X; p_X)$ is upper bounded by the cross entropy between p_X and q_X, denoted $H(p_X, q_X)$, for all distributions q_X (similarly for the conditional entropy $H(X|Y; p_{XY})$). They are equal iff $q_X = p_X$ and we have

$$H(X; p_X) = \inf_{q_X} H(p_X, q_X) \tag{15}$$

Note that the upper-bound guarantee for the cross-entropy estimator (15) yields neither an upper bound nor a lower bound guarantee for a difference of entropies (14). However, we give theoretical evidence below that, unlike lower bound estimators, upper bound cross-entropy estimators can meaningfully estimate large entropies from feasible samples.

Cross entropy estimation. The statistical limitations on distribution-free high-confidence lower bounds on entropy do not arise for cross-entropy upper bounds. For upper bounds we can show that naive sample estimates of the cross-entropy loss produce meaningful (large entropy) results. The empirical cross-entropy loss computed on samples $x_1 \dots x_N$ from a population distribution p_X is

$$\widehat{H}^N(p_X, q_X) = \frac{1}{N} \sum_{i=1}^{N} -\ln q_X(x_i) \tag{16}$$

where q_X is viewed as a model of p_X. We can bound the true loss of q_X by ensuring a minimum probability $e^{-F_{\max}}$ where $F_{\max}$ is then the maximum possible log loss in the cross-entropy objective.[2] Given a loss bound of $F_{\max}$, $\widehat{H}^N(p_X, q_X)$ is just the standard sample mean estimator of an expectation of a bounded variable. In this case we have the following standard confidence interval derived from the Chernoff bound.

Theorem 5.1. *For any population distribution p_X, and model distribution q_X with $-\ln q_X(x)$ bounded to the interval $[0, F_{\max}]$, with probability at least $1 - \delta$ over the draw of $x_1 \dots x_N \sim p_X$ we have*

$$H(p_X, q_X) \in \widehat{H}^N(p_X, q_X) \pm F_{\max} \sqrt{\frac{\ln \frac{2}{\delta}}{2N}}$$

[2] In language modeling a loss bound exists for any model that ultimately backs off to a uniform distribution on characters.

Thus unlike high-confidence distribution-free lower bounds, high-confidence distribution-free upper bounds on entropy can approach the true cross entropy at the modest sample rate of $1/\sqrt{N}$ even when the true cross entropy is large.[3]

5.2 Experiments

We present experiments with the proposed difference-of-entropies (DoE) estimator to gain a better understanding of its empirical behavior. First, we compare DoE with existing lower-bound estimators in a standard synthetic setting based on correlated Gaussians. Next, we apply DoE on the task of measuring mutual information between related articles and translation pairs and show an evidence of large mutual information.

In the following, DoE computes an empirical estimate of (14) by taking iid samples $(x_1, y_1) \dots (x_N, y_N)$ from p_{XY} and computing

$$\begin{aligned} \widehat{H}(p_X, q_X) &= \inf_{q_X} \frac{1}{N} \sum_{i=1}^{N} -\log q_X(x_i) \\ \widehat{H}(p_{X|Y}, q_{X|Y}) &= \inf_{q_{X|Y}} \frac{1}{N} \sum_{i=1}^{N} -\log q_{X|Y}(x_i|y_i) \\ \widehat{I}(X, Y; p_{XY}) &= \widehat{H}(p_X, q_X) - \widehat{H}(p_{X|Y}, q_{X|Y}) \end{aligned} \tag{17}$$

where each minimization corresponds to fitting a probabilistic model on the samples with the cross-entropy loss.

5.2.1 Synthetic Experiments

Following Poole *et al.* (2019), we define random variables $X, Y \in \mathbb{R}^d$ where (X_i, Y_i) are standard normal with correlation $\rho \in [-1, 1]$. It can be checked that $I(X, Y) = -(d/2)\ln(1 - \rho^2)$. We use $d = 128$ and vary ρ to experiment with different values of mutual information. We compare with the following lower-bound estimators: the DV bound (2), MINE (i.e., DV with an "improved gradient estimator") (Belghazi *et al.*, 2018), NWJ (Nguyen *et al.*, 2010), NWJ estimated by optimizing Jensen-Shannon divergence (NWJ (JS)), CPC (Oord *et al.*, 2018), and a nonlinear interpolation between NWJ and CPC (NWJ+CPC). We refer the reader to Poole *et al.* (2019) for a detailed exposition of these estimators. We parameterize the distributions in DoE (i.e., q_X and $q_{X|Y}$ in (17)) by isotropic Gaussian (correct) or logistic (misspecified).

Table 1 shows mutual information estimates given by these estimators. All estimators are trained for 3,000 steps where at every step they use $N = 128$ samples drawn from p_{XY} to update their weights. We tune the hyperparame-

[3] It is also possible to give PAC-Bayesian bounds on $H(p_X, q_X^\theta)$ as functions of the parameters θ of q_X. See the supplementary material for details.

Table 1: Estimates of mutual information under different estimators. Each estimator is trained for 3,000 steps where at every step it receives $N = 128$ samples of (X, Y) and optimizes its weights; it is fully tuned over hyperparameter choices with respect to its final estimate. In each row, we boldface the estimate closest to the ground-truth mutual information.

DV	MINE	NWJ	NWJ (JS)	CPC	CPC+NWJ	DoE (Gaussian)	DoE (Logistic)	$I(X,Y)$	$\ln N$
2.72	2.57	1.99	1.50	2.73	2.77	4.19	**4.13**	4.13	4.85
10.27	9.38	9.25	5.55	4.82	8.18	18.38	**18.42**	18.41	4.85
61.96	34.56	50.46	13.41	4.85	10.45	**104.18**	104.16	106.29	4.85

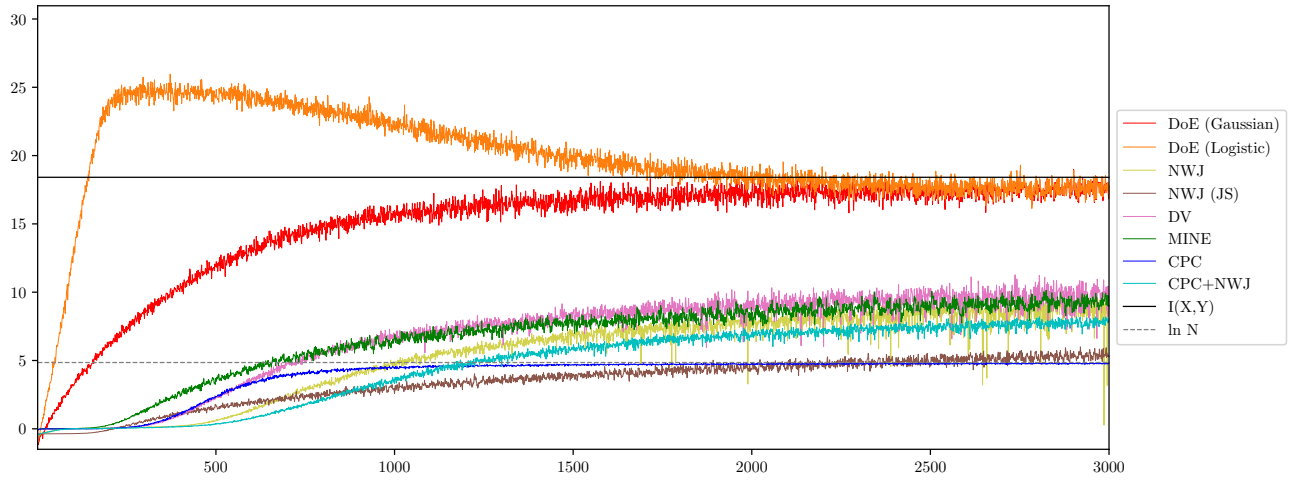

Figure 2: A plot of mutual information estimates during training. At every training step (x-axis), each estimator receives a (shared) minibatch of $N = 128$ samples, computes the current estimate (y-axis), and updates weights. We show the best-case scenario for each estimator by fully tuning its hyperparameters with respect to the final estimate.

ters of each estimator (e.g., learning rate, number of hidden units, initialization strategies, estimator-specific hyperparameters such as the mixing weights in NWJ+CPC and MINE) to minimize $|I(X,Y) - \hat{m}|$ where $\hat{m}$ is the final estimate. Thus the results assume an oracle that gives an optimal configuration. All details can be found in the code: `https://github.com/karlstratos/doe`.

It is clear that (1) DoE obtains the most accurate estimates whether $I(X,Y) > \ln N$ or $I(X,Y) \leq \ln N$, (2) DoE is the only estimator that achieves accurate estimates when underlying mutual information is large, and (3) either Gaussian or logistic parameterization of DoE yields accurate estimates. Furthermore, Table 1 does not show the fact that DV, MINE, and NWJ are highly unstable (especially in the large mutual information setting). In particular, while it seems that they can achieve estimates much larger than $\ln N$, it is not representative of their general performance which is fraught with numerical overflow/underflow problems leading to completely off estimates. Poole *et al.* (2019) discuss the large variance of these estimators. In contrast, DoE is based on the standard cross-entropy loss and very stable.

Figure 2 shows a plot of the best training session for each estimator. DoE is a clear outlier as the only accurate estimator of mutual information. Unlike lower-bound estimators, DoE can approach $I(X,Y)$ from above or below.

5.2.2 Mutual Information Between Articles and Translations

We consider two datasets for the choice of (X, Y):

1. Related news article pairs extracted from the Who-Did-What dataset (Onishi *et al.*, 2016)
2. English-German translation pairs extracted from the IWSLT 2014 dataset

We expect that mutual information is large in either setting: a random article (sentence) has large entropy, but given a related article (translation) the uncertainty is drastically reduced. We use a standard LSTM language model for q_X and a standard attention-based translation model for $q_{X|Y}$. More details of the experiments can be found in the supplementary material.

Table 2 shows the estimates of mutual information on the test portion of data. We use log base 2 to accommodate

Table 2: Estimates of mutual information (in bits) on article pairs and translation pairs based on the difference of entropies (17).

distribution p_{XY}	(17)
related article pairs	120.34
shuffled article pairs	-2.38
translation pairs	54.72
shuffled translation pairs	-2.64

the bit interpretation of entropies (rather than nats). We see that mutual information is estimated to be over 120 bits on related article pairs and 54 bits on translation pairs. Mutual information is estimated to be close to zero for shuffled pairs: this shows that the estimator can also handle small mutual information.

6 RELATED WORK

We make a few additional remarks on related work to better contextualize our work. In the continuous setting, a classical approach to estimating mutual information is based on computing the average log of the distance to the k-th nearest neighbor in samples (Kraskov *et al.*, 2004). Gao *et al.* (2015) show that this estimator suffers exponential sample complexity and propose more refined nearest-neighbor methods (Gao *et al.*, 2015). In contrast, we establish that serious statistical limitations are inherent to the measurement of mutual information no matter what estimator is used.

There is a line of work that develops efficient estimators for entropy by assuming distributions with very small support—distributions with support smaller than the sample size. Valiant and Valiant (2011) show that it is possible to achieve an optimal sample rate of $O(n/\ln n)$ where n is the support size. Past work on analyzing minimax bounds likewise assume small support (Jiao *et al.*, 2015; Han *et al.*, 2015; Kandasamy *et al.*, 2015). In this case, the entropy of the distribution cannot be larger than the log of the number of samples. This is in agreement with, but does not imply, our results. We are interested in the large entropy setting, such as a distribution over all possible images or articles. We cannot have the number of samples equal to the number of possible images or articles.

As discussed in depth in Section 2, we are motivated by the approach in MINE which measures the DV bound to estimate and maximize mutual information (Belghazi *et al.*, 2018). CPC is another notable example that illustrates the statistical limitations of measuring mutual information (Oord *et al.*, 2018). CPC maximizes a lower bound on mutual information through noise contrastive estimation: it is shown that this lower bound cannot be larger than $\ln k$ where k is number of negative samples used in the contrastive choice. Complementary to our work, a recent work by Poole *et al.* (2019) investigates tradeoffs between bias and variance in estimating variational bounds on mutual information.

There is a class of representation learning methods such as Brown clustering (Brown *et al.*, 1992) and the information bottleneck method (Tishby *et al.*, 1999) that maximize a lower bound on mutual information given by the data processing inequality (DPI). In these methods, we learn "coding" functions (C, C') by optimizing the objective

$$\max_{C,C'} I(C(X), C'(Y); p_{XY}) \leq I(X, Y; p_{XY})$$

where the inequality is by the DPI. Information theoretic co-training (McAllester, 2018) considers a similar lower bound and has been shown to be useful for label induction in speech and text (Stratos, 2019). Measuring these lower bounds is subject to the same limitations presented in this paper.

7 CONCLUSIONS

Maximizing mutual information is well motivated as a method of unsupervised pretraining of representations that maintain semantic signal while dropping uninformative noise. However, measuring and maximizing mutual information from finite data is a difficult training objective. In this paper, we have shown serious statistical limitations inherent to measuring lower bounds on various information theoretic measures including KL divergence, entropy, and mutual information. We have also given theoretical arguments that representing mutual information as a difference of entropies, and estimating those entropies by minimizing cross-entropy loss, is a more statistically justified approach than maximizing a lower bound on mutual information.

Unfortunately cross-entropy upper bounds on entropy fail to provide either upper or lower bounds on mutual information—mutual information is a difference of entropies. We cannot rule out the possible existence of superintelligent models, models beyond current expressive power, that dramatically reduce cross-entropy loss. Lower bounds on entropy can be viewed as proofs of the nonexistence of superintelligence. We should not surprised that such proofs are infeasible.

References

Arora, S., Risteski, A., and Zhang, Y. (2018). Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. (2018). Mutual information neural estimation. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on*

Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540, StockholmsmÃd'ssan, Stockholm Sweden. PMLR.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, **7**(6), 1129–1159.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18**(4), 467–479.

Donsker, M. and Varadhan, S. (1983). Asymptotic evaluation of certain markov process expectations for large time, iv. *Communications on Pure and Applied Mathematics*, **36**(2), 183–212.

Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*.

Gao, S., Ver Steeg, G., and Galstyan, A. (2015). Efficient estimation of mutual information for strongly dependent variables. In *Artificial Intelligence and Statistics*, pages 277–286.

Han, Y., Jiao, J., and Weissman, T. (2015). Minimax estimation of discrete distributions under l_1 loss. *IEEE Transactions on Information Theory*, **61**(11), 6343–6354.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, **61**(5), 2835–2885.

Kandasamy, K., Krishnamurthy, A., Poczos, B., Wasserman, L., *et al.* (2015). Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, pages 397–405.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review E*, **69**(6), 066138.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

McAllester, D. (2013). A pac-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307:2118*.

McAllester, D. (2018). Information theoretic co-training. *arXiv preprint arXiv:1802.07572*.

McAllester, D. A. and Schapire, R. E. (2000). On the convergence rate of good-turing estimators. In *COLT*, pages 1–6.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, **56**(11), 5847–5861.

Onishi, T., Wang, H., Bansal, M., Gimpel, K., and McAllester, D. (2016). Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Orlitsky, A. and Suresh, A. T. (2015). Competitive distribution estimation: Why is good-turing good. In *Advances in Neural Information Processing Systems*, pages 2143–2151.

Orlitsky, A., Santhanam, N., and Zhang1, J. (2003). Always good turing: Asymptotically optimal probability estimation. *Science*, **302**(5644).

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. (2019). On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180.

Stratos, K. (2019). Mutual information maximization for simple and accurate part-of-speech induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Valiant, G. and Valiant, P. (2011). Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 685–694. ACM.

A PROOF OF THEOREM 2.1

For any distribution r_X over X, we can write

$$\begin{aligned} D_{\mathrm{KL}}(p_X||q_X) &= \mathop{\mathbb{E}}_{x\sim p_X}\left[\ln\frac{r_X(x)}{q_X(x)}\right] + D_{\mathrm{KL}}(p_X||r_X) \\ &\geq \mathop{\mathbb{E}}_{x\sim p_X}\left[\ln\frac{r_X(x)}{q_X(x)}\right] \end{aligned} \tag{18}$$

Let $f : \mathcal{X} \to \mathbb{R}$ be a bounded function and define

$$r_X(x) = \frac{q_X(x)e^{f(x)}}{\mathop{\mathbb{E}}_{x\sim q_X}\left[e^{f(x)}\right]} \qquad \forall x \in \mathcal{X}$$

which is a valid distribution over X. Plugging this into the lower bound in (18), we have

$$\mathop{\mathbb{E}}_{x\sim p_X}\left[\ln\frac{r_X(x)}{q_X(x)}\right] = \mathop{\mathbb{E}}_{x\sim p_X}\left[f(x)\right] - \ln \mathop{\mathbb{E}}_{x\sim q_X}\left[e^{f(x)}\right] \tag{19}$$

By (18), the supremum of (19) over the choice of f is precisely the KL divergence between p_X and q_X. It can be easily verified that an optimal f is given by

$$f(x) = \ln\frac{p_X(x)}{q_X(x)} \qquad \forall x \in \mathcal{X}$$

Since (19) is invariant to translation of f, without loss of generality we can assume that the range of f is bounded in $[0, F_{\max}]$ for some constant $F_{\max}$.

B MUTUAL INFORMATION AS THE SUPREMUM OVER BINNINGS

We now show that the mutual information $I(X, Y; p_{XY})$ for X and Y continuous can be expressed as the supremum of $I(C(X), C'(Y); p_{XY})$ over discrete binnings of the continuous space. We first consider the case where $X, Y \in \mathbb{R}$ and where the mutual information can be written as a Riemann integral over densities.

$$\begin{aligned} &I(X, Y; p_{XY}) \\ &= \int p_{XY}(x, y) \ln\frac{p_{XY}(x, y)}{p_X(x)p_Y(y)}\, dx\, dy \\ &= \lim_{\epsilon\to 0} \sum_{i,j\in\mathbb{Z}} p_{XY}(i\epsilon, j\epsilon)\ \ln\frac{p_{XY}(i\epsilon, j\epsilon)}{p_X(i\epsilon)p_Y(j\epsilon)}\epsilon^2 \end{aligned}$$

where $\mathbb{Z}$ is the set of all integers. For each $i \in \mathbb{Z}$, define the half-open interval $C_{i,\epsilon} := [i\epsilon, (i + 1)\epsilon)$. The probability of the interval is approximately $\epsilon p_X(i\epsilon)$ under p_X (similarly for p_Y and p_{XY}). Therefore we can write the last expression as

$$\begin{aligned} &\lim_{\epsilon\to 0} \sum_{i,j\in\mathbb{Z}} p_{XY}(C_{i,\epsilon} \times C_{j,\epsilon})\ \ln\frac{p_{XY}(C_{i,\epsilon} \times C_{j,\epsilon})}{p_X(C_{i,\epsilon})p_Y(C_{j,\epsilon})} \\ &= \lim_{\epsilon\to 0} \sum_{i,j\in\mathbb{Z}} p_{I_\epsilon J_\epsilon}(i, j)\ \ln\frac{p_{I_\epsilon J_\epsilon}(i, j)}{p_{I_\epsilon}(i)p_{J_\epsilon}(j)} \\ &= \lim_{\epsilon\to 0} I(I_\epsilon, J_\epsilon; p_{I_\epsilon J_\epsilon}) \end{aligned}$$

where (I_ϵ, J_ϵ) denote the indices (i, j) such that $x \in C_{i,\epsilon}$ and $y \in C_{j,\epsilon}$ for $(x, y) \sim p_{XY}$.

This proof immediately generalizes to higher dimensions where the mutual information can be expressed as a Riemann integral. We believe that this statement remains true for arbitrary measures on product spaces where the mutual information is finite. However the proof for this extremely general case appears to be nontrivial.

C PAC-BAYESIAN BOUNDS

The PAC-Bayesian bounds apply to "broad basin" losses and loss estimates such as the following:

$$\begin{aligned} H_\sigma(S, q_X^\theta) &= \mathop{\mathbb{E}}_{x\sim p_X}\left[\mathop{\mathbb{E}}_{\epsilon\sim N(0,\sigma I)}\left[-\ln q_X^{\theta+\epsilon}(x)\right]\right] \\ \widehat{H}_\sigma(S, q_X^\theta) &= \frac{1}{|S|}\sum_{x\in S} \mathop{\mathbb{E}}_{\epsilon\sim N(0,\sigma I)}\left[-\ln q_X^{\theta+\epsilon}(x)\right] \end{aligned}$$

Under mild smoothness conditions on $q_X^\theta(x)$ as a function of θ we have

$$\begin{aligned} \lim_{\sigma\to 0} H_\sigma(p_X, q_X^\theta) &= H(p_X, q_X^\theta) \\ \lim_{\sigma\to 0} \widehat{H}_\sigma(S, q_X^\theta) &= \widehat{H}(S, q_X^\theta) \end{aligned}$$

An L_2 PAC-Bayesian generalization bound (McAllester, 2013) gives that for any parameterized class of models and any bounded notion of loss, and any $\lambda > 1/2$ and $\sigma > 0$, with probability at least $1 - \delta$ over the draw of S from p_X^N we have the following simultaneously for all parameter vectors θ.

$$\begin{aligned} &H_\sigma(p_X, q_X^\theta) \\ &\leq \frac{1}{1-\frac{1}{2\lambda}}\left(\widehat{H}_\sigma(S, q_X^\theta) + \frac{\lambda F_{\max}}{N}\left(\frac{||\theta||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right) \end{aligned}$$

It is instructive to set $\lambda = 5$ in which case the bound becomes.

$$\begin{aligned} &H_\sigma(p_X, q_X^\theta) \\ &\leq \frac{10}{9}\left(\widehat{H}_\sigma(S, q_X^\theta) + \frac{5F_{\max}}{N}\left(\frac{||\theta||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right) \end{aligned}$$

	train (tgt)	train (src)
# articles	68348	68348
vocab size	100001	87941
# words	20271664	19072167
avg length	296	279
max length	400	400
min length	10	12

Table 3: Training statistics of the article pairs

	train (tgt)	train (src)
# sentences	160239	160239
vocab size	24726	35445
# words	3275729	3100720
avg length	20	19
max length	175	172
min length	2	2

Table 4: Training statistics of the translation pairs

While this bound is linear in $1/N$, and tighter in practice than square root bounds, note that there is a small residual gap when holding λ fixed at 5 while taking $N \rightarrow \infty$. In practice the regularization parameter λ can be tuned on holdout data. One point worth noting is the form of the dependence of the regularization coefficient on $F_{\max}$, N and the basin parameter σ.

It is also worth noting that the bound can be given in terms of "distance traveled" in parameter space from an initial (random) parameter setting θ_0.

$$H_\sigma(p_X, q_X^\theta) \leq \frac{10}{9}\left(\widehat{H}_\sigma(S, q_X^\theta) + \frac{5F_{\max}}{N}\left(\frac{||\theta-\theta_0||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right)$$

Evidence is presented in Dziugaite and Roy (2017) that the distance traveled bounds are tighter in practice than traditional L_2 generalization bounds.

D EXPERIMENT DETAILS

Article pairs. We take pairs from the Who-Did-What dataset (Onishi *et al.*, 2016). The pairs in this dataset were constructed by drawing articles from the LDC Gigaword newswire corpus. A first article is drawn at random and then a list of candidate second articles is drawn using the first sentence of the first article as an information retrieval query. A second article is selected from the candidates using criteria described in Onishi *et al.* (2016), the most significant of which is that the second article must have occurred within a two week time interval of the first. The training statistics of this dataset after preprocessing is given in Table 3.

Translation pairs. Our translation pairs consists of English-German sentence pairs extracted from the IWSLT 2014 dataset. The training statistics of this dataset after preprocessing is given in Table 4.

Model. We train an LSTM encoder-decoder model where the decoder doubles as both the decoder of a translation model and a language model. The decoder is a left-to-right 2-layer LSTM in which a single word embedding matrix is used for both input embeddings and the softmax predictions. When this model is trained as a language model on PTB using standard hyperparameter values it achieves test perplexity of 72.26. The encoder is a separate left-to-right 2-layer LSTM using the same word embeddings as the decoder. We use the input-feeding attention archietecture of Luong *et al.* (2015).

The model is trained using SGD and batch size 10 with no BPTT-style truncation. The dimension of the input/hidden states is 900 (thus 1800 for the input-feeding decoder). We use step-wise dropout with rate 0.65 on word embeddings and hidden states. The model is trained for 40 epochs and the model that achieves the best validation perplexity is selected. The sequence-level cross entropy is estimated as $\mathrm{SQXENT} = \frac{1}{M}\mathrm{NLL}$ where NLL is the negative log likelihood of the corpus and M is the total number of sequences in the corpus.

Mutual information is estimated by taking the difference in SQXENT between the language model and the translation model (17).For article pairs, we obtain

$$\widehat{I}(X, Y; p_{XY}) = 1131.74 - 1048.33 = 83.41$$

in nats which translates to 120.34 bits. For translation pairs, we obtain

$$\widehat{I}(X, Y; p_{XY}) = 81.73 - 43.80 = 37.9$$

in nats which translates to 54.72 bits.