On variational lower bounds of mutual information

Ben Poole[1], **Sherjil Ozair**[1,2], **Aäron van den Oord**[3], **Alexander A. Alemi**[1], **George Tucker**[1]
[1]Google Brain, [2]MILA, [3]DeepMind
{pooleb, sherjilozair, avdnoord, alemi, gjt}@google.com

Abstract

Estimating and maximizing mutual information (MI) is core to many objectives in machine learning, but tractably lower bounding MI in high dimensions is challenging. Recent work has introduced variational lower bounds with neural networks to attack this problem, but the tradeoffs and relationships between these techniques remains unclear. Here, we present several results that begin to demystify these techniques: we show that the bias-corrected gradient in MINE (Belghazi et al., 2018) can be derived as an unbiased gradient of a new lower bound on MI, present a stabler Jensen-Shannon-based training algorithm for the critic, provide a new interpretation of contrastive predictive coding (CPC, van den Oord et al. (2018)) and prove this variant is a lower bound on MI, and demonstrate the batch-size dependence of CPC. Empirically, we show that the effectiveness of these bounds depends on properties of the data being modeled and the structure of the critic, with no one bound uniformly dominating.

Introduction. Estimating the relationship between pairs of variables is a fundamental problem in science and engineering. Quantifying the degree of the relationship requires a metric that captures a notion of dependency. Here, we focus on mutual information (MI), denoted $I(x; y)$, which is a reparameterization-invariant measure of dependency:

$$I(x, y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x, y)}{p(x)p(y)}\right] = \mathbb{E}_{p(x,y)}\left[\log \frac{p(y|x)}{p(y)}\right].$$

Estimating MI is challenging when we have samples but not direct access to the underlying distribution (Paninski, 2003; Anonymous, 2019). A related problem appears in representation learning, where we are given samples from one distribution, $x \sim p(x)$ and are interested in learning a stochastic mapping $p_\theta(y|x)$ that captures as much information as possible about x subject to constraints on the mapping (Bell & Sejnowski, 1995; Krause et al., 2010; Hu et al., 2017; Hjelm et al., 2018; Alemi et al., 2017; McAllester, 2018). Learning $p_\theta(y|x)$ requires computing gradients of an unbiased estimate or lower bound on MI so that we can increase $I(x; y)$, but may not require directly estimating MI. Unfortunately, even given a tractable $p_\theta(y|x)$, directly estimating MI is intractable due to the $p(y)$ term. While many parametric and non-parametric (Nemenman et al., 2004; Kraskov et al., 2004) techniques have been proposed to address the MI estimation and maximization problems, few of them scale up to the dataset size and dimensionality encountered in modern machine learning problems.

To overcome these scaling difficulties, recent work combines variational bounds (Donsker & Varadhan, 1983; Barber & Agakov, 2003; Nguyen et al., 2010) with deep learning (Alemi et al., 2016, 2017; van den Oord et al., 2018; Hjelm et al., 2018; Belghazi et al., 2018). These papers present differentiable and tractable mechanisms for estimating mutual information by introducing a neural-network critic that estimates a conditional distribution $p(y|x), p(x|y)$ or a density ratio $\frac{p(x,y)}{p(x)p(y)}$.

While these bounds have been useful for MI estimation and representation learning, their properties and tradeoffs are not well understood. Here we present several new results that begin to demystify these bounds and their relationships. In Table 1, we summarize several estimators and their properties.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

Lower Bound		L	∇L	$\perp$ BS	Var.	Norm.
I_{BA}	Barber & Agakov (2003)	✗	✓	✓	✓	✗
I_{DV}	Donsker & Varadhan (1983)	✗	✗	–	–	–
I_{NWJ}	Nguyen et al. (2010)	✓	✓	✓	✗	✓
I_{MINE}	Belghazi et al. (2018)	✗	✓	✓	✗	✓
I_{CPC}	van den Oord et al. (2018)	✓	✓	✗	✓	✗
I_{JS}	Ours	✓	✓	✓	✗	✓
I_{EB}	Ours	✓	✓	✓	✗	✓
I_{TCPC}	Ours	✓	✓	✗	✓	✗

Table 1: Characterization of mutual information lower bounds. Estimators can have a tractable (✓) or intractable (✗) objective (L), tractable (✓) or intractable (✗) gradients (∇L), be dependent (✗) or independent (✓) of batch size ($\perp$ BS), have high (✗) or low (✓) variance (Var.), and requires a normalized (✗) vs unnormalized (✓) critic (Norm.). See Table 2 for the objectives and parameters.

Data processing inequality. We can combine any estimator of MI with an initial preprocessing step to reduce the dimensionality of the data. This allows us to turn a complex high-dimensional estimation problem into a more tractable lower-dimensional problem. By data processing inequality, this is still a lower bound on MI:

$$I(x; y) \geq I(r(x); s(y)). \tag{1}$$

We can jointly maximize the processing functions r and s alongside paramters of our model or MI estimator to form a tighter bound on $I(x; y)$.

New bound explains bias-corrected gradient in MINE. In Appendix B, we show how to derive several existing and new bounds in a unified framework. In particular, we derive a new lower bound on MI that depends on a learned critic $f(x, y)$ and baseline $a(y) > 0$:

$$I(x, y) \geq I_{\text{EB}}(x, y; f, a) \triangleq \mathbb{E}_{p(x,y)}\left[\log f(x, y)\right] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}\left[f(x, y)\right]}{a(y)} + \log(a(y)) - 1\right], \tag{2}$$

Unlike I_{DV} (the Donsker & Varadhan (1983) lower bound), we can compute unbiased gradient estimates of this bound with respect to $a(y), f,$ and encoder parameters $p(y|x)$ jointly. This bound holds for any choice of $a > 0$. Choosing a to be the scalar moving average estimate of $\mathbb{E}_{p(x)p(y)}\left[\log f(x, y)\right]$ gives an objective whose gradients exactly match the bias-corrected gradients used in I_{MINE} (Belghazi et al., 2018). However, the optimal a depends on y, thus simply taking the moving average estimate may be insufficient. Alternatively, choosing a to be the fixed constant e, yields exactly I_{NWJ} from Nguyen et al. (2010) (also known as MINE-f from Belghazi et al. (2018)).

Reinterpreting CPC. While CPC focuses on the use of an unnormalized critic, we can also leverage learned or known normalized conditional densities $e(y|x)$ to build contrastive lower bounds on MI:

$$I \geq I_{\text{TCPC}} \triangleq \mathbb{E}_{p^K(x)}\left[\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{y_i \sim p(y_i|x_i)}\left[\log \frac{e(y_i|x_i)}{\frac{1}{K}\sum_{j=1}^{K} e(y_i|x_j)}\right]\right], \tag{3}$$

where $p^K(x)$ is the distribution corresponding to drawing a minibatch of K independent samples from the data density, $p(y_i|x_i)$ is the conditional distribution (that only needs to be sampled, not evaluated), and $e(y_i|x_j)$ is a variational distribution that approximates $p(y_i|x_j)$ (see Appendix D for derivation). This objective is a special case of I_{CPC}, where $f(x_i, y_j) = e(y_j|x_i)$ is a normalized density. Maximizing I_{TCPC} in terms of e gives $e^*(y|x) = p(y|x)$. In the case that $p(y|x)$ is known (for example in representation learning where $p_\theta(y|x)$ is a stochastic mapping from data to latents), we can simply use that for $e(y|x)$ to recover an estimate of MI without an additional critic! Thus we can think of the critic in CPC in two ways: (1) learning an *encoder* with an optimum of $p(y|x)$, or (2) learning a *density ratio* with an optimum of $\frac{p(x|y)}{p(x)}$ (as presented in van den Oord et al. (2018)).

A bias-variance tradeoff in MI estimation. In Fig. 1, we evaluate several MI estimators on the toy correlated Gaussian task from Belghazi et al. (2018) where (x, y) are drawn from a 20-d Gaussian distribution with correlation ρ (see Appendix E for details). We find that all the non-contrastive unnormalized critic estimates of MI exhibit high variance, and are challenging to tune for even this

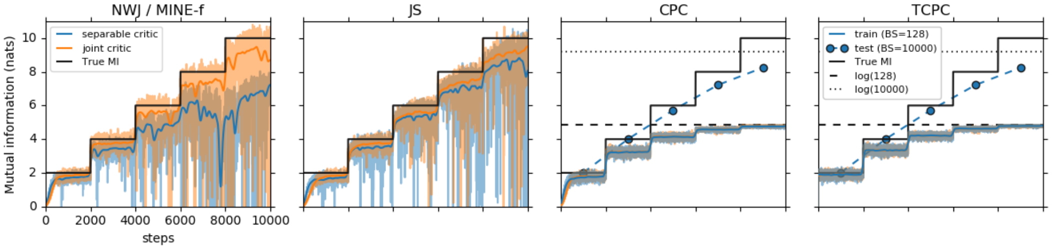

Figure 1: Performance of bounds at estimating mutual information. The dataset $p(x, y; \rho)$ is a correlated Gaussian with the correlation ρ stepping over time. Critics are trained to maximize each lower bound on MI, and the objective (light) and smoothed objective (dark) are plotted for each technique and critic type. The non-contrastive bounds have higher variance than CPC, and I_{JS} outperforms I_{NWJ} at higher MIs. I_{CPC} and I_{TCPC} are poor estimators of MI with the small training batch size of 128, but when evaluating with a larger batch size (10000) the MI estimators are on par or better than the non-contrastive approaches. Using a joint critic (orange) outperforms a separable critic (blue) for I_{JS} and I_{NWJ}, but has no effect on I_{CPC} and I_{TCPC}. Note that TCPC has no learnable parameters and uses the true $p(y|x; \rho)$ for the critic.

simple toy task. However, contrastive approaches based on CPC are low variance, but have estimates that saturate at $\log(\text{batch size})$. None of the estimators we have tried exhibit good estimates of MI at high rates *and* low variance, supporting the theoretical findings of Anonymous (2019).

Stabilizing critics with JS training. Updating the critic f using gradients of I_{NWJ} can be unstable due to the exp of the critic logits in the I_{NWJ} objective. Instead, one can update the critic f to maximize a lower bound on the Jensen-Shannon (JS) divergence, and use the density ratio estimate from the critic to plug into the lower bound on the KL divergence given by I_{NWJ}. We call this approach I_{JS} as we update the critic using the JS, but still evaluate using MI. This approach is similar to Poole et al. (2016); Mescheder et al. (2017) where the JS critic is used in a Monte-Carlo approximation of the f-divergence, but here we plug into I_{NWJ} to ensure we compute a lower bound (see Appendix C for details).

Efficiency-accuracy tradeoffs for critic architectures. One major difference between CPC and MINE is the structure of the critic architecture. CPC uses a separable critic $f(x, y) = h(x)^T g(y)$ which allows one to do $2N$ forward passes through a neural network for a batch size of N. However, Belghazi et al. (2018) use a joint critic, where x, y are concatenated and fed as input to one network, thus requiring N^2 forward passes. For the toy problem, we found that separable critics (blue) increased the variance of the estimator and downwardly biased the estimate of MI compared to joint critics (blue) when using I_{NWJ} or I_{JS}. However, using joint critics on large datasets is intractable. Future work should look into designing better critic architectures that share the efficiency of separable critics approach without sacrificing the increased expressivity of the joint critics.

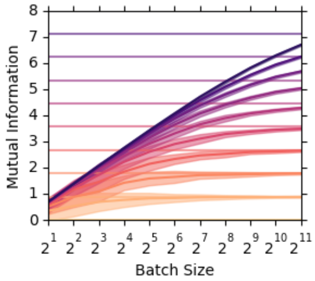

Figure 2: Scaling of I_{CPC} with batch size given the optimal critic $f(x, y; \rho) = p(y|x; \rho)$.

Batch size dependence of CPC. As noted in van den Oord et al. (2018), the mutual information estimate I_{CPC} is upper bounded by $\log(\text{batch size})$. However, we can learn the critic with a small batch size and evaluate the critic with a larger batch size. Given that the optimal critic for the CPC loss does not depend on the batch size, we can expect larger batch sizes to improve our MI estimate (van den Oord et al., 2018). In Fig. 1, the estimate of MI improves with batch size. We study this behavior further in Fig. 2, by fixing a dataset of a particular mutual information (adjusting the correlation of the multivariate Gaussian from the toy problem), and computing I_{CPC} for different batch sizes using the *optimal* critic. While we need an enormous batch size to approximate MI (exponential in the true MI), the accuracy of the learned critic at small batch sizes points to gradients of this bound still being useful in learning representations that maximize information. Future work should address how accurate the gradients of I_{CPC} with small batch sizes are for learning.

Open problems in MI estimation. None of the approaches we considered here are capable of providing good estimates of MI with practical batch sizes. Future work should identify whether such estimators are impossible (Anonymous, 2019), or whether certain distributional assumptions or neural network inductive biases can be leveraged to build tractable estimators. Another question is whether estimating gradients of MI is easier than estimating MI itself? Maximizing I_{BA} is feasible even though we do not have access to the constant data entropy. If we do not care about MI estimation and only care about MI maximization are there better approaches?

Another open question is whether mutual information is a useful objective for representation learning. Recent work from Hjelm et al. (2018) propose to optimize for the JS from the joint to factorial distribution instead of the KL. While deviating from mutual information maximization loses a number of connections to information theory, it may provide other mechanisms for learning features that are useful for downstream tasks. In future work, we will evaluate these estimators on larger-scale representation learning tasks to address these questions.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken elbo, 2017.

Anonymous. Formal limitations on the measurement of mutual information. In *Submitted to International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BkedwoC5t7`. under review.

D Barber and FV Agakov. The im algorithm: A variational approach to information maximization. In *NIPS*, pp. 201–208. MIT Press, 2003.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Devon Hjelm, and Aaron Courville. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 530–539, 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.

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

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.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. *arXiv preprint arXiv:1702.08720*, 2017.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

Andreas Krause, Pietro Perona, and Ryan G Gomes. Discriminative clustering by regularized information maximization. In *Advances in neural information processing systems*, pp. 775–783, 2010.

David McAllester. Information theoretic co-training. *arXiv preprint arXiv:1802.07572*, 2018.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

Ilya Nemenman, William Bialek, and Rob de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111, 2004.

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

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Ben Poole, Alexander A Alemi, Jascha Sohl-Dickstein, and Anelia Angelova. Improved generator objectives for gans. *arXiv preprint arXiv:1612.02780*, 2016.

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

A Objectives

MI Bound	Parameters	Objective
I_{BA}	$q(x\|y)$, tractable decoder	$\mathbb{E}_{p(x,y)}\left[\log q(x\|y) - \log p(x)\right]$
I_{DV}	$f(x,y)$, critic	$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \log\left(\mathbb{E}_{p(x)p(y)}\left[f(x,y)\right]\right)$
I_{NWJ}	$f(x,y)$	$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \frac{1}{e}\mathbb{E}_{p(x)p(y)}\left[f(x,y)\right]$
I_{MINE}	$f(x,y)$, EMA$(\log f)$	I_{DV} for evaluation, $I_{\text{EB}}(f, \text{EMA}(\log f))$ for gradient
I_{CPC}	$f(x,y)$	$\mathbb{E}_{p^K(x,y)}\left[\frac{1}{K}\sum_{i=1}^K \log \frac{f(y_i,x_i)}{\frac{1}{K}\sum_{j=1}^K f(y_i,x_j)}\right]$
I_{JS}	$f(x,y)$	I_{NWJ} for evaluation, f-GAN JS for gradient
I_{EB}	$f(x,y)$, $a(y)>0$	$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}[f(x,y)]}{a(y)} + \log(a(y)) - 1\right]$
I_{TCPC}	$e(y\|x)$, tractable encocder	I_{CPC} with $f(x,y) = e(y\|x)$

Table 2: Parameters and objectives for mutual information estimators.

B Deriving the energy-based bounds

Here we present a new derivation of the lower bounds in Barber & Agakov (2003); Donsker & Varadhan (1983); Nguyen et al. (2010); Belghazi et al. (2018) and describe the new objective whose gradient is the bias-corrected radient used in MINE Belghazi et al. (2018).

A commonly used bound due to Barber & Agakov (2003) follows from the inequality

$$0 \leq \mathbb{E}_{p(y)}\left[KL(p(x|y)||q(x|y))\right] = \mathbb{E}_{p(x,y)}\left[\log p(x|y) - \log q(x|y)\right],$$

which holds with equality when $q(x|y) = p(x|y)$. With this, we arrive at the bound by Barber & Agakov (2003):

$$I(x,y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x|y)}{p(x)}\right] \geq \mathbb{E}_{p(x,y)}\left[\log q(x|y) - \log p(x)\right]. \tag{4}$$

Unfortunately, evaluating this objective is intractable as it requires being able to compute the data entropy via the density $p(x)$. However, we can compute gradients of this objective with respect to the encoder $p(y|x)$ and variational decoder $q(x|y)$ and thus we can use this objective when we are learning an encoder $p(y|x)$ to maximize mutual information as in Alemi et al. (2017).

To derive bounds that can be tractably evaluated, we will turn towards an energy-based view of the variational lower bound. Suppose we can specify $q(x|y)$ up to a normalizing constant, choosing a parameterization of the form:

$$q(x|y) = \frac{p(x)}{Z(y)} f(x,y)$$

for which

$$Z(y) = \int dx\, p(x) f(x, y) = \mathbb{E}_{p(x)}\left[f(x, y)\right]$$

Plugging this into equation 4 gives:

$$I(X;Y) \geq \mathbb{E}_{p(x,y)}\left[\log \frac{q(x|y)}{p(x)}\right] = \mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \mathbb{E}_{p(y)}\left[\log \mathbb{E}_{p(x)}\left[f(x,y)\right]\right] \triangleq I_{\text{EB1}} \quad (5)$$

The challenge with this bound is that unlike Eq. 4, computing unbiased gradient estimates is non-trivial. Naturally, we would like to move the $\log$ term inside the expectation, however, this produces a bound in the wrong direction. Fortunately, we can use a well-known upper bound on $\log$ expressions. By the concavity of $\log$, we have $\log(x) \leq \frac{x}{a} + \log(a) - 1$ for all $x, a > 0$. As a result, we have that

$$\log \mathbb{E}_{p(x)}\left[f(x,y)\right] \leq \frac{\mathbb{E}_{p(x)}\left[f(x,y)\right]}{a(y)} + \log(a(y)) - 1$$

where the variational parameter $a(y)$ can depend on y and the bound is tight when $a(y) = \mathbb{E}_{p(x)}\left[f(x,y)\right]$. To tighten, the bound, we would minimize the bound with respect to $a(y)$. Putting this together with Eq. 5, we get

$$\begin{aligned} I(x,y) &\geq \mathbb{E}_{p(x,y)}\left[\log f(x|y)\right] - \mathbb{E}_{p(y)}\left[\log \mathbb{E}_{p(x)}\left[f(x,y)\right]\right] \\ &\geq \mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}\left[f(x,y)\right]}{a(y)} + \log(a(y)) - 1\right], \end{aligned} \quad (6)$$

which we can compute unbiased gradient estimates of, which would allow us to maximize the bound with respect to $a(y), f$, and model parameters jointly (conveniently, all of the optimization is in the same direction). To implement this, we can learn a parameterized function $a(y)$ to jointly maximize the lower bound.

Note that Eq. 5 is an upper bound on the MINE bound (Belghazi et al., 2018)

$$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \mathbb{E}_{p(y)}\left[\log \mathbb{E}_{p(x)}\left[f(x,y)\right]\right] \geq \mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \log \mathbb{E}_{p(x)p(y)}\left[f(x,y)\right],$$

where $\log f(x,y) = T_\theta$ in the notation of Belghazi et al. (2018).

Applying a similar log inequality yields a lower bound on MINE

$$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \log \mathbb{E}_{p(x)p(y)}\left[f(x,y)\right] \geq \mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}\left[f(x,y)\right]}{a} + \log(a) - 1\right].$$

This bound holds for any choice of $a > 0$, so choosing a to be a moving average estimate of $\mathbb{E}_{p(x)p(y)}\left[\log f(x,y)\right]$ is valid and results in precisely the same update as proposed in (Belghazi et al., 2018) as the bias corrected optimization procedure for MINE. Alternatively, choosing $a = e$, yields

$$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \log \mathbb{E}_{p(x)p(y)}\left[f(x,y)\right] \geq \mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}\left[f(x,y)\right]}{e}\right],$$

which is the I_{NWJ}/MINE-f bound.

C Deriving I_{JS}

Given that optimizing can be unstable, we can instead optimize the critic using the lower bound on Jensen-Shannon divergence as in GANs, and use the density ratio estimate from the JS critic to construct a critic for the KL lower bound.

The optimal critic for KL is given by (Nowozin et al., 2016):

$$T^*(x) = 1 + \log \frac{p(x)}{q(x)}.$$

If we use the f-GAN formulation for the JS divergence, then we can read out the density ratio from the real-valued logits $V(x)$:

$$\frac{p(x)}{q(x)} \approx \exp\left(V(x)\right)$$

In Poole et al. (2016); Mescheder et al. (2017), they plug in this estimate of the density ratio into a Monte-Carlo approximation of the f-divergence. However, this is no longer a bound on the f-divergence, it is just an approximation. Instead, we can construct a critic for the KL divergence and use that to get a lower bound:

$$T_{KL}(x) = 1 + V(x)$$

Plugging into I_{NWJ}, (aka the KL lower bound from f-GAN we get):

$$KL(p\|q) \geq \mathbb{E}_{x\sim p}\left[T_{KL}(x)\right] - \mathbb{E}_{x\sim q}\left[\exp(T_{KL}(x) - 1)\right] = 1 + \mathbb{E}_{x\sim p}\left[V(x)\right] - \mathbb{E}_{x\sim q}\left[\exp(V(x))\right]$$

Note that if the log density ratio estimate $V(x)$ is exact, i.e. $V(x) = \frac{p(x)}{q(x)}$, then the last term, $\mathbb{E}_{x\sim q}\left[\exp(V(x))\right]$ will be one, and the first term is exactly $KL(p\|q)$.

In our case, p is $p(x, y)$ and q is $p(x)p(y)$, so this objective becomes

$$I(X;Y) \geq \mathbb{E}_{x,y\sim p(x,y)}\left[V(x,y)\right] - \mathbb{E}_{x,y\sim p(x)p(y)}\left[\exp(V(x,y))\right].$$

D I_{TCPC} derivation

Here we derive a lower bound on mutual information by splitting the the objective into minibaches and lower bounding the objective on each minibatch. We believe these results represent a new variational lower bound on MI, and unlike van den Oord et al. (2018) does not rely on any approximations.

D.1 Known encoder

First, we consider the representation learning case where the mapping from x to y, $p(y|x)$ is known and can be evaluated. Given minibatches of size K, we can write the mutual information as:

$$I(x;y) = \mathbb{E}_{x_{1:K}\sim\mathcal{D}}\left[\frac{1}{K}\sum_{i=1}^{K}\text{KL}\left(p(y|x_i)\|p(y)\right)\right] \tag{7}$$

To establish a lower bound on the mutual information, we will lower bound each term in the expectation by replacing the intractable marginal $p(y)$ with an approximation based on samples from the minibatch:

$$m(y;x_{1:K}) = \frac{1}{K}\sum_{i=1}^{K}p(y|x_i) \tag{8}$$

To show this yields a lower bound, we rewrite Eq. 7 by multiplying by $m(y;x_{1:K})/m(y;x_{1:K})$ and splitting into two terms:

$$\frac{1}{K}\sum_{i=1}^{K}\text{KL}\left(p(y|x_i)\|p(y)\right) = \frac{1}{K}\sum_{i=1}^{K}\text{KL}\left(p(y|x_i)\|m(y;x_{1:K})\right) + \text{KL}(m(y;x_{1:K})\|p(y)) \tag{9}$$

As the second term, $\text{KL}(m(y;x_{1:K})\|p(y))$ is non-negative, dropping this term yields a tractable lower bound on mutual information:

$$I(x;y) \geq I_{\text{TCPC}} \triangleq \mathbb{E}_{x_{1:K}\sim\mathcal{D}}\left[\frac{1}{K}\sum_{i=1}^{K}\text{KL}\left(p(y|x_i)\|\frac{1}{K}\sum_{i=1}^{K}p(y|x_i)\right)\right] \tag{10}$$

D.2 Unknown encoder

If the encoder $p(y|x)$ is not known or is intractable, we can introduce a variational encoder $e(y|x)$ to approximate it. This occurs when we are given samples (x, y) and are interested in MI estimation, or when we are interested in learning a mapping from x to y that does not have a tractable density $p(y|x)$.

We aim to show that if we use an $e(y|x) \neq p(y|x)$ that the objective is strictly worse than Eqn. 10, i.e.

$$\frac{1}{K}\sum_{i=1}^{K}\int dy\, p(y|x_i)\log\frac{p(y|x_i)}{\frac{1}{K}\sum_{i=1}^{K}p(y|x_i)} \geq \frac{1}{K}\sum_{i=1}^{K}\int dy\, p(y|x_i)\log\frac{e(y|x_i)}{\frac{1}{K}\sum_{i=1}^{K}e(y|x_i)} \tag{11}$$

The optimization problem is effectively over K densities, $e_1, ..., e_K$. We can compute the gradient for one e_j (ignoring the constraint that it integrates to 1):

$$\nabla_{e_j}\sum_{i=1}^{K}\int dy\, p(y|x_i)\log\frac{e_i(y|x_i)}{\sum_{i=1}^{K}e_i(y|x_i)} = -\int dy\,\frac{\sum_i p(y|x_i)}{\sum_i e_i(y|x_i)} + \int dy\,\frac{p(y|x_j)}{e_j(y|x_j)} \tag{12}$$

$$\implies \int dy\,\frac{\sum_i p(y|x_i)}{\sum_i e_i^*(y|x_i)} = \int dy\,\frac{p(y|x_j)}{e_j^*(y|x_j)} \tag{13}$$

Thus $e_i(y|x_i) \propto p(y|x_i)$ is probably an extremum of the objective. Future work should more rigorously address whether this unknown encoder case is provably a minimum, and thus whether CPC is truly a lower bound on mutual information.

D.3 Relation to CPC

The minibatch objective is identical to CPC when using a tractable encoder and a Monte-Carlo approximation of the KL! Lets consider sampling $y_i \sim p(y_i|x_i)$ for each element i in the minibatch. Then we have:

$$\mathcal{L}_K = \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{y_i\sim p(y_i|x_i)}\left[\log\frac{e(y_i|x_i)}{\frac{1}{K}\sum_{j=1}^{K}e(y_i|x_j)}\right]$$

We know that drawing a single sample for the inner expectations will give us an unbiased estimate, i.e. if

$$\hat{\mathcal{L}}_K = \frac{1}{K}\sum_{i=1}^{K}\log\frac{e(y_i|x_i)}{\frac{1}{K}\sum_{j=1}^{K}e(y_i|x_j)} \tag{14}$$

then $\mathbb{E}_{y_{1:K}}\left[\hat{\mathcal{L}}_K\right] = \mathcal{L}_K$.

Furthermore Eq. 14 is exactly the equation for CPC up to the constant $\log K$ with $f(x_j, y_i) = e(y_i|x_j)$:

$$\hat{\mathcal{L}}_K - \log K = \frac{1}{K}\sum_{i=1}^{K}\log\frac{f(x_i, y_i)}{\sum_{j=1}^{K}f(x_j, y_i)} \tag{15}$$

Here we know that the optimal critic is $f^*(x_j, y_i) = p(y_i|x_j)$, which only depends on the conditional, and does not depend on the marginal $p(y)$.

E Experimental details

Dataset. For each dimension, we sampled (x_i, y_i) from a correlated Gaussian with mean 0 and correlation of ρ. We used a dimensionality of 20, i.e. $x \in \mathbb{R}^{20}, y \in \mathbb{R}^{20}$. Given the correlation coefficient ρ, and dimensionality $d = 20$, we can compute the true mutual information: $I(x, y) = -\frac{d}{2}\log(1-\rho^2)$. For Fig. 1, we increase ρ over time to show how the estimator behavior depends on the true mutual informaiton.

Architectures. We experimented with two forms of architecture: separable and joint. Separable architectures independently mapped x and y to an embedding space and then took the inner product, i.e. $f(x, y) = h(x)^T g(y)$ as in (van den Oord et al., 2018). Joint critics concatenate each x, y pair before feeding it into the network, i.e. $f(x, y) = h([x, y])$ as in (Belghazi et al., 2018). In practice, separable critics are much more efficient as we only have to perform $2N$ forward passes through neural networks for a batch size of N vs. N^2 for joint critics. All networks were fully-connected networks with ReLU activations.

Optimization. All lower bounds were optimized with a batch size of 64 and Adam. We experimented with other optimizers and gradient clipping but found that it hurt results on the toy problem.

$$I(X;Y) \geq \log K - \mathcal{L}_K$$
$$\mathcal{L}_K \geq 0$$