Towards Robust Scale-Invariant Mutual Information Estimators

Anonymous authors
Paper under double-blind review

Appendix K

Proofs of theoretical results

Proposition 1. Let $\widehat{I}^n_{\text{bin}}(X;T)$ denote the mutual information estimated using a fixed number of bins per dimension, with bin edges defined by the minimum and maximum values of the data. Then, for all $\alpha \in \mathbb{R}^+$,

$$\widehat{I}^n_{\text{bin}}(\alpha X;\alpha T) = \widehat{I}^n_{\text{bin}}(X;T) \quad \& \quad \widehat{I}^n_{\text{bin}}(X;\alpha T) = \widehat{I}^n_{\text{bin}}(X;T).$$

Proof. For $X \in \mathbb{R}^d$, define $X_{\min}, X_{\max} \in \mathbb{R}^d$ as the componentwise extrema. Under scaling by $\alpha > 0$, we have:

$$\min(\alpha X) = \alpha X_{\min} \quad \& \quad \max(\alpha X) = \alpha X_{\max},$$

so the bin edges scale by α, and the number of bins remains fixed. Thus, each bin in X corresponds bijectively to a bin in αX with identical counts. Therefore, the empirical distribution over bins, and hence $\widehat{I}^n_{\text{bin}}$, is invariant under such scaling. □

Proposition 2. It holds that $\widehat{I}^n_{KSG}(\alpha X;\alpha T) = \widehat{I}^n_{KSG}(X;T)$, $\forall \alpha \in \mathbb{R}^+$.

Proof. Let ψ denote the digamma function(Abramowitz, 1974), and

$$n_{\alpha x,i,\infty} = \sum_{j\neq i} \mathbb{I}\{||\alpha X_i - \alpha X_j||_\infty \leq \rho_{k,i,\infty}\}, \tag{1}$$

where $\rho_{k,i,\infty}$ is the sup-norm distance of $\{\alpha X_i, \alpha T_i\}$ to its k-nearest neighbor in the joint $\{\alpha X, \alpha T\}$ space. Here $||\alpha X_i - \alpha X_j||_\infty$ represents the X-dimensions only distance.

Then, the KSG estimate of mutual information (equation 3 from (Kraskov et al., 2004) is

$$\widehat{I}^n_{KSG}(\alpha X;\alpha T) = \psi(k) + \psi(n) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_{\alpha x,i,\infty}) + \psi(n_{\alpha t,i,\infty})\right). \tag{2}$$

Let $\rho'_{k,i,\infty}$ denote the sup-norm distance of $\{X_i, T_i\}$ to its k-nearest neighbor in the joint $\{X, T\}$ space for the unscaled variables. It is trivial to see that $\rho_{k,i,\infty} = \alpha\rho'_{k,i,\infty}$.

We then have,

$$n_{\alpha x,i,\infty} = \sum_{j\neq i} \mathbb{I}\{||\alpha X_i - \alpha X_j||_\infty \leq \alpha\rho'_{k,i,\infty}\} = \sum_{j\neq i} \mathbb{I}\{||X_i - X_j|| \leq \rho'_{k,i,\infty}\} = n_{x,i,\infty}. \tag{3}$$

One can similarly show that $n_{\alpha t,i,\infty} = n_{t,i,\infty}$.

Thus, $\psi(n_{\alpha x,i,\infty}) = \psi(n_{x,i,\infty})$ and $\psi(n_{\alpha t,i,\infty}) = \psi(n_{t,i,\infty})$, and it follows that $\widehat{I}^n_{KSG}(\alpha X;\alpha T) = \widehat{I}^n_{KSG}(X;T)$. □

Proposition 3. Let $\{(X_i, T_i)\}_{i=1}^n$ be drawn i.i.d. from a bounded distribution on $\mathbb{R}^{d_X} \times \mathbb{R}^{d_T}$ that is absolutely continuous. Then, it holds almost surely that

$$\lim_{\alpha \to 0^+} \widehat{I}^n_{KSG}(X; \alpha T) = \lim_{\alpha \to \infty} \widehat{I}^n_{KSG}(X; \alpha T) = -\frac{1}{k}.$$

Thus, $\widehat{I}^n_{KSG}(X; \alpha T)$ need not be equal to $\widehat{I}^n_{KSG}(X; T)$.

Proof. Following the proof of Proposition 2, let $\rho_{k,i,\infty}$ denote the sup-norm distance from $(X_i, \alpha T_i)$ to its k-th nearest neighbor in the joint space $(X, \alpha T)$. Define

$$n_{\alpha t,i,\infty} = \sum_{j \neq i} \mathbb{I}\{||\alpha T_i - \alpha T_j||_\infty \leq \rho_{k,i,\infty}\}, \tag{4}$$

and similarly for $n_{x,i,\infty}$. The KSG estimator's estimate here will be:

$$\widehat{I}^n_{KSG}(X; \alpha T) = \psi(k) + \psi(n) - \frac{1}{k} - \frac{1}{n} \sum_{i=1}^n \left(\psi(n_{x,i,\infty}) + \psi(n_{\alpha t,i,\infty})\right). \tag{5}$$

First, we observe that

$$\lim_{\alpha \to 0^+} \left\|(X_i, \alpha T_i) - (X_j, \alpha T_j)\right\|_\infty = \lim_{\alpha \to 0^+} \max\{\|X_i - X_j\|_\infty,\ \alpha \|T_i - T_j\|_\infty\} = \|X_i - X_j\|_\infty. \tag{6}$$

Furthermore, as X and T are absolutely continuous, it holds almost surely that the k-nearest neighbor distance in X space will be greater than zero and finite.

In the limit $\alpha \to 0^+$, all pairwise distances $||\alpha T_i - \alpha T_j||_\infty \to 0$, and thus $n_{\alpha t,i,\infty} = \sum_{j \neq i} \mathbb{I}\{||\alpha T_i - \alpha T_j||_\infty \leq \rho_{k,i,\infty}\} \to n$.

Next, from equation 6 it follows that $\rho_{k,i,\infty}$ becomes the k-nearest neighbor distance in X space as $\alpha \to 0^+$, and thus we can write

$$\lim_{\alpha \to 0^+} n_{x,i,\infty} = \sum_{j \neq i} \mathbb{I}\{||X_i - X_j||_\infty \leq \rho_{k,i,\infty}\} = k. \tag{7}$$

This enables us to replace the terms in equation 5 yielding

$$\lim_{\alpha \to 0^+} \psi(k) + \psi(n) - \frac{1}{k} - \frac{1}{n} \sum_{i=1}^n \left(\psi(k) + \psi(n)\right) = -\frac{1}{k}. \tag{8}$$

Lastly, by global scale invariance (Proposition 2), $\widehat{I}^n_{KSG}(X; \alpha T) = \widehat{I}^n_{KSG}\left(\frac{1}{\alpha} X;\ T\right)$, and thus,

$$\lim_{\alpha \to 0} \widehat{I}^n_{KSG}(X; \alpha T) = \lim_{\alpha \to 0} \widehat{I}^n_{KSG}(\frac{1}{\alpha} X; T) = \lim_{\beta \to \infty} \widehat{I}^n_{KSG}(\beta X; T) = -\frac{1}{k}. \tag{9}$$

□

Proposition 4. It holds that $\widehat{I}^n_{MINE-opt}(X; \alpha T) = \widehat{I}^n_{MINE-opt}(X; T)\ \forall \alpha \in \mathbb{R}^+$.

Proof. Let f be any neural network function used in the MINE objective:

$$\mathbb{E}_{(X,T) \sim P(X,T)}[f(X, T)] - \log \mathbb{E}_{(X,T) \sim P(X) \times P(T)}[e^{f(X,T)}]. \tag{10}$$

Define a corresponding function f' for the scaled variable αT by setting the first-layer weights on T in f' to $W'_T = W_T/\alpha$, and keeping all other parameters identical. Then $f'(X, \alpha T) = f(X, T)$, so both expectations in the MINE objective remain unchanged after the transformation.

As this holds for any f, the supremum over all such functions is preserved under scaling of T. Hence, $\widehat{I}^n_{MINE\text{-}opt}(X; \alpha T) = \widehat{I}^n_{MINE\text{-}opt}(X; T)$ for all $\alpha > 0$. □

Proposition 5. Consider the MINE optimization problem with input data $S = \{(\alpha X_1, Y_1), \ldots, (\alpha X_n, Y_n)\}$ where $X \in \mathbb{R}^{d_x}$, $Y \in \mathbb{R}^{d_y}$, $(X, Y) \sim P(X, Y)$ are bounded RVs and $\alpha \in \mathbb{R}^+$ is a scaling factor. We consider a neural network of depth $L+1$ having $h_1, h_2, \ldots, h_L$ ReLU-activated hidden neurons in the respective layers. The network is trained via gradient descent on the MINE loss function in Belghazi et al. (2018) for a fixed number of iterations T, with a learning rate schedule $\eta(t) < \infty$ for all $t \leq T$. Let the weights between the i^{th} node of the $l+1^{th}$ hidden layer and the j^{th} node of the l^{th} hidden layer after t iterations be denoted by $w^l_{ji}(t)$. Assume that the initialized weights are bounded, i.e., $\forall\ (l, i, j)$, $|w^l_{ji}(0)| \leq \epsilon$ for some $\epsilon > 0$. Lastly, let $w^0_{ji}[X]$ denote the first layer weights that are attached to X. We then have, $\forall (i, j)$,

$$\lim_{\alpha \to 0^+} \left| w^0_{ji}[X](T) \right| \leq \epsilon. \tag{11}$$

Proof. Let $Z^l_j(\alpha x, y; t)$ denote the output of the j^{th} node at layer l in response to $(\alpha x, y)$ as the input at iteration t, where (x, y) is a sampled instance of $P(X, Y)$. Note that when $l = 0$, Z^l_j will be the inputs to the network $(\alpha x, y)$ itself. As we are given the dataset S, the weight update rule for w^l_{ji}, at iteration t is then $w^l_{ji}(t+1) = w^l_{ji}(t) + \Delta w^l_{ji}(t)$, where

$$\Delta w^l_{ji}(t) = -\eta(t) \frac{1}{n} \sum_{k=1}^{n} Z^l_j(\alpha X_k, Y_k; t) \delta^{l+1}_i(\alpha X_k, Y_k; t), \tag{12}$$

Here, $\delta^{l+1}_i(\alpha X_k, Y_k; t)$ denotes the backpropagation error signal at the i^{th} neuron of the $(l+1)^{\text{th}}$ layer at iteration t, computed in response to the input pair $(\alpha X_k, Y_k)$. For notational simplicity, we omit the input instance arguments in the following expressions, and write the error signal as $\delta^{l+1}_i(t)$. Let $a(\cdot)$ be the ReLU activation with its derivative $a'(\cdot) \in \{0, 1\}$. Then we have,

$$\delta^{l-1}_j(t) = a'(z^{l-1}_j(t)) \left(\sum_{i=1}^{h_l} \delta^l_i(t) w^{l-1}_{ji}(t) \right), \tag{13}$$

where $z^{l-1}_j(t)$ denotes the pre-activation output of the j^{th} node of the $(l-1)^{th}$ layer itself. Note that $Z^l_j(t) = a(z^l_j(t))$ for $l \geq 1$. As $|a'| \leq 1$, this yields:

$$|\delta^{l-1}_j(t)| \leq \sum_{i=1}^{h_l} |\delta^l_i(t) w^{l-1}_{ji}(t)|. \tag{14}$$

Let us assume the whole network function can be denoted as $f_W(X, Y) : \mathbb{R}^{d_x + d_y} \to \mathbb{R}$. Note that the network outputs a single real number, and the last layer does not have any activation function (which is ReLU for the other layers). In the context of MINE's optimization problem, the network minimizes the following loss function:

$$\widehat{I}_{\text{MINE}}(X; Y) = -\frac{1}{n} \sum_{i=1}^{n} f_W\left(\alpha X_i, Y_i\right) + \log \left(\frac{1}{n} \sum_{i=1}^{n} e^{f_W\left(\alpha X_i, \tilde{Y}_i\right)} \right). \tag{15}$$

The error signal at the last layer, $\delta^{L+1}(t)$ is the derivative of the loss w.r.t the network output. However, as the MINE optimization effectively has two distributions $P(X, Y)$ and $P(X)P(Y)$ of input, we consider the error signal of these distributions separately. The loss function for the input $X_j, Y_j \sim P(X, Y)$ is

$-\frac{1}{n} f_W(\alpha X_j, Y_j)$, which yields an error signal $\delta^{L+1}(t) = -1/n$. Whereas, the loss function for the input $X_j, \tilde{Y}_j \sim P(X)P(Y)$ is $\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{f_W(\alpha X_i, \tilde{Y}_i)}\right)$, which yields an error signal

$$\delta^{L+1}(t) = \frac{d\left(\log\left(\frac{1}{n}\sum_{i=1}^{n} e^{f_W(\alpha X_i, \tilde{Y}_i)}\right)\right)}{df_W\left(\alpha X_j, \tilde{Y}_j\right)} = \frac{e^{f_W(\alpha X_j, \tilde{Y}_j)}}{\sum_{i=1}^{n} e^{f_W(\alpha X_i, \tilde{Y}_i)}} \leq 1. \tag{16}$$

Thus, across both cases, we have that the error signal $\left|\delta^{L+1}(t)\right| \leq 1$.

Next, we consider the case when $|\alpha| \leq A$. where $A \in \mathbb{R}^+$ is a finite real. Within this setting, using the principle of induction, we now will show that the error signal $|\delta_j^l(t)| \leq D_t \ \forall(l, j)$ and the weights $|w_{ji}^l(t)| \leq B_t$ $\forall(l, i, j)$ for some finite reals D_t, B_t.

First, let us assume weights at iteration t satisfy $|w_{ji}^l(t)| \leq B_t \ \forall(l, i, j)$ for some $B_t \in \mathbb{R}^+$. Then, it first follows from equation 14 that

$$|\delta_j^{l-1}(t)| \leq \sum_{i=1}^{h_l} B_t |\delta_i^l(t)|, \tag{17}$$

and as $\left|\delta^{L+1}(t)\right| \leq 1$, this ultimately yields

$$|\delta_j^l(t)| \leq (B_t)^{L+1-l} \prod_{v=l+1}^{L} h_v. \tag{18}$$

If we set $D_t = \max_l\{(B_t)^{L+1-l}\prod_{v=l+1}^{L} h_v\}$, we have that $|\delta_j^l(t)| \leq D_t \ \forall(l, j)$. Note that here D_t only depends on B_t and the network architecture. Following the weight update rule at iteration t from equation 12, we have:

$$\begin{aligned}
|w_{ji}^l(t+1)| &= \left| w_{ji}^l(t) - \frac{\eta(t)}{n}\sum_{k=1}^{n} Z_j^l(\alpha X_k, Y_k; t)\delta_i^{l+1}(t) \right|, \\
&\leq |w_{ji}^l(t)| + \frac{\eta(t)}{n}\sum_{k=1}^{n} \left|Z_j^l(\alpha X_k, Y_k; t)\right| \left|\delta_i^{l+1}(t)\right|, \\
&\leq B_t + \eta(t) D_t \left(\frac{1}{n}\sum_{k=1}^{n} \left|Z_j^l(\alpha X_k, Y_k; t)\right|\right).
\end{aligned} \tag{19}$$

Next, as X and Y are bounded, we can assume that every dimension $|x_i| \leq K$ and $|y_i| \leq K$ for some $K \in \mathbb{R}^+$. Now, as $|\alpha| \leq A$, the inputs αX and Y are bounded. Furthermore, as the weights $|w_{ji}^l(t)| \leq B_t$, the outputs of the layers $|Z_j^l(\alpha X_k, Y_k; t)|$ will be bounded by a finite scalar which can be written as a fixed function of B_t, K and the hidden layer counts $h_1, \ldots, h_L$. Thus,

$$|w_{ji}^l(t+1)| \leq B_t + \eta(t) D_t g(B_t, K, h_1, \ldots, h_L). \tag{20}$$

We can set $B_{t+1} = B_t + \eta(t) D_t g(B_t, K, h_1, \ldots, h_L)$, and the above then implies that $|w_{ji}^l(t+1)| \leq B_{t+1}$ $\forall(l, i, j)$.

Lastly, as the weights at iteration 0 are finite ($|w_{ji}^l(0)| \leq \epsilon$), via the principle of induction we have that for any finite t, $|w_{ji}^l(t)| \leq B_t$ for some finite B_t, when $|\alpha| \leq A$. Let us consider $B = \max_{t\in\{0,1,\ldots,T\}} B_t$. Then for the entire duration of training and for any $|\alpha| \leq A$, we have $|w_{ji}^l(t)| \leq B$.

Thus, plugging this into equation 14, we obtain, $\forall(l, j, t)$,

$$|\delta_j^{l-1}(t)| \leq \sum_{i=1}^{h_l} B |\delta_i^l(t)|. \tag{21}$$

Subsequently, following the fact that $\left|\delta^{L+1}(t)\right| \leq 1$, we get

$$\left|\delta_i^1(t)\right| \leq B^L \prod_{v=2}^{L} h_v. \tag{22}$$

Let $C = KB^L \prod_{v=2}^{L} h_v$. With this, following equation 12, we have that $|\Delta w_{ji}^0[X](t)| \leq \eta(t)\alpha C$. This yields

$$|w_{ji}^0[X](T)| \leq \left|w_{ji}^0[X](0)\right| + \alpha \left(\sum_{t=1}^{T} |\eta(t)C|\right), \tag{23}$$

$$\leq \epsilon + \alpha C \left(\sum_{t=1}^{T} \eta(t)\right). \tag{24}$$

Thus, when $\alpha \to 0^+$, we have that $\epsilon + \alpha C \left(\sum_{t=1}^{T} \eta(t)\right) \to \epsilon$, yielding $\lim_{\alpha \to 0^+} \left|w_{ji}^0[X](T)\right| \leq \epsilon$. □

Proposition 6. We consider the same setting as Proposition 5 for the MINE estimation problem. There, it holds that $\lim_{\alpha \to 0^+} \widehat{I}^n_{MINE-sgd}(X; \alpha T) = 0$. Thus, $\widehat{I}^n_{MINE-sgd}(X; \alpha T)$ need not be equal to $\widehat{I}^n_{MINE-sgd}(X; T)$.

Proof. Let f^* denote the neural network function learned via stochastic gradient descent (SGD) that maximizes the following MINE objective for $\widehat{I}^n_{MINE\text{-}sgd}(X; \alpha T)$:

$$\mathbb{E}_{X,\alpha T \sim P(X,\alpha T)} \left[f^*(X, \alpha T)\right] - \log \ \mathbb{E}_{X,\alpha T \sim P(X) \times P(\alpha T)} \left[e^{f^*(X,\alpha T)}\right].$$

Let W_T be the first-layer weights in f^* connected to αT. From Proposition 5, as $\alpha \to 0^+$, we have $|W_T|_{ij} \leq \epsilon$, thus yielding finite weights. Hence, the contribution of αT to the output of a neuron of the first hidden layer of f^* becomes $\alpha T \epsilon d_T \to 0$ as $\alpha \to 0^+$, where d_T is the dimensionality of T. Therefore, $f^*(X, \alpha T) = f^*(X)$ becomes independent of T in the limit $\alpha \to 0$.

In this case, both expectations in the above objective reduce to empirical averages of the same function $f^*(X)$, and Jensen's inequality implies

$$\frac{1}{n} \sum f^*(X_i) - \log \left(\frac{1}{n} \sum e^{f^*(X_i)}\right) \leq 0,$$

with equality if and only if $f^*(X)$ is constant. Thus, the maximized objective converges to zero.

Hence, $\lim_{\alpha \to 0^+} \widehat{I}^n_{MINE\text{-}sgd}(X; \alpha T) = 0$, showing that $\widehat{I}^n_{MINE\text{-}sgd}(X; \alpha T)$ need not equal $\widehat{I}^n_{MINE\text{-}sgd}(X; T)$. □

Proposition 7. We consider the same setting as Proposition 5 for the MINE estimation problem. There, it holds that $\lim_{\alpha \to 0^+} \widehat{I}^n_{MINE-sgd}(X; \alpha T) = 0$. Thus, $\widehat{I}^n_{MINE-sgd}(X; \alpha T)$ need not be equal to $\widehat{I}^n_{MINE-sgd}(X; T)$.

Proof. Let f^* denote the neural network function learned via stochastic gradient descent (SGD) that maximizes the following MINE objective for $\widehat{I}^n_{MINE\text{-}sgd}(X; \alpha T)$:

$$\mathbb{E}_{X,\alpha T \sim P(X,\alpha T)} \left[f^*(X, \alpha T)\right] - \log \ \mathbb{E}_{X,\alpha T \sim P(X) \times P(\alpha T)} \left[e^{f^*(X,\alpha T)}\right].$$

Let W_T be the first-layer weights in f^* connected to αT. From Proposition 5, as $\alpha \to 0^+$, we have $|W_T|_{ij} \leq \epsilon$, thus yielding finite weights. Hence, the contribution of αT to the output of a neuron of the first hidden layer

of f^* becomes $\alpha T \epsilon d_T \to 0$ as $\alpha \to 0^+$, where d_T is the dimensionality of T. Therefore, $f^*(X, \alpha T) = f^*(X)$ becomes independent of T in the limit $\alpha \to 0$.

In this case, both expectations in the above objective reduce to empirical averages of the same function $f^*(X)$, and Jensen's inequality implies

$$\frac{1}{n} \sum f^*(X_i) - \log \left(\frac{1}{n} \sum e^{f^*(X_i)} \right) \leq 0,$$

with equality if and only if $f^*(X)$ is constant. Thus, the maximized objective converges to zero.

Hence, $\lim_{\alpha \to 0^+} \widehat{I}^n_{MINE\text{-}sgd}(X; \alpha T) = 0$, showing that $\widehat{I}^n_{MINE\text{-}sgd}(X; \alpha T)$ need not equal $\widehat{I}^n_{MINE\text{-}sgd}(X; T)$. □

References

Milton Abramowitz. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*. Dover Publications, Inc., USA, 1974. ISBN 0-486-61272-4.

Mohamed Ishmael Belghazi et al. Mutual information neural estimation. In Jennifer Dy and Andreas Krause (eds.), *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540. PMLR, July 2018.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, June 2004. Publisher: American Physical Society.