Variational lower bounds on mutual information based on nonextensive statistical mechanics

Anonymous authors
Paper under double-blind review

Abstract

This paper aims to address the limitations of mutual information estimators based on variational optimization. By redefining the cost using generalized functions from nonextensive statistical mechanics we raise the upper bound of previous estimators and enable the control of the bias variance trade off. Variational based estimators outperform previous methods especially in high dependence high dimensional scenarios found in machine learning setups. Despite their performance, these estimators either exhibit a high variance or are upper bounded by log(batch size). Our approach inspired by nonextensive statistical mechanics uses different generalizations for the logarithm and the exponential in the partition function. This enables the estimator to capture changes in mutual information over a wider range of dimensions and correlations of the input variables whereas previous estimators saturate them.

1 Introduction

Understanding the relationship between two variables is a fundamental problem in machine learning, finance, signal processing and other fields. To quantify such a relationship, we use mutual information, which measures the mutual dependence between two variables. The mutual information $I(X,Y)$ represents the ratio of two probabilities that can account for the nonlinear dependence and is defined as follows:

$$I(X,Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x|y)}{p(x)}\right] \tag{1}$$

It is a major challenge to estimate the mutual information in practical scenarios that involve limited samples without knowing the distributions or higher-order statistics McAllester & Statos (2018); Paninski (2003). For instance, existing methods such as the k-NN based Kraskov et al. (2004) and its variations Gao et al. (2015); Wang et al. (2009); Lord et al. (2018), or KDE based Khan et al. (2007); Suzuki et al. (2008) calculate the mutual information by estimating the probability density from the available samples. Although these approaches perform well in the low dimensional and low dependence case, they do not scale well when either the dimension of the variables or the dependence between variables increases. Such scenarios are often encountered in machine learning setups.

Estimators based on variational bounds, Belghazi et al. (2018); Poole et al. (2018); Nguyen et al. (2010); Oord et al. (2018); Zhang (2007); Foster & Grassberger (2011), perform much better in this scenarios. These estimators are inspired by the Donsker & Varadhan (1983) representation which states that there exists a function from the sample space to real number that satisfies the following equality:

$$I(X,Y) = \sup_{f:\Omega\to\mathbb{R}} \mathbb{E}_{p(x,y)}\left[f(x,y)\right] - \log \mathbb{E}_{p(y)}\left[e^{f(x,y)}\right] \tag{2}$$

Estimators based on variational bounds replace the function f in the above equation with a neural network trained to maximize a lower bound of the mutual information. The training process terminates when the lower bounds exhibit convergence, and these bounds are then interpreted as the estimated mutual information values. This NN-based approach requires good representations of lower bounds and guaranteed convergence for a wide range of dependence between the input variables which leads to numerous challenges.

Current state-of-the-art estimators, when applied to high dimensional high dependence scenarios, either exhibit a high variance or are bounded by $\log(K)$, where K is the batch size. In this work:

1. We propose new variational lower bounds on the mutual information inspired by methods from nonextensive statistical mechanics.
2. We review generalized versions of the logarithm and exponential function, define a generalized version of the partition function, and use them to control the trade off between variance and bias of the estimator.
3. We outperform previous estimators in capturing the trend when varying the correlation and the dimension of the input variables by using different generalizations for the logarithm and partition function.

2 VARIATIONAL ESTIMATORS

In what follows, we present most of the recent developments on variational lower bounds for estimation of mutual information. For a thorough review of variational estimation of mutual information we refer the reader to the work of Poole et al. (2018).

Most commonly, when estimating the mutual information, both $p(x|y)$ and $p(x)$ distributions in equation 1 are unknown. The only thing we have are samples from these distributions. In this case, to calculate the mutual information using the variational method we introduce a variational distribution $\hat{p}(x|y)$ to learn from samples. Multiplying and dividing equation 1 by this distribution and rewriting the expression we obtain the Barber & Agakov (2003) lower bound on the mutual information:

$$\begin{aligned} I(X,Y) &= \mathbb{E}_{p(x,y)}\left[\log \frac{p(x|y)\cdot\hat{p}(x|y)}{p(x)\cdot\hat{p}(x|y)}\right] \\ &= \mathbb{E}_{p(x,y)}\left[\log \frac{\hat{p}(x|y)}{p(x)}\right] + \underbrace{\mathbb{E}_{p(y)}\left[\log \frac{p(x|y)}{\hat{p}(x|y)}\right]}_{D_{\text{KL}}(p(x|y)||\hat{p}(x|y))} \\ &\geq \mathbb{E}_{p(x,y)}\left[\log \frac{\hat{p}(x|y)}{p(x)}\right] \end{aligned} \tag{3}$$

The last inequality holds since the Kullback-Liebler divergence is always non-negative. The bound is tight when $\hat{p}(x|y) = p(x|y)$, i.e., our variational distribution converges to the true distribution. To further simplify this variational optimization problem we make $\hat{p}(x|y)$ have the form in Poole et al. (2018):

$$\hat{p}(x|y) = p(x)\frac{e^{f(x,y)}}{Z(y)}, \text{ and } Z(y) = \mathbb{E}_{p(x)}\left[e^{f(x,y)}\right] = \frac{1}{K}\sum_{i=1}^{K} e^{f(x_i,y)} \tag{4}$$

where f is the critic, Z is the partition function, and K the number of samples. Equation 4 can also be written as $\hat{p}(x|y) = p(x)\cdot K\sigma(\boldsymbol{x};y)$ where $\sigma(\boldsymbol{x};y)$ is the softmax function. In this form we see that softmax function in $\hat{p}(x|y)$ maps the output of the critic to a probability. However, since the number of states (outputs) is unknown, its average is used instead of the partition function.

Replacing equation 4 into equation 3, nesting the expectations Rainforth et al. (2018), and bounding $\log Z(y) \leq \frac{Z(y)}{a(y)} + \log a(y) - 1\ \forall x, a(y) > 0$, which is a tight bound for $a(y) = Z(y)$, results in the Nguyen et al. (2010) bound, $I_{\text{NWJ}}(X,Y)$:

$$\begin{aligned} I(X,Y) &\geq \mathbb{E}_{p(x,y)}\left[f(x,y)\right] - \mathbb{E}_{p(y)}\left[\log Z(y)\right] \\ &\geq \mathbb{E}_{p(x,y)}\left[f(x,y)\right] - \mathbb{E}_{p(y)}\left[\frac{\mathbb{E}_{p(x)}[e^{f(x,y)}]}{a(y)} + \log a(y) - 1\right] \\ &\geq \frac{1}{K}\sum_{i=1}^{K} f(x_i,y_i) - \frac{1}{e\cdot K}\sum_{i=1}^{K} e^{f(x_i,y_j)} \triangleq I_{\text{NWJ}} \end{aligned} \tag{5}$$

where $a(y)$ was set to be the natural constant and the second expectation computed with shuffled y values. This bound reaches the true mutual information when the critic approaches the optimal critic $f^*(x,y) = 1 + \log \frac{p(x|y)}{p(x)}$. I_{NWJ} estimator performs well in high dimensional high dependence scenarios but exhibits high variance due to the exponential function in the partition function which amplifies the critic errors.

One way to reduce the variance is to make the critic depend on multiple samples and use a Monte Carlo estimator for the partition function. To accomplish this we set the optimal critic to depend on K additional samples $f^*(x_{1:K}, y) = 1 + \log \frac{e^{f(x,y)}}{m(x_{1:K},y)}$ where $m(x_{1:K}, y) = \frac{1}{K}\sum_{i=0}^{K} e^{f(x_i,y)}$. Replacing this new multisample critic in equation 5 and averaging over the batch after iteratively fixing each sample as x_1, we obtain the noise contrastive estimator $I_{\text{NCE}}(X, Y)$ introduced by Oord et al. (2018):

$$I(X,Y) \geq \frac{1}{K}\sum_{i=1}^{K} \log \frac{e^{f(x_i,y_i)}}{\frac{1}{K}\sum_{j=1}^{K} e^{f(x_i,y_j)}} \triangleq I_{\text{NCE}} \tag{6}$$

Although, as a result of averaging, the I_{NCE} estimator exhibits a much lower variance compared to I_{NWJ}, this estimator is upper bounded by $I_{\text{NCE}} \leq \log K$ where K is the batch size used to train the critic. This bound can be verified by extracting the K from the denominator outside the sum, $I(X,Y) \geq \log K + \langle \log \sigma(x_i, \boldsymbol{y})\rangle_i$ where the average is always non-positive since softmax returns a probability which is less or equal to 1.

One possible solution to solve the high variance of I_{NWJ} and the upper bound of I_{NCE}, proposed by Poole et al. (2018), is to use a nonlinear interpolation between the two. To accomplish that, the optimal critic for I_{NWJ}, equation 5, is set to $f^*(x_{1:k}, y) = 1 + \log \frac{e^{f(x_1,y)}}{\alpha m(x_{1:k},y)+(1-\alpha)q(y)}$ where $m(x_{1:k}, y)$ is $\frac{1}{k}\sum_{i=0}^{k} e^{f(x_i,y)}$, $q(y)$ is the secondary critic, and $\alpha \in [0,1]$ interpolates between I_{NWJ} ($\alpha = 0$) and I_{NCE} ($\alpha = 1$). Plugging this critic in equation 5 we obtain Poole et al. (2018) bound I_α:

$$\begin{aligned} I(X,Y) \geq 1 + \mathbb{E}_{p(x_{1:K})p(y|x_1)}&\left[\log \frac{e^{f(x_1,y)}}{\alpha m(x_{1:k},y) + (1-\alpha)q(y)}\right] \\ - \mathbb{E}_{p(x_{1:K})p(y)}&\left[\frac{e^{f(x_1,y)}}{\alpha m(x_{1:k},y) + (1-\alpha)q(y)}\right] \triangleq I_\alpha \end{aligned} \tag{7}$$

Here, similar to I_{NWJ} case, the second expectation is calculated using y values independently sampled from x. The nonlinear interpolation in this case increases the upper bound of the estimator from $\log K$ to $\log(K/\alpha)$.

Although these estimators perform much better in high dimensional high dependence scenarios than the ones based on k-NN or KDE, they still have their own limitations. In particular, I_{NWJ} has a high variance due to the exponential in the partition function that limits its use. I_{NCE} solves the variance problem by averaging the partition function which results in the estimator being bounded by $log(K)$ where K is the batch size used to train the critic. For the most common values of the batch size, the upper bound of this estimator is between 6 and 8, much lower than the values of mutual information encountered in the machine learning setup. I_α estimator, which is a combination of the two, can reach higher values than I_{NCE}, however, the variance is not far from I_{NWJ} variance and it requires training two critics.

In what follows we first present a generalization of the exponential and logarithm function from the nonextensive statistical mechanics. Then we use these functions to derive a generalized version of the previous bounds and show how, by using the generalization parameter q, we can overcome their limitations.

3 NONEXTENSIVE STATISTICAL MECHANICS

The most known universality class in statistical mechanics is the Boltzmann–Gibbs class. In this class the entropy is defined in terms of a probability set, and in particular, for a thermodynamic system at equilibrium it is calculated as:

$$S_{\mathrm{BG}} = -k_{\mathrm{B}} \sum_{i=0}^{W} p_i \log(p_i) \text{ with } p_i = \frac{e^{-\beta E_i}}{Z_{\mathrm{BG}}} \text{ and } Z_{\mathrm{BG}} = \sum_{j=1}^{W} e^{-\beta E_j} \tag{8}$$

where E_i is the energy of the system in state i, $\beta = 1/k_{\mathrm{B}}T$ is the thermodynamic beta, and Z_{BG} is the partition function that encodes the distribution of the probabilities among states. This expression found much attention in many fields outside statistical mechanics and most notably in information theory. In the information theoretic context, the Bolzmann constant $k_{\mathrm{B}} = 1$ and the pair $-\beta E_i$ is an unnormalized function for which the entropy is calculated. This is exactly the case of expression 4 where we normalize the critic using the partition function.

The Boltzmann-Gibbs entropy was generalized in the work of Tsallis (1988) which introduces a parameter q, named generalization, to control weight of the individual probabilities. Tsallis generalized entropy is defined as:

$$S_q = k\frac{1 - \sum_{i=1}^{W} p_i^q}{q-1} = -k\frac{1}{W}\sum_{i=1}^{W} \log_q(p_i) \tag{9}$$

where $\log_q$ is the q-logarithm which converges to the classical logarithm when $q \to 1$. The generalization q can be viewed as a “bias”. When $q < 1$ smaller probabilities are amplified, whereas when $q > 1$ the larger ones are amplified. Although a majority of systems are well described by the classical entropy and partition function, more complex system such as brain activity Tong et al. (2002), financial markets Michael & Johnson (2003), and black holes Majhi (2017) often do not follow the classical laws and are better captured by the generalized version.

Over the years, Tsallis entropy evolved into a separate branch of statistical mechanics known as nonextensive statistical mechanics Tsallis (2009). This new branch extensively makes use of generalized logarithm and exponential shown in Figure 1a and 1b defined as:

$$\begin{aligned} \exp_q(x) &= [1 + (1-q)x]_+^{\frac{1}{1-q}} \\ \log_q(x) &= \frac{x^{1-q} - 1}{1-q} \end{aligned} \tag{10}$$

where $[z]_+ = \max(z, 0)$. The following identities, which we will use in the next section to derive new variational bounds, are true: $\exp_q(x+y) = \exp_q(x) \otimes_q \exp_q(y)$ and $\log_q(xy) = \log_q(x) \oplus_q \log_q(y)$ where $\oplus_q$ and $\otimes_q$ are the q-addition and q-multiplication. For the definition of the q-operators and a brief introduction to q-algebra, consult Appendix B.

Similar to generalized entropy, the generalized mutual information is defined as:

$$I_q(X, Y) = \mathbb{E}_{p(x,y)}\left[\log_q \frac{p(x|y)}{p(x)}\right] \tag{11}$$

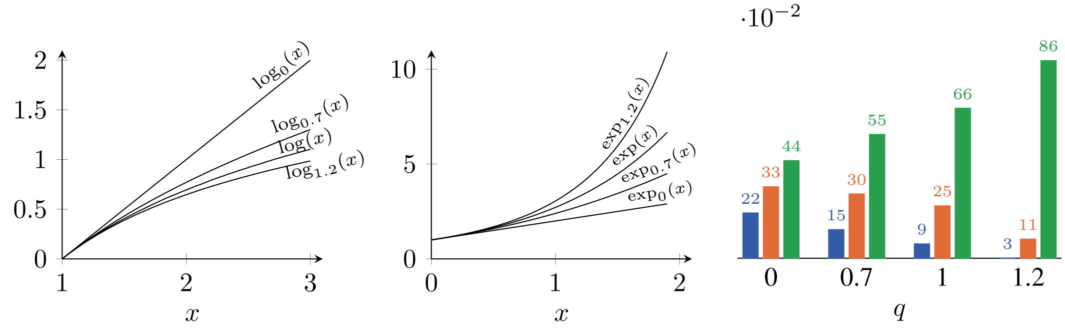
Moreover, we can define a generalized version of the partition function using the generalized exponential:

$$Z_q(x) = \frac{1}{N}\sum_{i=1}^{N} \exp_q(x_i) \tag{12}$$

The advantage of the generalized partition function is the freedom to choose how much weight large values will receive when mapping to probability. Figure 1c shows how the triplet $(1, 2, 3)$ is mapped into probabilities for different values of q. We will use this in the next section to improve the perfomance of the estimators.

4 Variational estimators based on Tsallis statistics

In this section, we derive a generalized version of previously introduced bounds using Tsallis statistics in order to overcome the limitations of the previous variational bounds. Although, the generalized mutual information differs in value with the classical mutual information due to the generalization.



(a)

(b)

(c)

Figure 1: Generalized functions. **(a)** q-generalization of the logarithm function. When the generalization parameter $q \to 1$ we recover the classical logarithm $log(x)$, for $q = 0$ a linear function that passes through $(1, 0)$, and for other values of q a nonlinear interpolation between the two. **(b)** q-generalization of the exponential function. The behaviour of $\exp_q(x)$ is similar to $\log_q(x)$ with respect to the generalization parameter q. For the same value of q the two functions are the inverse of each other, i.e., $\exp_q(\log_q(x)) = x$ and $\log_q(\exp_q(x)) = x$. **(c)** Normalizing the triplet $(1, 2, 3)$ using the q-generalization of the partition function for different values of q. Higher the value of q, higher the probability mass is attributed to the largest value in the set.

The two have the same properties and capture the dependence between input variables. To obtain a variational bound on the generalized mutual information we start from the definition, expression 11, which we multiply and divide by the variational distribution $\hat{p}(x|y)$:

$$\begin{aligned} I_q(X,Y) &= \mathbb{E}_{p(x,y)}\left[\log_q \frac{p(x|y)\cdot\hat{p}(x|y)}{p(x)\cdot\hat{p}(x|y)}\right] \\ &= \mathbb{E}_{p(x,y)}\left[\log_q \frac{\hat{p}(x|y)}{p(x)}\right] \oplus_q \underbrace{\mathbb{E}_{p(y)}\left[\log_q \frac{p(x|y)}{\hat{p}(x|y)}\right]}_{D_{\mathrm{KL}_q}(p(x|y)||\hat{p}(x|y))} \\ &\geq \mathbb{E}_{p(x,y)}\left[\log_q \frac{\hat{p}(x|y)}{p(x)}\right] \end{aligned} \tag{13}$$

Again the inequality holds since the $D_{\mathrm{KL}_q}(p(x|y)||\hat{p}(x|y))$ is a positive number for all values of q and is tight when $\hat{p}(x|y) = p(x|y)$ as 0 is a neutral element for the generalized q-addition. Again, this expression can be further simplified by selecting $\hat{p}(x|y)$ be of the form:

$$\hat{p}(x|y) = p(x)\frac{\exp_q(f(x,y))}{Z_q(y)}, \text{ and } Z_q(y) = \mathbb{E}_{p(x)}\left[\exp_q(f(x,y))\right] = \frac{1}{N}\sum_{i=1}^{N}\exp_q(f(x_i,y)) \tag{14}$$

where Z_q is the generalized partition function. For $q = 0$ we recover the counting measure, for $q = 1$ we obtain the softmax function, and for other value we obtain a nonlinear interpolation between the two. In what follows we set the q of the partition function equal to the Tsallis divergence, however, there is no reason to have them equal as the partition function has the role of a normalizer. Replacing equality 14 into 13 we obtain:

$$I(X,Y) \geq \mathbb{E}_{p(x,y)}\left[f(x,y)\right] \ominus_q \mathbb{E}_{p(y)}\left[\log_q Z_q(y)\right] \tag{15}$$

Next, to obtain the Nguyen, Wainwright and Jordan bound, I_{NWJ}, we bound the partition function using the following inequality $\log_q Z_q(y) \leq \frac{Z_q(y)}{a(y)} + \log_q a(y) - 1 \; \forall x, q, a(y) > 0$ which is tight when $a(y) = Z_q(y)$. Setting $a(y) = \exp_q(y)$ yields the bound:

$$I(X,Y) \geq \mathbb{E}_{p(x,y)}\left[f(x,y)\right] \ominus_q \exp_q(-1)\, \mathbb{E}_{p(y)}\left[Z_q(y)\right] \tag{16}$$

which reaches the true mutual information when the critic approaches the optimal critic $f^*(x,y) = 1 \oplus_q \log_q \frac{p(x|y)}{p(x)}$ that can be verified by replacing this expression for critic in equation 16.

To obtain the q-equivalent for the multisample bound I_{NCE} we replace the optimal critic with the multisample optimal critic $f^*(x_{1:k}, y) = 1 \oplus_q \log_q \frac{\exp_q(f(x,y))}{m(x_{1:K},y)}$ with $m(x_{1:K}, y) = \frac{1}{K}\sum_{i=0}^{K} \exp_q(f(x_i, y))$ in inequality 16 which results in:

$$I(X,Y) \geq 1 \oplus_q \mathbb{E}_{p(x_{1:K})p(y|x_1)}\left[\log_q \frac{\exp_q(f(x,y))}{m(x_{1:K},y)}\right] \ominus_q \mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{\exp_q(f(x,y))}{m(x_{1:K},y)}\right] \tag{17}$$

By iteratively reindexing each sample in $x_{1:K}$ as x_1 and averaging the result, the last term becomes:

$$\frac{1}{K}\sum_{i=1}^{K} \mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{\exp_q(f(x,y))}{m(x_{1:K},y)}\right] = \mathbb{E}_{p(x_{1:K})p(y)}\left[\frac{\frac{1}{K}\sum_{i=1}^{K}\exp_q(f(x,y))}{m(x_{1:K},y)}\right] = 1 \tag{18}$$

Replacing expression 18 into 17, and then as in previous step iteratively reindexing each sample in $x_{1:K}$ as x_1 and averaging the result we obtain:

$$\begin{aligned} I(X,Y) &\geq \frac{1}{K}\sum_{i=1}^{K} \mathbb{E}\left[\log_q \frac{\exp_q(f(x,y))}{m(x_{1:K},y)}\right] \\ &\geq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} \log_q \frac{\exp_q(f(x_i,y_i))}{\frac{1}{K}\sum_{j=0}^{K}\exp_q(f(x_i,y_j))}\right] \end{aligned} \tag{19}$$

which is our proposed lower bound on mutual information based on nonextensive statistical mechanics which we name accordingly I_{NES}. Similar to the original I_{NCE}, equation 6, our estimator is upper bounded by $\log_q(K)$. However, we can overcome this limitation by choosing a generalization $q < 1$ for the logarithm. In addition, since the generalization of the logarithm is independent of the generalization of the partition function we can set them independently. Thus, we control how the critic output is mapped to a probability using the generalization q of the partition function and the estimator bound through the generalization q of the logarithm. To show this, we rewrite expression 19 as:

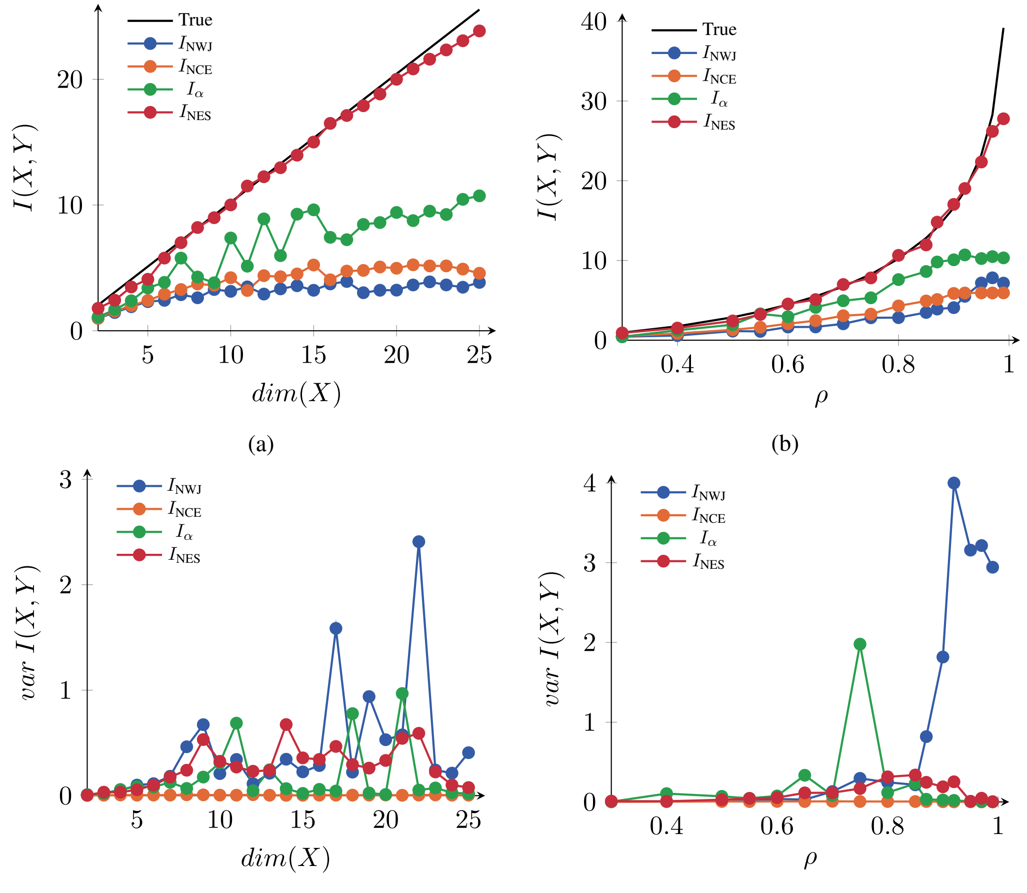
$$I(X,Y) \geq \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} \log_{q_1} \frac{\exp_{q_2}(f(x_i,y_i))}{\frac{1}{K}\sum_{j=0}^{K}\exp_{q_2}(f(x_i,y_j))}\right] \triangleq I_{\text{NES}} \tag{20}$$

where q_1 is the logarithm generalization and q_2 is the partition function generalization.

5 Experimental Results

We use Gaussian distributed random vectors to test our estimator and compare its performance with previous works. For the particular case when the vector elements with the same index have correlation ρ and 0 when indices do not match, the mutual information can be calculated using the following formula:

$$I(X,Y) = \frac{\dim(X)}{2}(1-\rho)^2 \tag{21}$$



(a)

(b)

(c)

(d)

Figure 2: Estimated mutual information from 3000 samples and using a batch size of 512. The two parameters of I_{NES} are 0.6 for the logarithm and 2 for the partition function, and $\alpha = 0.01$ for the I_α estimator. **(a)** The influence of the dimension on the estimated mutual information for a correlation ρ of 0.8. I_{NWJ} and I_{NCE} estimators saturate for $dim > 10$, I_α captures the trend but fluctuates as it require both critics to converge. I_{NES} performs the best although the returned of value of this estimator is not directly comparable with other estimators due to the generalized logarithm. **(b)** Impact of the vector correlation on the estimated mutual information for input vectors of dimension 10. All estimators follow the trend up to a correlation ρ of 0.9. Only I_{NWJ} and I_{NES} perform beyond this limit with the later capturing the trend even for a correlation of 0.99. **(c)** Variance of the estimated bound for different dimensions of the input vector. I_{NES} exhibits a higher variance than I_{NCE} and I_α due to the generalization $q > 1$ used in the partition function. **(d)** The influence of the correlation on the variance of the estimated mutual information.

For each experiment we use 3000 samples to estimate the mutual information and train the critic in batches of 512 samples to maximize the variational bounds introduced in the previous section. The result of these experiments are shown in Figure 2.

We observe that for low dimensions $dim(X) < 10$, a majority of the estimators can capture the trend, Figure 2a. However, for dimensions $dim(X) > 10$, I_{NWJ} and I_{NCE} saturate. Although both I_{NES} and I_α do follow the trend of the true mutual information for all input dimensions, I_{NES} outperforms I_α which uses two critics and requires both of them to converge. Note, the returned value of I_{NES} estimator is not a classical mutual information as it uses a generalized logarithm, however, it has the same properties as the classical mutual information.

In contrast to the input dimension experiment, estimators capture the trend much better over a wider range of correlations between inputs, Figure 2b. The two estimators that saturate first in this case are I_{NCE} and I_α for correlations $\rho > 0.9$. I_{NES} captures the trend and remains consistent across the whole range up to $\rho > 0.99$. From Figures 2c and 2d we can see that the performance of the I_{NES} estimators comes with a cost, it exhibits higher variance than I_{NCE} but lower than the one of I_{NWJ}.

6 Conclusion

The main goal of the current work was to improve the performance of mutual information estimators in the high dependence high dimensional scenarios which are often encountered in machine learning setups. We reviewed previous variational lower bounds and extended them using generalized logarithm and exponential functions from nonextensive statistical mechanics. One of the most significant findings to emerge from this study was that we can control the trade off between the bias and the variance of the estimator by independently tuning the generalizations q of the logarithm and partition function. As a result, we are able to better capture the trend when varying the correlation and dependence of the input variables. This method greatly improves upon the I_{NCE} estimator which has a low variance but a high bias and I_α estimator which requires two critics that can sometimes be challenging to train. The major limitation of the proposed estimator is that its results are not equal in value with the classical mutual information due to use of generalization. Despite that, I_{NES} still captures the dependence between the input variable and can be applicable to machine learning problems where a mutual information estimator is needed such as feature selection, variational autoencoders, and generative adversarial networks.

References

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.

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.

David V Foster and Peter Grassberger. Lower bounds on mutual information. *Physical Review E*, 83 (1):010101, 2011.

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.

Shiraj Khan, Sharba Bandyopadhyay, Auroop R Ganguly, Sunil Saigal, David J Erickson III, Vladimir Protopopescu, and George Ostrouchov. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2): 026209, 2007.

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

Warren M Lord, Jie Sun, and Erik M Bollt. Geometric k-nearest neighbor estimation of entropy and mutual information. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(3):033114, 2018.

Abhishek Majhi. Non-extensive statistical mechanics and black hole entropy from quantum geometry. *Physics Letters B*, 775:32–36, 2017.

David McAllester and Karl Statos. Formal limitations on the measurement of mutual information. *arXiv preprint arXiv:1811.04251*, 2018.

Fredrick Michael and MD Johnson. Financial market dynamics. *Physica A: Statistical Mechanics and its Applications*, 320:525–534, 2003.

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.

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

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A Alemi, and George Tucker. On variational lower bounds of mutual information. In *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

Tom Rainforth, Robert Cornish, Hongseok Yang, and Andrew Warrington. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pp. 4264–4273, 2018.

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.

S Tong, A Bezerianos, J Paul, Y Zhu, and N Thakor. Nonextensive entropy measure of eeg following brain injury from cardiac arrest. *Physica A: Statistical Mechanics and its Applications*, 305(3-4): 619–628, 2002.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.

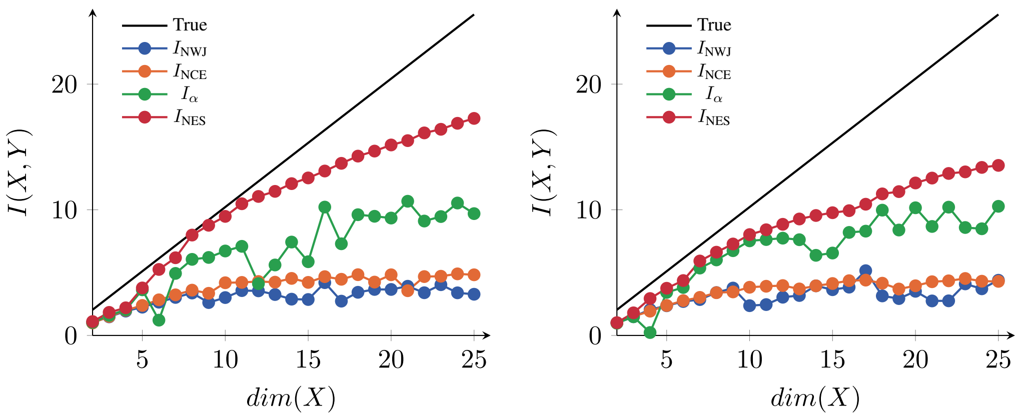
Constantino Tsallis. *Introduction to nonextensive statistical mechanics: approaching a complex world*. Springer Science & Business Media, 2009.

Sabir Umarov, Constantino Tsallis, and Stanly Steinberg. On a q-central limit theorem consistent with nonextensive statistical mechanics. *Milan journal of mathematics*, 76(1):307–328, 2008.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5): 2392–2405, 2009.

Zhengmin Zhang. Estimating mutual information via kolmogorov distance. *IEEE Transactions on Information Theory*, 53(9):3280–3282, 2007.

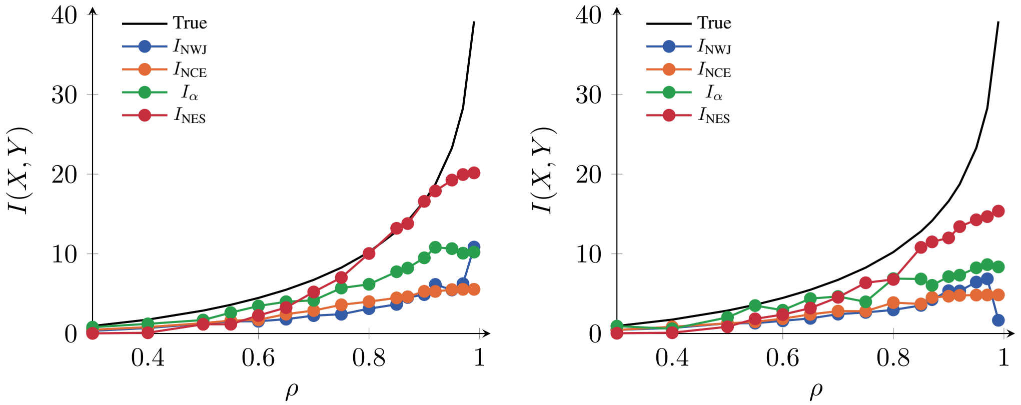
A Additional experiments



(a) Batch size 256

(b) Batch size 128

Figure 3: Impact of the input dimension on mutual information estimation using 3000 samples.



(a) Batch size 256

(b) Batch size 128

Figure 4: Impact of the correlation between vectors on mutual information estimation using 3000 samples.

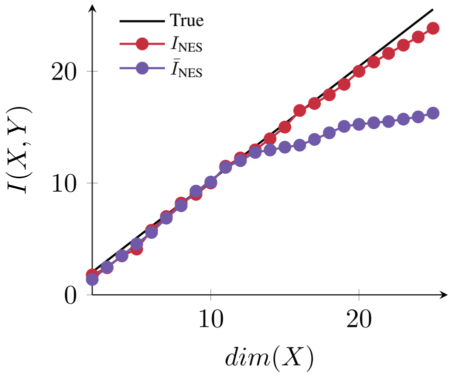


Figure 5: Comparison of two version of I_{NES}. For $\bar{I}_{\text{NES}}$ both generalizations, logarithm and exponential, are equal and set to 0.75. For I_{NES} the two generalization are distinct and set to 0.6 and 2 respectively. Mutual information was estimated using 3000 samples with a correlation $\rho = 0.8$.

B Q-ALGEBRA

Brief introduction to q-algebra. For a more in depth treatment consult the works of Tsallis (2009); Umarov et al. (2008).

Addition and subtraction operations in q-algebra are defined as follow:

$$\begin{aligned} x \oplus_q y &= x + y + (1-q)xy \\ x \ominus_q y &= \frac{x-y}{1+(1-q)y} \end{aligned} \tag{22}$$

Similar to the classical algebra, above operations have 0 as the neutral element:

$$\begin{aligned} x \oplus_q 0 &= x \\ x \ominus_q 0 &= x \end{aligned} \tag{23}$$

Multiplication and division are have the form:

$$\begin{aligned} x \otimes_q y &= \left[x^{1-q} + y^{1-q} - 1\right]_+^{\frac{1}{1-q}} \\ x \oslash_q y &= \left[x^{1-q} - y^{1-q} + 1\right]_+^{\frac{1}{1-q}} \end{aligned} \tag{24}$$

where $[z]_+ = \max(z, 0)$. For both operators 1 is a neutral element:

$$\begin{aligned} x \otimes_q 1 &= x \\ x \oslash_q 1 &= x \end{aligned} \tag{25}$$

Exponential and logarithm are defined as:

$$\begin{aligned} \exp_q(x) &= [1 + (1-q)x]_+^{\frac{1}{1-q}} \\ \log_q(x) &= \frac{x^{1-q} - 1}{1-q} \end{aligned} \tag{26}$$

with properties:

$$\begin{aligned} \exp_q(0) &= 1 \\ \log_q(1) &= 0 \\ \exp_q(\log_q(x)) &= x \\ \log_q(\exp_q(x)) &= x \end{aligned} \tag{27}$$

However, exponential and logarithmic identities have a slightly different form:

$$\begin{aligned} \exp_q(x+y) &= \exp_q(x) \otimes_q \exp_q(y) \\ \exp_q(x \oplus_q y) &= \exp_q(x) \cdot \exp_q(y) \\ \log_q(xy) &= \log_q(x) \oplus_q \log_q(y) \\ \log_q(x \otimes_q y) &= \log_q(x) + \log_q(y) \end{aligned} \tag{28}$$