A. Summary of mutual information lower bounds

In Table 1, we summarize the characteristics of lower bounds on MI. The parameters and objectives used for each of these bounds is presented in Table 2.

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_{NCE}	van den Oord et al. (2018)	✓	✓	✗	✓	✓
I_{JS}	Appendix D	✓	✓	✓	✗	✓
I_α	Eq. 11	✓	✓	✗	✓	✓

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.).

Lower 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}	$f(x,y)$, $\text{EMA}(\log f)$	I_{DV} for evaluation, $I_{\text{TUBA}}(f, \text{EMA}(\log f))$ for gradient
I_{NCE}	$f(x,y)$	$\mathbb{E}_{p^K(x,y)}\left[\frac{1}{K}\sum_{i=1}^{K}\log\frac{f(y_i,x_i)}{\frac{1}{K}\sum_{j=1}^{K}f(y_i,x_j)}\right]$
I_{JS}	$f(x,y)$	I_{NWJ} for evaluation, f-GAN JS for gradient
I_{TUBA}	$f(x,y)$, $a(y) > 0$	$\mathbb{E}_{p(x,y)}\left[\log f(x,y)\right] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}[f(x,y)]}{a(y)} + \log(a(y)) - 1\right]$
I_{TNCE}	$e(y\|x)$ tractable encocder	I_{NCE} with $f(x,y) = e(y\|x)$
I_α	$f(x,y)$, α, $q(y)$	$1 + \mathbb{E}_{p(x_{1:K},y)}\left[\log\frac{e^{f(x_1,y)}}{\alpha m(y;x_{1:K})+(1-\alpha)q(y)}\right]$ $-\mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{e^{f(x_1,y)}}{\alpha m(y;x_{1:K})+(1-\alpha)q(y)}\right]$

Table 2. Parameters and objectives for mutual information estimators.

B. 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. 2, we increase ρ over time to show how the estimator behavior depends on the true mutual information.

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.

	Mutual Information				
	2.0	4.0	6.0	8.0	10.0
Gaussian, unstructured					
I_α	1.9	3.8	5.7	7.4	8.9
I_{NCE}	1.9	3.6	4.9	5.7	6.0
I_{JS}	1.2	3.0	4.8	6.5	8.1
I_{NWJ}	1.6	3.5	5.2	6.7	8.0
Cubic, unstructured					
I_α	1.7	3.6	5.4	6.9	8.2
I_{NCE}	1.7	3.2	4.1	4.6	4.8
I_{JS}	1.0	2.8	4.5	6.1	7.6
I_{NWJ}	1.5	3.2	4.7	5.9	6.9
Gaussian, known $p(y\|x)$					
I_{NCE} (Eq. 12)	1.9	3.3	4.2	4.6	4.8
I_{NWJ} (Eq. 14)	2.0	4.0	6.0	8.0	10.0

Table 3. Hyperparameter-optimizes results on the toy Gaussian and Cubic problem of Fig. 2.

C. Additional experiments

C.1. Exhaustive hyperparameter sweep.

To better evaluate the tradeoffs between different bounds, we performed more extensive experiments on the toy problems in Fig. 2. For each bound, we optimized over learning rate, architecture (separable vs. joint critic, number of hidden layers (1-3), hidden units per layer (256, 512, 1024, 2048), nonlinearity (ReLU or Tanh), and batch size (64, 128, 256, 512). In Table 3, we present the estimate of the best-performing hyperparameters for each technique. For both the Gaussian and Cubic problem, I_α outperforms all approaches at all levels of mutual information between X and Y. While the absolute estimates are improved after this hyperparameter sweep, the ordering of the approaches is qualitatively the same as in Fig. 2. We also experimented with the bounds that leverage known conditional distribution, and found that Eq. 14 that leverages a known $p(y|x)$ is highly accurate as it only has to learn the marginal $q(y)$.

C.2. Effective bias-variance tradeoffs with I_α

To better understand the effectiveness of I_α at trading off bias for variance, we plotted bias vs. variance for 3 levels of mutual information on the toy 20-dimensional Gaussian problem across a range of architecture settings. In Fig. 6, we see that I_α is able to effectively interpolate between the high-bias low-variance I_{NCE}, and the low-bias high-variance I_{NWJ}. We also find that I_{JS} is competitive at high rates, but exhibits higher bias and variance than I_α at lower rates.

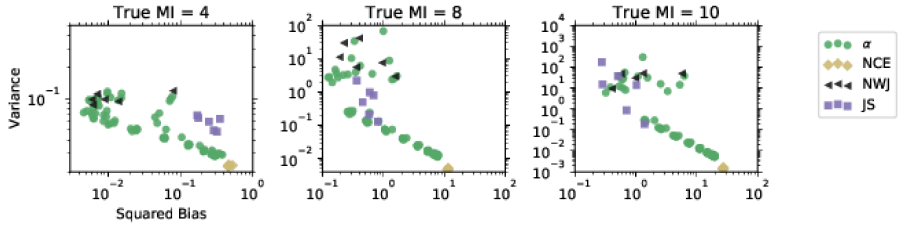

Figure 6. I_α effectively interpolates between I_{NCE} and I_{NWJ}, trading off bias for variance.

In addition to I_α, we compared to two alternative interpolation procedures, neither of which showed the improvements of I_α:

1. I_α interpolation: multisample bound that uses a critic with linear interpolation between the batch mixture $m(y; x_{1:K}$ and the learned marginal $q(y)$ in the denominator (Eqn. 11).

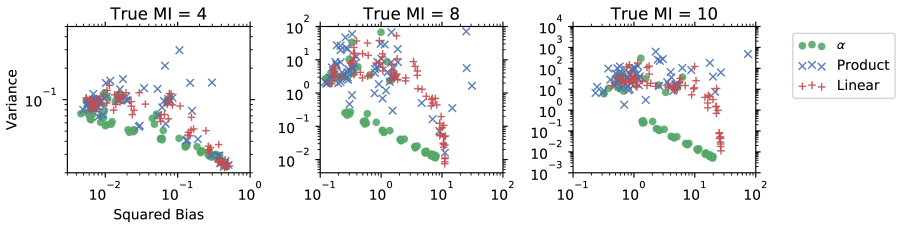

Figure 7. Comparing I_α to other interpolations schemes.

2. Linear interpolation: $\alpha I_{\text{NCE}} + (1-\alpha) I_{\text{NWJ}}$

3. Product interpolation: same as I_α, but uses the product $m(y; x_{1:K})^\alpha q(y)^{(1-\alpha)}$ in the denominator.

We compared these approaches in the same setting as Fig. 6, evaluating the bias and variance for various hyperparameter settings at three different levels of mutual information. In Fig. 7, we can see that neither the product or linear interpolation approaches reduce the bias or variance as well as I_α.

D. I_{JS} derivation

Given the high-variance of I_{NWJ}, optimizing the critic with this objective can be challenging. Instead, we can optimize the critic using the lower bound on Jensen-Shannon (JS) divergence as in GANs and Hjelm et al. (2018), and use the density ratio estimate from the JS critic to construct a critic for the KL lower bound.

The optimal critic for I_{NWJ}/f-GAN KL that saturates the lower bound on $KL(p\|q)$ is given by (Nowozin et al., 2016):

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

If we use the f-GAN formulation for parameterizing the critic with a softplus activation, 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, $T_{KL}(x) = 1 + V(x)$, and use that to get a lower bound using the I_{NWJ} objective:

$$\begin{aligned} KL(p\|q) &\geq \mathbb{E}_{x\sim p}\left[T_{KL}(x)\right] - \mathbb{E}_{x\sim q}\left[\exp(T_{KL}(x) - 1)\right] && (16)\\ &= 1 + \mathbb{E}_{x\sim p}\left[V(x)\right] - \mathbb{E}_{x\sim q}\left[\exp(V(x))\right] && (17)\end{aligned}$$

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

For the special case of mutual information estimation, p is the joint $p(x, y)$ and q is the product of marginals $p(x)p(y)$, yielding:

$$I(X;Y) \geq 1 + \mathbb{E}_{p(x,y)}\left[V(x,y)\right] - \mathbb{E}_{p(x)p(y)}\left[\exp(V(x,y))\right] \triangleq I_{\text{JS}}. \quad (18)$$

E. Alternative derivation of I_{TNCE}

In the main text, we derive I_{NCE} (and I_{TNCE}) from a multi-sample variational lower bound. Here we present a simpler and more direct derivation of I_{TNCE}. Let $p(x)$ be the data distribution, and $p(x_{1:K})$ denote K samples drawn iid from $p(x)$. Let $p(y|x)$ be a stochastic encoder, and $p(y)$ be the intractable marginal $p(y) = \int dx\, p(x)p(y|x)$. First, we can write the mutual

information as a sum over K terms each of whose expectation is the mutual information:

$$I(X;y) = \mathbb{E}_{x_{1:K}}\left[\frac{1}{K}\sum_{i=1}^{K}\mathrm{KL}(p(y|x_i)\|p(y))\right] = \mathbb{E}_{x_{1:K}}\left[\frac{1}{K}\sum_{i=1}^{K}\int dy\, p(y|x_i)\log\frac{p(y|x_i)}{p(y)}\right] \tag{19}$$

$$\tag{20}$$

Let $m(y;x_{1:K}) = \frac{1}{K}\sum_{i=1}^{K} p(y|x_i)$ be the minibatch estimate of the intractable marginal $p(y)$. We multiply and divide by m and then simplify:

$$I(X;y) = \mathbb{E}_{x_{1:K}}\left[\frac{1}{K}\sum_{i=1}^{K}\int dy\, p(y|x_i)\log\frac{p(y|x_i)m(y;x_{1:K})}{m(y;x_{1:K})p(y)}\right] \tag{21}$$

$$= \mathbb{E}_{x_{1:K}}\left[\frac{1}{K}\sum_{i=1}^{K}\left[\int dy\, p(y|x_i)\log\frac{p(y|x_i)}{m(y;x_{1:K})} + \int dy\, p(y|x_i)\log\frac{m(y;x_{1:K})}{p(y)}\right]\right] \tag{22}$$

$$= \mathbb{E}_{x_{1:K}}\left[\frac{1}{K}\sum_{i=1}^{K}\mathrm{KL}(p(y|x_i)\|m(y;x_{1:K})) + \int dy\, \frac{1}{K}\sum_{i=1}^{K} p(y|x_i)\log\frac{m(y;x_{1:K})}{p(y)}\right] \tag{23}$$

$$= \mathbb{E}_{x_{1:K}}\left[\left(\frac{1}{K}\sum_{i=1}^{K}\mathrm{KL}(p(y|x_i)\|m(y;x_{1:K}))\right) + \mathrm{KL}(m(y;x_{1:K})\|p(y))\right] \tag{24}$$

$$\geq \mathbb{E}_{x_{1:K}}\left[\frac{1}{K}\sum_{i=1}^{K}\mathrm{KL}(p(y|x_i)\|m(y;x_{1:K}))\right] \tag{25}$$