DEMI:DISCRIMINATIVE ESTIMATOR OF MUTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating mutual information between continuous random variables is often intractable and extremely challenging for high-dimensional variables. Recent progress has leveraged neural networks to optimize variational lower bounds on mutual information. Although showing promise for this difficult problem, the variational methods have been theoretically and empirically proven to have serious statistical limitations: 1) most of the approaches cannot make accurate estimates when the underlying mutual information is either low or high; 2) the resulting estimators may suffer from high variance. Our approach is based on training a classifier that provides the probability that a data sample pair is drawn from the joint distribution or from the product of its marginal distributions. We use this probabilistic prediction to estimate mutual information. We show theoretically that our method and other variational approaches are equivalent when they achieve their optimum, while our approach does not optimize a variational bound. Empirical results demonstrate high accuracy and a good bias/variance tradeoff using our approach.

Mutual information (MI) measures the information that two random variables share. MI quantifies the statistical dependency — linear and non-linear — between two variables. This property has made MI a crucial measure in machine learning. In particular, recent work in unsupervised representation learning has built on optimizing MI between latent representations and observations (Chen et al., 2016; Zhao et al., 2018; Oord et al., 2018; Hjelm et al., 2018; Tishby & Zaslavsky, 2015; Alemi et al., 2018; Ver Steeg & Galstyan, 2014). Maximization of MI has long been a default method for multi-modality image registration (Maes et al., 1997), especially in medical applications (Wells III et al., 1996), though in most work the dimensionality of the random variables is very low. Here coordinate transformations on images are varied to maximize their MI.

Estimating MI from finite data samples has been challenging and is intractable for most continuous probabilistic distributions. Traditional MI estimators (Suzuki et al., 2008; Darbellay & Vajda, 1999; Kraskov et al., 2004; Gao et al., 2015) do not scale well to modern machine learning problems with high-dimensional data. This impediment has motivated the construction of variational bounds for MI (Nguyen et al., 2010; Barber & Agakov, 2003); in recent years this has lead to maximization procedures that use deep learning architectures to parameterize the space of functions, exploiting the expressive power of neural networks (Song & Ermon, 2019; Belghazi et al., 2018; Oord et al., 2018).

As pointed out in McAllester & Stratos (2020), optimizing lower bounds on MI has serious statistical limitations. Specifically, any high-confidence distribution-free lower bound cannot be larger than $O(\ln N)$, where N is the number of samples. This implies that if the underlying MI is high, it cannot be accurately and reliably estimated by variational methods like MINE (Belghazi et al., 2018).

Song & Ermon (2019) further categorized the state-of-the-art variational methods into "generative" and "discriminative" approaches, depending on if they estimate the probability densities or the density ratios. Song & Ermon (2019) has shown that the "generative" approaches perform poorly when the underlying MI is small and "discriminative" approaches perform poorly when MI is large. They point out that certain approaches like MINE (Belghazi et al., 2018) are prone to high variances.

We propose a simple discriminative approach based on an auxillary variable that avoids the limitations of previous discriminative methods. Instead of direct density estimation or attempting to predict one data variable from another, our method need only estimate the likelihood that sample

is drawn from the joint distribution versus the product of marginal distributions. We show that accurate performance on this classification task provides an estimate of the log density ratio. This can greatly simplify the MI estimation task in comparison with generative approaches: estimating a single likelihood ratio may be easier than estimating three distributions (the joint and the two marginals). Moreover, classification tasks are generally amicable to deep learning, while density estimation remains challenging in many cases. Song & Ermon (2019) point out that the estimation of the partitition function induces large variance in most discriminative methods; our approach avoids this entirely. Our empirical results bear out these conceptual advantages.

In Section 1, we derive our approach to estimating MI. Section 2 discusses related approaches, including MINE; this is followed by an empirical evaluation in Section 3 and Section 4. Our experimental results demonstrate the advantages of our discriminative-classification-based MI estimation, which has higher accuracy than the state-of-the-art variational approaches and a good bias/variance tradeoff.

1 METHODS

Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be two random variables generated by joint distribution $p : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$. Mutual Information (MI)

$$I(x;y) \triangleq \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \tag{1}$$

is a measure of dependence between x and y. Let $\mathcal{D} = \{(x_i, y_i)_{i=1}^n\}$ be a set of n independent identically distributed (i.i.d.) samples from $p(x,y)$. The law of large numbers implies

$$\hat{I}_p(\mathcal{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(x_i,y_i)}{p(x_i)p(y_i)} \to I(x;y) \quad \text{as } n \to \infty, \tag{2}$$

which suggests a simple estimation strategy via sampling. Unfortunately, the joint distribution $p(x,y)$ is often unknown and therefore the estimate in (2) cannot be explicitly computed. Here we develop an approach to accurately approximating the estimate $\hat{I}_p(\mathcal{D})$ based on discriminative learning.

In our development, we will find it convenient to define a Bernoulli random variable $z \in \{0,1\}$ and to "lift" the distribution $p(x,y)$ to the product space $\mathcal{X} \times \mathcal{Y} \times \{0,1\}$. We thus define a family of distributions parametrized by $\alpha \in (0,1)$ as follows:

$$p_*(x,y|z=1;\alpha) = p(x,y), \tag{3}$$
$$p_*(x,y|z=0;\alpha) = p(x)p(y), \tag{4}$$
$$p_*(z=1;\alpha) = 1 - p_*(z=0;\alpha) = \alpha. \tag{5}$$

Using Bayes' rule, we obtain

$$\frac{p_*(z=1|x,y)}{p_*(z=0|x,y)} = \frac{p_*(x,y,z=1)}{p_*(x,y,z=0)} = \frac{p_*(x,y|z=1)\,p_*(z=1)}{p_*(x,y|z=0)\,p_*(z=0)} = \frac{p(x,y)}{p(x)p(y)} \cdot \frac{\alpha}{1-\alpha}, \tag{6}$$

which implies that the estimate in (2) can be alternatively expressed as

$$\hat{I}_p = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_*(z=1|x_i,y_i)}{p_*(z=0|x_i,y_i)} - \log \frac{\alpha}{1-\alpha} \tag{7}$$
$$= \frac{1}{n}\sum_{i=1}^{n} \text{logit}\left[p_*(z=1|x_i,y_i)\right] - \text{logit}[\alpha], \tag{8}$$

where $\text{logit}[u] \triangleq \log \frac{u}{1-u}$ is the log-odds function.

Our key idea is to approximate the posterior distribution $p_*(z=1|x,y)$ by a classifier that is trained to distinguish between the joint distribution $p(x,y)$ and the product distribution $p(x)p(y)$ as described below.

1.1 TRAINING SET CONSTRUCTION

We assume that we have access to a large collection $\hat{\mathcal{D}}$ of i.i.d. samples (x, y) from $p(x, y)$ and define $\hat{p}(x, y; \hat{\mathcal{D}})$, $\hat{p}(x; \hat{\mathcal{D}})$, and $\hat{p}(y; \hat{\mathcal{D}})$ to be the empirical joint and marginal distributions respectively induced by data set $\hat{\mathcal{D}}$.

We construct the training set $\mathcal{T} = \{(x^j, y^j, z^j)\}$ of m i.i.d. samples from our empirical approximation to the distribution $p_*(x, y, z)$. Each sample is generated independently of all others as follows. First, a value $z^j \in \{0, 1\}$ is sampled from the prior distribution $p_*(z)$ in (5). If $z^j = 1$, then a pair (x^j, y^j) is sampled randomly from the empirical joint distribution $\hat{p}(x, y; \hat{\mathcal{D}})$; otherwise value x^j is sampled randomly from the empirical marginal distribution $\hat{p}(x; \hat{\mathcal{D}})$ and value y^j is sampled randomly from the empirical marginal distribution $\hat{p}(x, y; \hat{\mathcal{D}})$, independently from x^j. This sampling is easy to implement as it simply samples an element from a set of unique values in the original collection $\hat{\mathcal{D}}$ with frequencies adjusted to account for repeated appearances of the same value.

It is straightforward to verify that any individual sample in the training set $\mathcal{T}$ is generated from distribution $p_*(x, y, z)$ up to the sampling of $\hat{\mathcal{D}}$. Where $\hat{\mathcal{D}}$ is small, multiple samples may not be jointly from $\hat{\mathcal{D}}$ but from some idiosyncratic subset; however, the empirical distribution induced by the set $\mathcal{T}$ converges to $p_*(x, y, z)$ as the size of available data $\hat{\mathcal{D}}$ and the size m of the training set $\mathcal{T}$ becomes large.

1.2 CLASSIFIER TRAINING FOR MUTUAL INFORMATION ESTIMATION

Let $q(z = 1|x, y; \theta, \mathcal{T})$ be a (binary) classifier parameterized by θ and derived from the training set $\mathcal{T}$. If $q(z = 1|x, y; \theta, \mathcal{T})$ accurately approximates the posterior distribution $p_*(z = 1|x, y; \alpha)$, then we can use this classifier q instead of $p_*(z = 1|x, y; \alpha)$ in (8) to estimate MI.

We follow the widely used maximum likelihood approach to estimating the classifier's parameters θ and form the cross-entropy loss function

$$\ell(\theta; \mathcal{T}) = -\frac{1}{m} \sum_{j=1}^{m} \log q(z^j|x^j, y^j; \theta, \mathcal{T}) \tag{9}$$

$$= -\frac{1}{m} \sum_{j=1}^{m} z^j \log q(z^j = 1|x^j, y^j; \theta, \mathcal{T}) + (1 - z^j)(1 - \log q(z^j = 1|x^j, y^j; \theta, \mathcal{T})) \tag{10}$$

to be minimized to determine the optimal value of parameters $\hat{\theta}$. Once the optimization is completed, we form the estimate

$$\hat{I}_q(\mathcal{D}, \mathcal{T}) = \frac{1}{n} \sum_{i=1}^{n} \text{logit} \left[q(z = 1|x_i, y_i; \hat{\theta}, \mathcal{T}) \right] - \text{logit}[\alpha] \tag{11}$$

that approximates the estimate in (8). Note that the estimate is computed using the data set $\mathcal{D}$, which is distinct from the training set $\mathcal{T}$.

1.3 ASYMPTOTIC ANALYSIS

As the size of available data $\hat{\mathcal{D}}$ and the size m of the training set $\mathcal{T}$ increase to infinity, the law of large numbers implies

$$\ell(\theta; \mathcal{T}) \to \mathbb{E}_{p_*(x,y,z)} \left[\log q(z|x, y; \theta, \mathcal{T}) \right], \tag{12}$$

and therefore

$$\hat{\theta} \triangleq \arg\max_{\theta} \ell(\theta; \mathcal{T}) \to \arg\max_{\theta} \mathbb{E}_{p_*(x,y,z)} \left[\log q(z|x, y; \theta, \mathcal{T}) \right]. \tag{13}$$

Thus, when the model capacity of the family $q(z|x, y; \theta)$ is large enough to include the original distribution $p_*(z|x, y)$, Gibb's inequality implies

$$q(z|x, y; \hat{\theta}, \mathcal{T}) \to p_*(z|x, y) \quad \text{and} \quad \hat{I}_q(\mathcal{D}, \mathcal{T}) \to I(x; y) \tag{14}$$

as both the training data and testing data grow.

2 RELATED WORK

2.1 CONNECTIONS TO OTHER MUTUAL INFORMATION ESTIMATORS

MINE and SMILE. Belghazi et al. (2018) introduced the Mutual Information Neural Estimation (MINE) method, wherein they proposed learning a neural network $f(x, y; \theta)$ that maximizes the objective function $J(f) = \mathbb{E}_{p(x,y)}\left[f(x, y; \theta)\right] - \log \mathbb{E}_{p(x)p(y)}\left[e^{f(x,y;\theta)}\right]$, which is the Donsker-Varadhan lower bound for the KL divergence. For analysis purposes, we define $\hat{q}(x, y; \theta) \triangleq \frac{1}{Z} e^{f(x,y;\theta)} p(x)p(y)$, where $Z = \mathbb{E}_{p(x)p(y)}\left[e^{f(x,y;\theta)}\right]$. By substituting into the definition of $J(\cdot)$ and invoking Gibb's inequality, we obtain

$$J(f) = \mathbb{E}_{p(x,y)}\left[\log \hat{q}(x, y; \theta)\right] - \mathbb{E}_{p(x,y)}\left[\log p(x)p(y)\right] \tag{15}$$

$$\leq \mathbb{E}_{p(x,y)}\left[\log p(x, y)\right] - \mathbb{E}_{p(x,y)}\left[\log p(x)p(y)\right] = I(x; y), \tag{16}$$

with equality if and only if $\hat{q}(x, y; \theta) \equiv p(x, y)$, i.e.,

$$f(x, y) = \log \frac{p(x, y)}{p(x)p(y)} + C, \tag{17}$$

where C is a constant that is absorbed into the partition function Z. Thus the objective function is a lower bound on MI and is maximized when the unspecified "statistics network" $f(x, y)$ is the log likelihood ratio of the joint distribution and the product of the marginals.

Song & Ermon (2019) introduced the Smoothed Mutual Information Lower Bound Estimator (SMILE) approach which is a modification of the MINE estimator. To alleviate the high variance of $f(x, y)$ in practice, the tilting factor $e^{f(x,y)}$ is constrained to the interval $[e^{-\tau}, e^{\tau}]$, for a tuned hyper-parameter τ. As $\tau \to \infty$, SMILE estimates converges to those produced by MINE.

The log likelihood ratio of the joint versus the marginals, which the $f(x, y)$ network from both these methods approximates, is the optimal classifier function for the task defined on our training set $\mathcal{T}$ above. Our parameterization of this ratio makes use of a classifier and the logit transformation. While analytically equivalent, the MINE and SMILE optimization procedures must instead search over ratio functions directly, optimizing $f(x, y) \approx p(x, y)/p(x)p(y)$ itself. Our experimental results demonstrate the advantage of using our estimator in (11).

Contrastive Predictive Coding. Oord et al. (2018) proposed a contrastive predictive coding (CPC) method that also maximizes a lower bound

$$J(f) = \mathbb{E}_{p(x,y)}\left[\frac{1}{N}\sum_{i=1}^{N} \log \frac{f(x_i, y_i; \theta)}{\frac{1}{N}\sum_{j=1}^{N} f(x_i, y_j; \theta)}\right] + \log N \leq I(x; y),$$

where $f(x, y; \theta)$ is a neural network and N is the batch size. CPC is not capable of estimating high underlying MI accurately– it is constrained by their batch size N, and this constraint scales logarithmically. In our approach, we do not estimate the likelihood ratio directly, instead we construct an auxiliary variable and "lift" the joint distribution, where we leverage the power of a discriminative neural network classifier. The logit transformation of our classifier response is used to approximate the log likelihood ratio in (1).

2.2 RESAMPLING METHODS FOR CHARACTERIZING JOINT DATA

Our approach, as well as other methods such as MINE, uses the given joint/paired data with derived "unpaired" data that capture the product $p(x)p(y)$ of the marginal distributions. The "unpaired" data can be synthesized via permutations or resampling of the paired data. This construction, which synthesizes "unpaired" data and then defines a metric to encourage "paired" data points to map closer than the "unpaired" data in a latent space, has previously been used in other machine learning applications, such as audio-video and image-text joint representation learning (Harwath et al., 2016; Chauhan et al., 2020), patch-based image registration (Simonovsky et al., 2016). Recent contrastive learning approaches (Tian et al., 2019; Hénaff et al., 2019; Chen et al., 2020; He et al., 2020) further leverage a machine learning model to differentiate "paired" and "unpaired" data mostly in the context of unsupervised representation learning. We formulate this construction motivated by mutual information estimation and provide a probabilistic interpretation of the resulting estimates.

3 EXPERIMENTS

We evaluate and compare our proposed method **DEMI** with **InfoNCE(CPC)** (Oord et al., 2018), **SMILE** (Song & Ermon, 2019) for $\tau = \{1, 5\}$, and **SMILE** with $\tau = \infty$, which is to equivalent **MINE** (Belghazi et al., 2018). For **SMILE** the two clipped ratios $\boldsymbol{\tau = 1}$ and $\boldsymbol{\tau = 5}$ under consideration are those described in Song & Ermon (2019). In all our experiments demonstrated in this paper, the hyper-parameter α, as defined in (5), is set to 1.

3.1 GAUSSIAN VARIABLES

We evaluate the performance of the MI estimators on two tractable toy problems:

- **Multivariate Gaussian variables.** We consider two correlated multivariate Gaussian variables $\boldsymbol{x}$ and $\boldsymbol{y}$. We evaluate the MI estimators on 20-d, 50-d, and 100-d Gaussian variables. As the mutual information $I(\boldsymbol{x}, \boldsymbol{y})$ is only dependent on the covariance between the two variables, we set the two marginal distributions $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ to normal distributions. We vary the correlation between two Gaussian variables to achieve different values of MI, as detailed in Poole et al. (2019).
- **Multivariate Gaussian variables with cubic transformation.** We use the same setup as in the Multivariate Gaussian variable case, except that we apply an element-wise cubic transformation on one of the variables.

Given two correlated Gaussian variables, we draw 160k i.i.d. samples for training the MI estimators and generate a different set of 10240 samples held out for testing, i.e., for estimating the MI using the trained estimators. We generate 10 independently drawn training and test sets for each two correlated Gaussian variables.

Each estimator uses the same architecture: a multi-layer perceptron with an initial concatenation layer for the x and y inputs, then two fully connected layers with ReLU activations, then a single output. This final layer uses a linear output for **MINE**, **SMILE**, and **InfoNCE**, and a logistic output for **DEMI**. We use 256 hidden units for each of the fully connected layers. We train each MI estimator for 20 epochs and set the mini-batch size to 64, using the Adam optimizer with learning rate parameter 0.0005. Architectures and settings are comparable with Song & Ermon (2019).

3.2 SELF-CONSISTENCY TESTS

We assess and compare the MI estimators using the self-consistency tests proposed in Song & Ermon (2019). We perform the tests on MNIST images (LeCun et al., 2010). The self-consistency tests examine some important properties that a "useful" MI estimator should have, because optimizing MI is more important for many downstream machine learning applications than estimating the exact value of MI.

The self-consistency tests examine: 1) capability of detecting independence, 2) monotonicity with data processing, 3) and additivity. We thus perform the following experiments, where the MNIST image set induces a data distribution and each MNIST image is a random variable that follows this data distribution:

- **MI estimation between one MNIST image and one row-masked image.** Given an MNIST image X, we mask out the bottom rows and leave the top t rows of the image, which creates $Y = h(X; t)$. The estimated MI $\hat{I}(X, Y)$ should be equal or very close to zero, if X and Y are independent. In this context, $\hat{I}(X, Y)$ should be close to 0 when t is small and be non-decreasing with t. We normalize this measurement to the final value at $t = 28$ (the last row), which should be the maximum information.
- **MI estimation between two identical MNIST images and two row-masked images.** Given an MNIST image X, we create two row-masked images: $Y_1 = h(X; t_1)$ and $Y_2 = h(X; t_2)$, where $t_1 = t_2 + 3$. Since additional data processing should not increase mutual information, $\hat{I}([X, X], [Y_1, Y_2])/\hat{I}(X, Y_1)$ should be close to 1.
- **MI estimation between two MNIST images and two row-masked images.** We randomly select two MNIST images and concatenate them: $[X_1, X_2]$, and mask the same number of

rows on them: $[h(X_1;t), h(X_2;t)] = [Y_1, Y_2]$. $\hat{I}([X_1, X_2], [Y_1, Y_2])/\hat{I}(X_1, Y_1)$ should be close to 2.

We have 60k MNIST images or concatnated images for training and a test set of 10k images. We train each MI estimator for 100 epochs and set the mini-batch size to 64. For all methods, we concatentate inputs, then convolve with a 5×5 kernel with stride 2 and 64 output channels, then

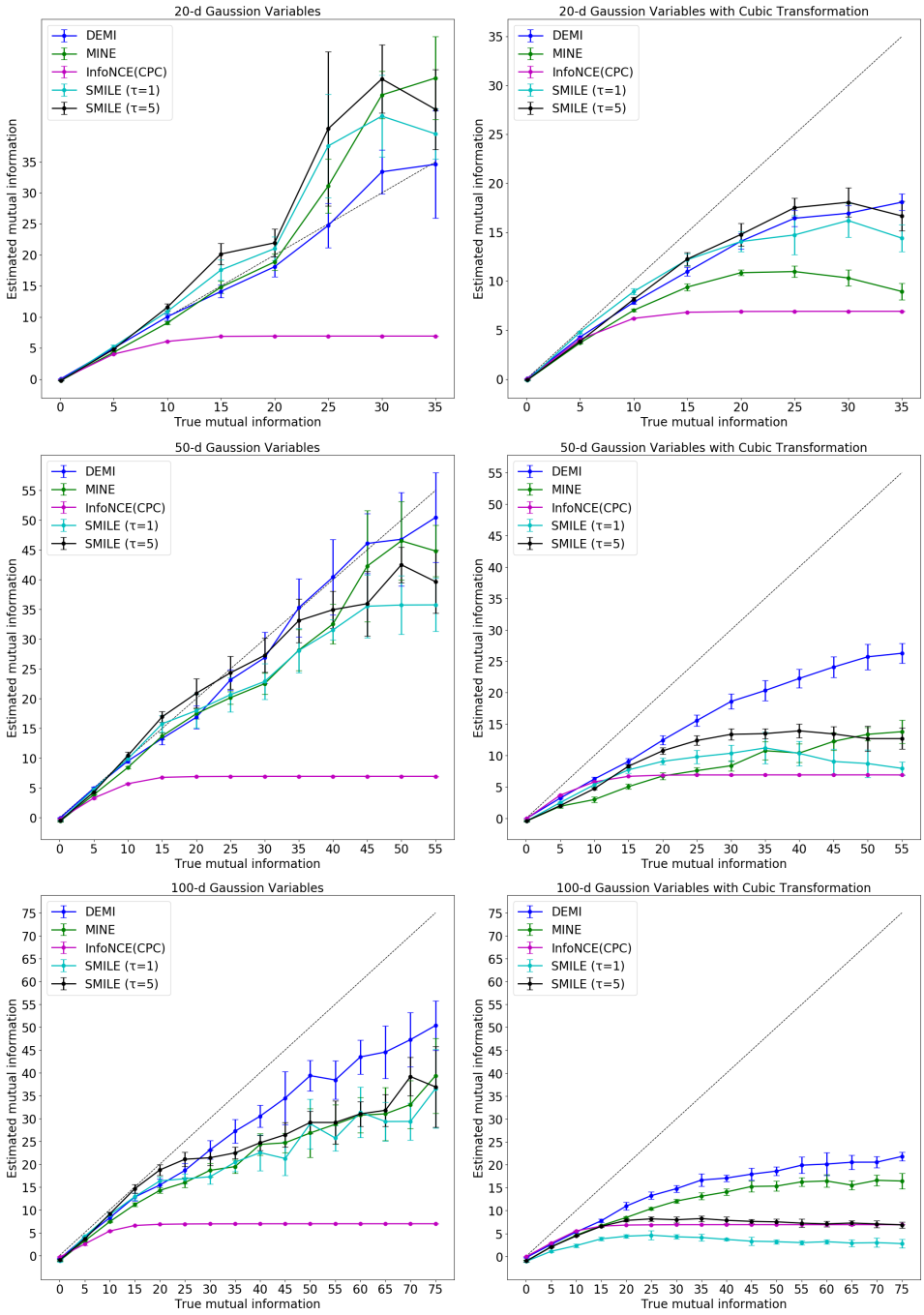

Figure 1: Mutual information estimation between multivariate Gaussian variables (left column) and between multivariate Gaussian variables with a cubic transformation (right column).

apply a fully connected layer with 1024 hidden units, which then maps to a single output. ReLU is applied after all but the last layer.

4 RESULTS

4.1 GAUSSIAN VARIABLES

In Figure 1, we show the estimated MI versus the true underlying MI for the experiments with joint Gaussians and joint Gaussians with a cubic transformation. For all experiments, InfoNCE(CPC) substantially underestimated MI. This is due to its log-batchsize ($\ln N$) maximum, which saturates quickly relative to the actual mutual information in these regimes.

Jointly Gaussian For 20-d Gaussian variables, MINE and SMILE with both parameter settings overestimated MI in comparison to DEMI, which provided estimates that were fairly close the truth.

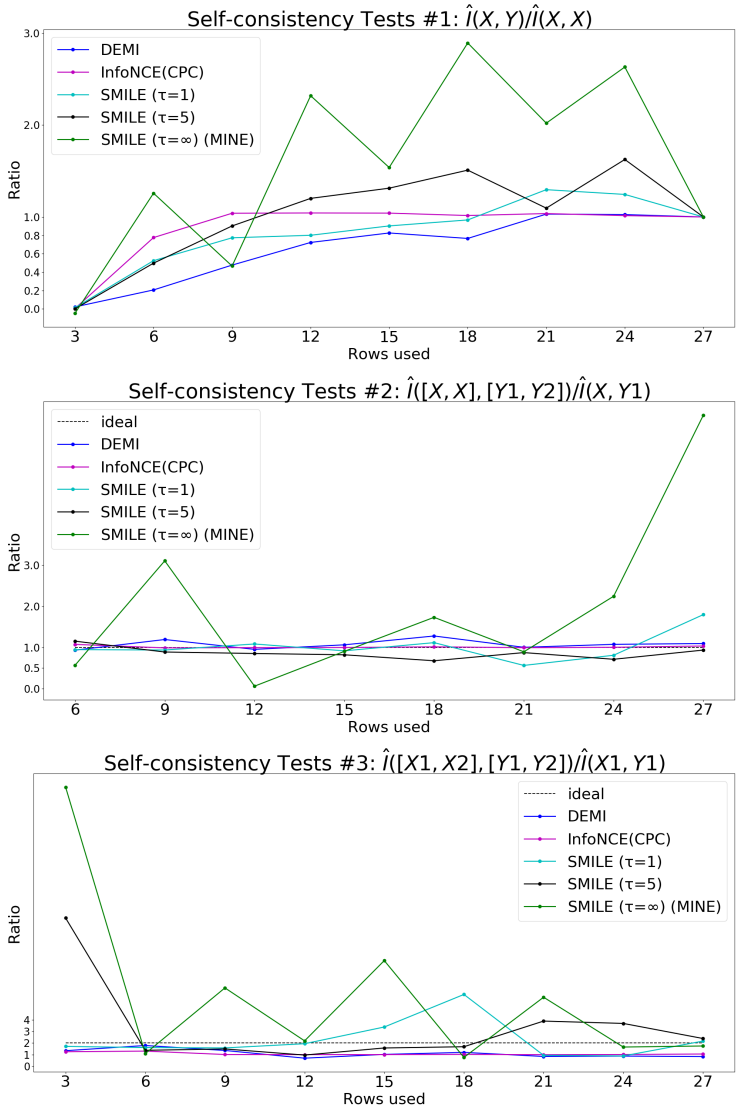

Figure 2: Results of the three self-consistency tests for SMILE ($\tau = 1, 5, \infty$), InfoNCE, and DEMI. "Monotonicity" is **top**, "data processing" is **middle**, and "additivity" is **bottom**.

Overestimation for these methods is further explored in Appendix A. For the 50-d joint Gaussian case, the situation was similar to the 20-d case, except that MINE and SMILE underestimated in comparison to DEMI, which again, was fairly close to the truth. For the 100-d joint Gaussian case, all approaches underestimated MI, with DEMI performing best.

Jointly Gaussian with Cubic Transformation For 20-d joint Gaussians with a cubic transformation all approaches underestimated MI, SMILE (τ = 5) and DEMI seemed to perform best. For the 50-d and 100-d cases all approaches understimated MI, with DEMI looking best.

In summary, for 5 of the 6 experiments DEMI appears to have performed best, in 1 experiment, DEMI and SMILE (τ=5) performed similarly.

4.2 Self-consistency Tests

In Figure 2 we plot the results of the three self-consistency metrics for each method. In general most methods perform well for the first measurement (monotonicity) with the exception of SMILE ($\tau = \infty$), which exhibits charateristicly high variance. Other settings of SMILE and DEMI both are relatively well behaved, though overall InfoNCE performs best. For the second metric ("data processing"), all methods perform well, again aside from SMILE ($\tau = \infty$) variance. In the third metric, SMILE $\tau = 1$ also exhibits a large bump in the center (where optimal should be constant 2 overall), but both InfoNCE and DEMI converge to 1 overalll, and no method performs optimally. In general InfoNCE performs best across between the first two measures, but DEMI and SMILE ($\tau = 1$) also do well in two of three.

5 Conclusion

We described a simple approach for estimating MI from joint data that is based on a neural network classifier that is trained to distinguish whether a sample pair is drawn from the joint distribution or the product of its marginals. The resulting estimator is the average over joint data of the logit transform of the classifier responses. Theoretically, the estimator converges to MI when the data sizes grow to infinity and the neural network capacity is large enough to contain the corresponding true conditional probability. We discussed close connections between our approach and the lower bound approaches of MINE and SMILE and InfoNCE(CPC). Unlike the difference-of-entropies (DoE) estimator described in McAllester & Stratos (2020), our approach does not make use of assumed distributions. Experiments on jointly Gaussian data, and similar experiments using a cubic transformation, show that our approach performed better than the MINE, SMILE and InfoNCE(CPC) in 5 out of 6 experiments; in these experiments we identified some convergence issues with SMILE for certain combinations of dimensionality and MI. In experiments with MNIST data, DEMI also performed well in MI self-consistency tests. Given it's simplicity and good performance, we believe that DEMI is a good candidate for use in research that optimizes MI for representation learning.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

David Barber and Felix V Agakov. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, pp. None, 2003.

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

Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. *arXiv preprint arXiv:2008.09884*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

Georges A Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 277–286, 2015.

David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pp. 1858–1866, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

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

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

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884, 2020.

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.

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

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180, 2019.

Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *International conference on medical image computing and computer-assisted intervention*, pp. 10–18. Springer, 2016.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2019.

Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, pp. 5–20, 2008.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2015.

Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. In *Advances in Neural Information Processing Systems*, pp. 577–585, 2014.

William M Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical image analysis*, 1(1):35–51, 1996.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. The information autoencoding family: A lagrangian perspective on latent variable generative models. *arXiv preprint arXiv:1806.06514*, 2018.

A LONG-RUN TRAINING BEHAVIOR OF SMILE

As shown in Section 4, SMILE somewhat overestimates the MI for the 20 dimensional Gaussian case in high MI regimes (~ 30 Nats or more). This did not occur at lower MI conditions or in higher dimensions.

Further investigation showed this problem to increase as training went on; to illustrate this, we set up a new experiment on the 20 dimensional Gaussian case. We ran each setting of SMILE ($\tau = 1, 5, \infty$) for 100000 training steps with batch size 64, drawing samples directly from the generating distributions. This means that the training set has effectively a very large size. We did this for three ground-truth MI values of 10,20, and 30. For comparison we also run the proposed method through the same.

This setup exactly mirrors the experiment in Song & Ermon (2019) Figure 1 in Section 6.1 of that paper, and uses their provided code and generation method, except that we replace their step-wise increasing MI schedule with a constant 10, 20, or 30 nat generator, and we run the experiment longer.

The curves for the first row of Figure 3 show good performance with relatively stable long-term behavior, particularly for $\tau = 1$. The curves in the third row of Figure 3 on the otherhand suggest that for certain distribution/domain combinations, even though SMILE and MINE are based on a lower bound of MI, they can both grossly overestimate it. This may be as McAllester & Stratos (2020) suggests due in part to a sensitivity of the estimate of $-\ln \mathbb{E}[e^{f(x,y)}]$ to outliers. The proposed method eventually overestimates as well in both the 20 and 30 nat cases, but does not have the strongly divergent behavior exhibited by SMILE (seen particularly strongly in $\tau = \infty$ settings).

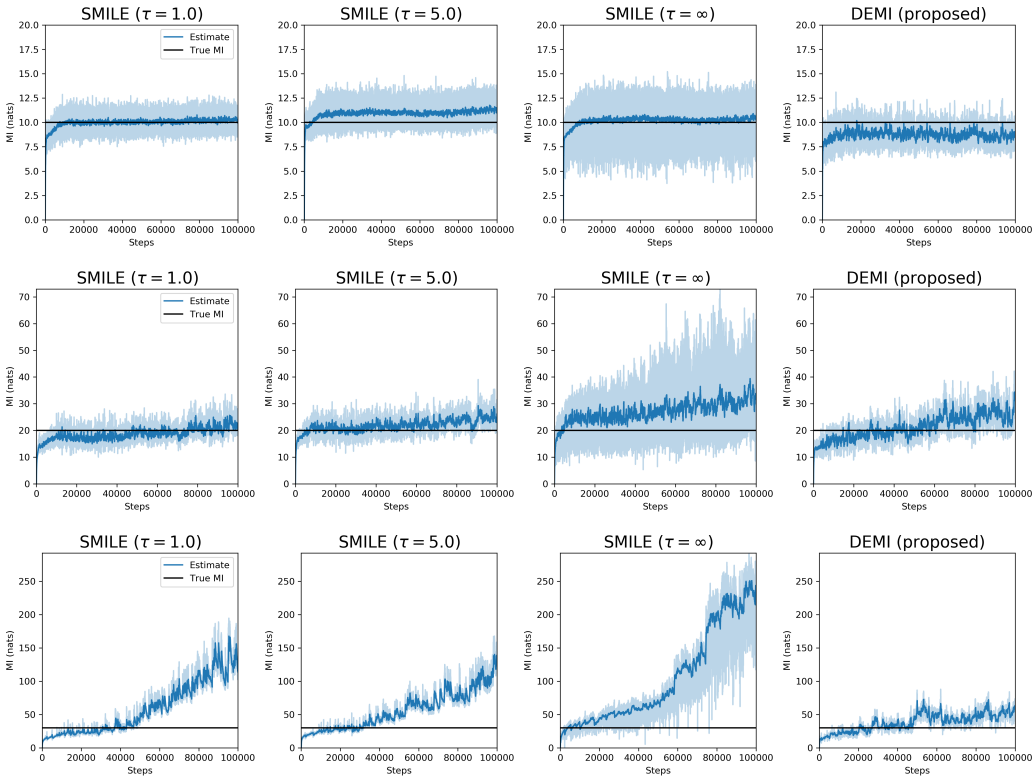

Figure 3: Long-run behavior of SMILE and DEMI for 10 (**top row**), 20 (**middle row**), and 30 (**bottom row**) Nats. Analytically SMILE converges to the MINE objective for $\tau \to \infty$. Smoothed trajectories are plotted in bold, exact trajectories are the semi-translucent curve, and the actual Mutual information is the black constant line.