FORMAL LIMITATIONS ON THE MEASUREMENT OF MUTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Maximum mutual information (MMI) predictive coding has recently been proposed as a compelling approach to self-supervised learning. Maximizing mutual information also arises in INFOMAX and the information bottleneck. Recent papers have used the Donsker-Varadhan (DV) lower bound on KL divergence as a tractable training surrogate when maximizing mutual information. We show here that any distribution-free high-confidence lower bound on mutual information requires a sample size exponential in the value of the bound. We also directly analyze the DV lower bound and show that the DV bound cannot be measured accurately without an exponential sample size. We propose instead to representing mutual information as a difference of entropies and of estimating entropies with cross-entropy upper bounds. A sample estimate of a cross-entropy converges to the true cross-entropy at the rate of $1/\sqrt{N}$.

1 INTRODUCTION

We consider the problem of maximizing mutual information. We are motivated by maximal mutual information (MMI) predictive coding McAllester (2018); Stratos (2018); Oord et al. (2018). Here the objective is to pretrain compressive representations of input signals, such as images or sound waves, from unlabeled data in such a way that the representation preserves predictive information (signal) but drops non-predictive information (noise).

One can formally define a version of MMI predictive coding by considering a population distribution on pairs (x, y) where we think of x as past raw sensory signals (images or sound waves) and y as a future sensory signal. We then consider the problem of learning stochastic coding functions C_x and C_y so as to maximize the mutual information $I(C_x(x), C_y(y))$ while limiting the entropies $H(C_x(x))$ and $H(C_y(y))$. Forms of MMI predictive coding have been independently introduced in McAllester (2018) under the name "information-theoretic cotraining" and in Oord et al. (2018) under the name "contrastive predictive coding" in reference to a particular optimization method.

A closely related framework is the information bottleneck (Tishby et al., 2000). Here one again assumes a population distribution on pairs (x, y). The objective is to learn a stochastic coding function C_x so as to maximize $I(C_x(x), y)$ while minimizing $I(C_x(x), x)$. Here one does not ask for a coding function on y and one does not limit $H(C_x(x))$.

Another related framework is INFOMAX (Linsker, 1988; Bell & Sejnowski, 1995; Hjelm et al., 2018). Here we consider a population distribution on a single random variable x. The objective is to learn a stochastic coding function C_x so as to maximize the mutual information $I(x, C_x(x))$ subject to some constraint or additional objective.

Both Oord et al. (2018) and Hjelm et al. (2018) maximize a lower bound on mutual information derived from the Donsker-Varaghan (DV) lower bound on KL-divergence (Donsker & Varadhan (1983)). This approach was developed in Belghazi et al. (2018) under the name MINE (Mutual Information Neural Estimation). Here we argue that using the DV lower bound for maximizing mutual information is fundamentally flawed in cases where the achievable mutual information is more than tens of bits.

In MMI predictive coding we seek to find code parameters that maximize the mutual information between two code values. In the remainder of this paper we consider a fixed setting of the code

parameters and consider methods of simply measuring the mutual information under the population distribution on codes determined by the fixed code parameters. We show that for any population sample of size N it is not possible to give a distribution-free high-confidence guarantee for a lower bound on mutual information larger than $2 \ln N$. Lower bounds are therefore poor estimates unless the sample size is exponential in the quantity of mutual information.

We argue that a better approach is to write mutual information as a difference of entropies and to estimate each of these entropies separately using cross-entropy upper bounds. In contrast with the DV lower bound, we show that when the log loss is bounded, and for a sufficiently expressive model class, the cross entropy upper bound converges to the true entropy at a rate of $1/\sqrt{N}$.

2 THE DONSKER-VARAGHAN LOWER BOUND

Mutual information can be written as a KL divergence.

$$I(X,Y) = KL(P_{X,Y}, P_X P_Y)$$

Here $P_{X,Y}$ is a joint distribution on the random variables X and Y and P_X and P_Y are the marginal distributions on X abd Y respectively. The DV lower bound applies to KL-divergence generally. To derive the DV bound we start with the following observation for any distributions P, Q, and G on the same support. Our theoretical analyses will assume discrete distributions.

$$\begin{aligned} KL(P,Q) &= E_{z\sim P} \ln \frac{P(z)}{Q(z)} \\ &= E_{z\sim P} \ln \left(\frac{G(z)}{Q(z)} \frac{P(z)}{G(z)} \right) \\ &= E_{z\sim P} \ln \frac{G(z)}{Q(z)} + KL(P,G) \\ &\geq E_{z\sim P} \ln \frac{G(z)}{Q(z)} \end{aligned} \tag{1}$$

Note that (1) achieves equality for $G(z) = P(z)$ and hence we have

$$KL(P,Q) = \sup_G E_{z\in P} \ln \frac{G(z)}{Q(z)} \tag{2}$$

Here we can let G be a parameterized model such that $G(z)$ can be computed directly. However, we are interested in $KL(P_{X,Y}, P_X P_Y)$ where our only access to the distribution P is through sampling. If we draw a pair (x, y) and ignore y we get a sample from P_X. We can similarly sample from P_Y. So we are interested in a KL-divergence $KL(P,Q)$ where our only access to the distributions P and Q is through sampling. Note that we cannot evaluate (1) by sampling from P because we have no way of computing $Q(z)$. But through a change of variables we can convert this to an expression restricted to sampling from Q. More specifically we define $G(z)$ in terms of an unconstrained function $F(z)$ as

$$G(z) = \frac{1}{Z} Q(z) e^{F(z)} \qquad Z = \sum_z Q(z) e^{F(z)} = E_{z\sim Q}\, e^{F(z)} \tag{3}$$

Substituting (3) into (2) gives

$$KL(P,Q) = \sup_F E_{z\sim P}\, F(z) - \ln E_{z\sim Q}\, e^{F(z)} \tag{4}$$

Equation (4) is the Donsker-Varadhan lower bound. Applying this to mutual information we get

$$\begin{aligned} I(X,Y) &= KL(P_{X,Y}, P_X P_Y) \\ &= \sup_F E_{x,y\sim P_{X,Y}}\, F(x,y) - \ln E_{x\sim P_X,\, y\sim P_Y}\, e^{F(x,y)} \end{aligned} \tag{5}$$

This is the equation underlying the MINE approach to maximizing mutual information (Belghazi et al., 2018). It would seem that we can estimate both terms in (5) through sampling and be able to maximize $I(X, Y)$ by stochastic gradient ascent on this lower bound.

3 STATISTICAL LIMITATIONS OF KL-DIVERGENCE LOWER BOUNDS

In this section we show that the DV bound (4) cannot be used to measure KL-divergences of more than tens of bits. In fact we will show that no high-confidence distribution-free lower bound on KL divergence can be used for this purpose.

As a first observation note that (4) involves $E_{z\sim Q}\ e^{F(z)}$. 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. It should be noted that the optimal value for $F(z)$ in (4) is $\ln(P(z)/Q(z))$ in which case the right hand side of (4) simplifies to $KL(P, Q)$. But for large KL divergence we will have that $F(z) = \ln(P(z)/Q(z))$ is typically hundreds of bits and this is exactly the case where $E_{z\sim Q}\ e^{F(z)}$ cannot be measured by sampling from Q. If $E_{z\sim Q}\ e^{F(z)}$ is dominated by events that will never occur in sampling from Q then the optimization of F through the use of (4) and sampling from Q cannot possibly lead to a function $F(z)$ that accurately models the desired function $\ln(P(z)/Q(z))$.

To quantitatively analyze the risk of unseen outlier events we will make use of the following simple lemma where we write $P_{z\sim Q}[\Phi[z]]$ for the probability over drawing z from Q that the statement $\Phi[z]$ holds.

Outlier Risk Lemma: For a sample $S \sim Q^N$ with $N \geq 2$, and a property $\Phi[z]$ such that $P_{z\sim Q}[\Phi[z]] \leq 1/N$, the probability over the draw of S that no $z \in S$ satisfies $\Phi[z]$ is at least 1/4.

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

We can use the outlier risk lemma to perform a quantitative risk analysis of the DV bound (4). For a fair comparison with our analysis of cross-entropy estimators below, we will limit the outlier risk by bounding $F(z)$ to the interval $[0, F_{\max}]$. The strongest value of (4) when computed from samples of P and Q will occur when we have $F(z) = F_{\max}$ for draws from P and $F(z) = 0$ for draws from Q. In this case plugging in sample estimates of $E_{z\sim P}\ F(z)$ and $E_{z\sim Q}\ e^{F(z)}$ gives an apparent lower bound on $KL(P, Q)$ of $F_{\max}$. But by the outlier risk lemma we cannot rule out the possibility of a value z that is unseen in the sample from Q but with $Q(z) = 1/N$ and $F(z) = F_{\max}$. For a distribution where such an event exists we have

$$E_{z\sim q}\ e^{F(z)} \geq \frac{1}{N} e^{F_{\max}}.$$

Plugging this into (4) gives.

$$\begin{aligned} KL(P, Q) &\geq F_{\max} - \ln \frac{e^{F_{\max}}}{N} \\ &= \ln N \end{aligned}$$

So taking unseen outlier risk into account the strongest possible distribution-free high confidence bound derivable from (4) is $\ln N$. Providing a distribution-free high-confidence lower bound on KL-divergence by use of the DV-bound requires a sample exponential in the size of desired bound.

Our negative results can be strengthened by consider the preliminary bound (1) where $G(z)$ is viewed as a model of $P(z)$. We can consider the extreme case of perfect modeling of the population P with a model $G(z)$ where $G(z)$ is computable. In this case we have essentially complete access to the distribution P. But even in this setting we have the following negative result.

Theorem 1 *Let B be any distribution-free high-confidence lower bound on KL(P,Q) computed with complete knowledge of P but only a sample from Q.*

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

$$KL(P, Q) \geq B(P, S, \delta).$$

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

$$B(P, S, \delta) \leq \ln N.$$

Proof. Consider distributions P and Q and $N \geq 2$. Define $\tilde{Q}$ by

$$\tilde{Q}(z) = \left(1 - \frac{1}{N}\right) Q(z) + \frac{1}{N} P(z).$$

We now have

$$KL(P, \tilde{Q}) \leq \ln N$$

and

$$P_{S \sim \tilde{Q}^N}[B(P, S, \delta) \leq KL(P, \tilde{Q})] \geq 1 - \delta.$$

The distribution $\tilde{Q}$ equals the marginal on z of a distribution on pairs (s, z) where s is the value of Bernoulli variable with bias $1/N$ such that if $s = 1$ then z is drawn from P and otherwise z is drawn from Q. By the outlier risk lemma the probability that all coins are zero is at least 1/4. Conditioned on all coins being zero the distributions $\tilde{Q}^N$ and Q^N are the same. Let $\mathrm{Pure}(S)$ represent the event that all coins are 0 and let $\mathrm{Small}(S)$ represent the event that $B(P, S, \delta) \leq \ln N$. We now have

$$\begin{aligned}
P_{S \sim Q^N}[\mathrm{Small(S)}] &= P_{S \sim \tilde{Q}^N}[\mathrm{Small}(S) | \mathrm{Pure}(S)] \\
&= \frac{P_{S \sim \tilde{Q}^N}[\mathrm{Pure}(S) \wedge \mathrm{Small(S)}]}{P_{S \sim \tilde{Q}^N}[\mathrm{Pure}(S)]} \\
&\geq \frac{P_{S \sim \tilde{Q}^N}[\mathrm{Pure}(S)] - P_{S \sim \tilde{Q}^N}[\neg \mathrm{Small}(S)]}{P_{S \sim \tilde{Q}^N}[\mathrm{Pure}(S)]} \\
&\geq \frac{P_{S \sim \tilde{Q}^N}[\mathrm{Pure}(S)] - \delta}{P_{S \sim \tilde{Q}^N}[\mathrm{Pure}(S)]} \\
&= 1 - \frac{\delta}{P_{S \sim \tilde{Q}^N}[\mathrm{Pure}(S)]} \\
&\geq 1 - 4\delta.
\end{aligned}$$

4 Statistical Limitations on Entropy Lower Bounds

Mutual information is a special case of KL-divergence. It is possible that tighter lower bounds can be given in this special case. In this section we show similar limitations on lower bounding mutual information. We first note that a lower bound on mutual information implies a lower bound on entropy. The mutual information between X and Y cannot be larger than information content of X alone.

$$I(X, Y) = H(X) - H(X|Y) \leq H(X)$$

So a lower bound on $I(X, Y)$ gives a lower bound on $H(X)$. We show, somewhat informally, that any distribution-free high-confidence lower bound on entropy requires a sample size exponential in the size of the bound.

Consider sampling N items IID from a population distribution P. The strongest possible sample — the one guaranteeing the highest entropy — is a sample in which no item occurs twice. This should happen when we draw images or new stories being careful to maintain independence. If the sample contains N items and the distribution is uniformly distributed on N^2 items then the probability of never drawing the same item twice is approximately $1/\sqrt{e}$. We leave this birthday paradox calculation to the reader. A more complete discussion of birthday paradox lower bounds

on entropy can be found in (Arora et al. (2018)). So if all we know is that no item occurred twice it is perfectly possible that the data has been drawn from a distribution that is uniform on N^2 items in which case $H(P) = 2\ln(N)$. So for any sample of size N one can never give a high confidence guarantee that $H(P) > 2\ln N$.

5 Upper Bounds as Entropy Estimators

Since mutual information can be expressed as a difference of entropies, the problem of measuring mutual information can be reduced to the problem of measuring entropies. In this section we show that, unlike high-confidence distribution-free lower bounds, high-confidence distribution-free upper bounds on entropy can approach the true entropy at modest sample sizes even when the true entropy is large. More specifically we consider the cross-entropy upper bound.

$$\begin{aligned} H(P) &= E_{x\sim P} \ln \frac{1}{P(x)} \\ &= E_{x\sim P} \ln \left[\frac{1}{G(x)} \frac{G(x)}{P(x)} \right] \\ &= H(P,G) - KL(P,G) \\ &\leq H(P,G) \end{aligned}$$

For $G = P$ we get $H(P,G) = H(P)$ and hence we have

$$H(P) = \inf_G \; H(P,G)$$

In practice P is a population distribution and G is model of P. For example P might be a population distribution on paragraphs and G might be an autoregressive RNN language model. In practice G will be given by a network with parameters Φ. In this setting we have the following upper bound entropy estimator.

$$\hat{H}(P) = \inf_\Phi \; H(P, G_\Phi) \tag{6}$$

The gap between $\hat{H}(P)$ and $H(P)$ depends on the expressive power of the model class.

The statistical limitations on distribution-free high-confidence lower bounds on entropy do not arise for upper bounds. For upper bounds we can show that naive sample estimates of the cross entropy loss produce meaningful (large entropy) results. We first define the cross-entropy estimator from a sample S.

$$\hat{H}(S,G) = \frac{1}{|S|} \sum_{x\in S} -\ln \; G(x)$$

We can bound the loss of a model G by ensuring a minimum probability $e^{-F_{\max}}$ where $F_{\max}$ is then the maximum possible log loss in the cross entropy objective. In language modeling a loss bound exists for any model that ultimately backs off to a uniform distribution on characters. Given a loss bound of $F_{\max}$ we have that $\hat{H}(S,G)$ is just the standard sample mean estimator of an expectation of a bounded variable. In this case we have the following standard confidence interval.

Theorem 2 *For any population distribution P, and model distribution G with $-ln\ G(x)$ bounded to the interval $[0, F_{\max}]$, with probability at least $1-\delta$ over the draw of $S \sim P^N$ we have*

$$H(P,G) \in \hat{H}(S,G) \pm F_{\max} \sqrt{\frac{\ln \frac{2}{\delta}}{2N}}$$

It is also possible to give PAC-Bayesian bounds on $H(P, G_\Phi)$ that take into account the fact that G_Φ is typically trained so as to minimize the empirical loss on the training data. The PAC-Bayesian bounds apply to"broad basin" losses and loss estimates such as the following.

$$
\begin{aligned}
H_\sigma(S, G_\Phi) &= E_{x\sim P}\ E_{\epsilon\sim N(0,\sigma I)}\ -\ln\ G_{\Phi+\epsilon}(x) \\
\hat{H}_\sigma(S, G_\Phi) &= \frac{1}{|S|}\sum_{x\in S} E_{\epsilon\sim N(0,\sigma I)}\ -\ln\ G_{\Phi+\epsilon}(x)
\end{aligned}
$$

Under mild smoothness conditions on $G_\Phi(x)$ as a function of Φ we have

$$
\begin{aligned}
\lim_{\sigma\to 0}\ H_\sigma(P, G_\Phi) &= H(P, G_\Phi) \\
\lim_{\sigma\to 0}\ \hat{H}_\sigma(S, G_\Phi) &= \hat{H}(S, G_\Phi)
\end{aligned}
$$

An L2 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^N we have the following simultaneously for all parameter vectors Φ.

$$
H_\sigma(P, G_\Phi) \leq \frac{1}{1-\frac{1}{2\lambda}}\left(\hat{H}_\sigma(S, G_\Phi) + \frac{\lambda F_{\max}}{N}\left(\frac{||\Phi||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right)
$$

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

$$
H_\sigma(P, G_\Phi) \leq \frac{10}{9}\left(\hat{H}_\sigma(S, G_\Phi) + \frac{5F_{\max}}{N}\left(\frac{||\Phi||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right)
$$

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 $N \to \infty$. 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 $F_{\max}$ and N.

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 Φ_0.

$$
H_\sigma(P, G_\Phi) \leq \frac{10}{9}\left(\hat{H}_\sigma(S, G_\Phi) + \frac{5F_{\max}}{N}\left(\frac{||\Phi-\Phi_0||^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)\right)
$$

Evidence is presented in (Dziugaite & Roy (2017)) that the distance traveled bounds are tighter in practice than traditional L2 generalization bounds.

6 MMI PREDICTIVE CODING

Recall that in MMI predictive coding we assume a population distribution on pairs (x, y) where we think of x as past raw sensory signals (images or sound waves) and y as a future sensory signal. We then consider the problem of learning stochastic coding functions C_x and C_y that maximizes the mutual information $I(C_x(x), C_y(y))$ while limiting the entropies $H(C_x(x))$ and $H(C_y(y))$. Here we propose representing the mutual information as a difference of entropies.

$$
I(C_x(x), C_y(y)) = H(C_y(y)) - H(C_y(y)|C_x(x))
$$

When the coding functions are parameterized by a function Ψ, the above quantities become a function of Ψ. We can then formulate the following nested optimization problem.

$$
\begin{aligned}
\Psi^* &= \operatorname*{argmax}_{\Psi}\ \hat{H}(C_y(y);\ \Psi) - \hat{H}(C_y(y)|C_x(x); \Psi) \\
\hat{H}(C_y(y);\ \Psi) &= \inf_{\Theta}\ H(C_y(y), G_\Theta;\ \Psi) \\
\hat{H}(C_y(y)|C_x(x);\ \Psi) &= \inf_{\Phi}\ H(C_y(y), G'_\Phi|c_x(x);\ \Psi)
\end{aligned}
$$

The above quantities are expectations over the population distribution on pairs (x, y). In practice we have only a finite sample form the population. But the preceding section presents theoretical evidence that, unlike lower bound estimators, upper bound cross-entropy estimators can meaningfully estimate large entropies from feasible samples.

7 CONCLUSIONS

Maximum mutual information (MMI) predictive coding seems well motivated as a method of unsupervised pretraining of representations that maintain semantic signal while dropping uninformative noise. However, the maximization of mutual information is a difficult training objective. We have 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 non-existence of superintelligence. We should not surprised that such proofs are infeasible.

REFERENCES

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do gans learn the distribution? some theory and empirics. *ICLR*, 2018.

Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R Devon Hjelm, and Aaron Courville. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

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

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

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv Preprint arXive:1703.11008*, 2017.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

David McAllester. A pac-bayesian tutorial with a dropout bound. *arXiv reprint arXiv:1307:2118*, 2013.

David McAllester. Information Theoretic Co-Training. *arXiv preprint arXiv:1802.07572*, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Karl Stratos. Accurate part-of-speech induction by inter-annotator agreement. *arXiv preprint arXiv:1804.07849*, 2018.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.