On Variational Generalization Bounds for Unsupervised Visual Recognition

Karush Suri[1], **Mahdi Haghifam**[1,2], **Ashish Khisti**[1]
[1]University of Toronto, [2]Vector Institute
karush.suri@mail.utoronto.ca

Abstract

Recent advancements in generalization bounds have led to the development of tight information theoretic and data-dependent measures. Although generalization bounds reduce bias in estimates, they often suffer from tractability during empirical evaluation. The lack of a uniform criterion for estimation of Mutual Information (MI) and selection of divergence measures in conventional bounds hinders utility to sparse distributions. To that end, we revisit generalization through the lens of variational bounds. We identify hindrances based on bias, variance and learning dynamics which prevent accurate approximations of data distributions. Our empirical evaluation carried out on large-scale unsupervised visual recognition tasks highlights the necessity for variational bounds as generalization objectives for learning complex data distributions. Approximated estimates demonstrate low variance and improved convergence in comparison to conventional generalization bounds. Lastly, based on observed hindrances, we propose a theoretical alternative which aims to improve learning and tightness of variational generalization bounds. The proposed approach is motivated by contraction theory and yields a lower bound on MI.

1 Introduction

Generalization bounds provide tight measures which facilitate the learning of distributions under sparse data. The work of Russo et. al. [1] has led to drastic improvements [2, 3] in bounding generalization error with information theoretic metrics. The surge of information theoretic metrics [2, 4] has further motivated improvements in bias reduction for control measurements [**?**]. While generalization bounds tighten the dynamics of sparse learning, a tighter approximation often hurts the performance in the presence of out-of distribution samples [5]. In many such scenarios, it is difficult to empirically evaluate the performance of the bound [6]. Additionally, the abundance of divergence metrics does not provide a selection criterion for an optimal information theoretic entity [7, 8]. This allows one to rethink the feasibility of conventional bounds in practical scenarios.

Variational bounds [9] are a class of probabilistic bounds which depict increasing potential for learning [5, 10, 11]. A typical variational bound utilizes a tractable data distribution which can be approximated with limited data samples. This property of variational measures motivates data-efficient learning [12]. Tractibility of variational bounds for information maximization and minimization allows multiple objective functions to be realisable in a given problem setting [9]. Variational bounds can then be flexibly modeled as lower and upper bounding measures of information [9]. However, large-scale utilization of multi-sample variational bounds is an open problem for unsupervised learning tasks [9]. Data-efficient learning in conjunction with tractable compatibility to data distributions presents variational bounds as suitable candidates for learning objective functions.

We revisit the regime of generalization bounds from the perspective of information theoretic and variational distributions. The work highlights the suitability of variational bounds in comparison to conventional generalization bounds which emphasize only on the bias in data estimates. Variational objectives tackle high bias as well as high variance estimates. Our main contributions are threefold-

- We revisit generalization in light of variational learning and identify hindrances which prevent accurate approximations of data distributions.
- We empirically demonstrate the suitability of variational generalization bounds on unsupervised visual recognition tasks wherein the data distribution is inherently challenging to approximate. Our evaluation highlights the necessity for variational generalization bounds.
- We conjecture a theoretical alternative which aims to address the hindrances discovered in learning variational generalization bounds. The proposed approach is motivated by contraction theory and yields a lower bound on MI.

2 Related Work

Variational Bounds: A number of methods [9, 5, 13, 11, 12] introduce variational bounds for information-based learning. MINE [5] presents the estimation of MI utilizing gradient descent for high-dimensional random variables. Suitability of MINE leads to improved adversarial generative models and supervised classification tasks. InfoMAX [13], extends the MI framework by simultaneously estimating and maximizing information between output representations and input prior distributions. InfoMAX scales well to unsupervised learning scenarios and sparse latent distributions. While, MINE and InfoMAX highlight the practical utility of information estimation, they do so at the cost of large data requirements from the input distribution. CPC [11] and CPCv2 [12] aid data-efficient learning by introducing the InfoNCE bound. The InfoNCE objective eliminates the need for explicit estimation of MI by providing a lower bound on MI. InfoNCE being a multi-sample bound [9], scales well in the number of data samples in-distribution. However, the objective is hindered by large batch sizes and is not tight for large values of MI. The recently proposed interpolation bounds [9] extend the InfoNCE setup towards a continuum of bounds which trade-off bias with variance. Additionally, the bound is tight for varying batch sizes. Our work is orthogonal to the proposed interpolation scheme and extends it to the generalization setup.

Generalization Bounds: The pivotal work of [1] provides a lower-bound on MI based on information-theoretic measures [14, 7]. The MI bound [15] is further improved as a result of tight lower bounds on MI minimizing the generalization error [2, 4]. Additional measures such as data-dependent estimates [3] and the specific choice of distributions [16] extend the application of lower bounds to stochastic learning dynamics [4, 17] and differential privacy [1**?**]. The bounds are further sharpened using conditional MI [18] in a sample-based framework [19] which extends the data-dependent scheme of [3]. A more suitable application is the setting of adaptive control [6] which is based on high stochasticity stemming from continuous measurements. The bound provided in [6] aims to address this problem with the introduction of $alpha$ divergence metrics [8] which serve as a lower bound on MI. While the bound is proven to be theoretically tight, its application and empirical evaluation remain an open problem in literature. We aim to leverage the theoretical contributions of [6] in order to provide a variational alternative which can be empirically realised.

Unsupervised Visual Recognition: One of the main applications of information-theoretic bounds is unsupervised learning for visual recognition tasks [10]. The information maximization framework [13] reduces local sparsity and motivates the learning of richer representations [11]. Multi-sample bounds such as InfoNCE in CPC [11] and CPCv2 [12] contrast augmented representations with actual input samples in order to maximize MI among local pixels. MOCO [20, 21] extends the setup of InfoNCE further with the pretext contrast as a dictionary-lookup task. InfoNCE bound is combined with a momentum encoder which maximizes MI as a slow moving average of input and augmented samples. SimCLR [22] builds on the MOCO framework by maximizing the internal agreement between representations. While unsupervised representation learning methods adopt lower bounds on MI, they do so at the cost of large batch sizes. Since the InfoNCE bound is lose at small batch sizes, large architectures lean towards pretraining alternatives [23] rather than improving lower bounds. The work of [10] adapts InfoNCE bound based on parameteric and non-parameteric learning of visual instance discrimination. The multi-sample classification is casted to a binary discrimination setup, hence providing improved generalization and consistent performance. Based on this insight, we adopt the instance discrimination setup of [10] for our experiments.

3 Preliminaries

We review the information-theoretic setup for generalization and variational bounds. Let X and Y be a pair of random variables denoting the input and output data distributions $p(x)$ and $p(y)$ respectively. The mutual information $I(X;Y)$ between X and Y is a reparameterization-invariant measure of dependency consisting of the joint distribution $p(x,y)$and can be mathematically expressed as follows,

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

Equation 1 can be further simplified by exapnding the expectation,

$$I(X;Y) = \sum_{y} p(x|y)p(y) \log \frac{p(x|y)}{p(x)} = \mathbb{E}_{p(y)}[D_{KL}(p(x|y)||p(x))] \tag{2}$$

D_{KL} in Equation 2 denotes the Kullback-Liebler (KL) Divergence [24] which is a divergence metric. D_{KL} belongs to the general class of $\phi-$divergence metrics $D_\phi(P(x)||Q(y))$ which quantify the similarity between any two data distributions $P(x)$ and $Q(y)$. The general form of a ϕ-divergence, with ϕ being a convex and lower semi-continuous function such that $\phi(1) = 0$, is expressed in Equation 3. Utilizing $\phi(t) = t \log t$ in Equation 3 yields $D_\phi(P(x)||Q(y)) = D_{KL}(P(x)||Q(y))$.

$$D_\phi(P(x)||Q(y)) = \sum_{y} Q(y)\phi(\frac{P(x)}{Q(y)}) \tag{3}$$

Generalization bounds make use of random variables with a cumulant-generation [25] function $\psi(\lambda) = \log \mathbb{E}[e^{\lambda x}]$ such that $\lambda \geq 0$. A random variable is called σ-sub-Gaussian if the argument of the $\log$ cumulant-generation function satisfies $\mathbb{E}[e^{\lambda x}] \leq e^{\frac{\lambda^2}{\sigma}^2} 2$ for all $\lambda \in \mathbb{R}$ with σ^2 as the variance proxy or variance factor of the distribution.

4 When Do Bounds Hurt Learning?

The work of [9] throws light on the behavior of tractable distributions with high dimensional random variables. Based on the empirical characteristics of these bounds, one can identify the hindrances faced in generalization of the leanring algorithm (see Figure 1).

High Variance: Normalized upper and lower bounds aid in tractability of variational distributions when the data to be learned is long-tailed. However, these bounds demonstrate high variance as a result of large MI estimates. A suitable alternative to normalized bounds is to adopt the framework of structured bounds. These bounds leverage the structure of the problem and yield a known conditional distribution $p(y|x)$ which is tractable as per the problem setting. Structured bounds are conveniently applicable to representation learning [11, 9] but do not necessarily scale to high-dimensional scenarios. Another alternative which provisions a conitional tractable distribution are reparameterization bounds. These bounds make use of an additional functional, known as the *critic*, which converts lower bounds on MI into upper bounds on KL divergence. The critic functional need not explicitly learn the mapping between x and y. However, reparameterization is only made feasible if the conditional distribution $p(y|x)$ is tractable.

High Bias: Unnormalized upper and lower bounds demonstrate high bias and hurt tractability of complex distributions. Primary reasons for instability in bounds is lack of a partition functional which normalizes MI estimates. [9] argues that requirement of a partition function presents high bias as a result of exponential distributions which may not be tractable. However, the work does not provide empirical evidence on their tractability which leaves the suitability of a normalization constant an pen question. A suitable alternative to address biased estimates is the adoption of density ratios which train the critic functional using a divergence metric. The Jensen-Shanon Divergence (JSD) is one such scheme which yields a lower-biased estimate of optimal critic. While training critics is theoretically suitable, empirical evaluations [9] demonstrate unstable convergence of exponential gradients.

A Failure to Learn: Biased and noisy estimates are the key hindrances in learning tractable distributions. To that end, [9] aptly proposes a continuum of multi-sample interpolation bounds which trade-off bias with variance. A simpler form of critic when applied to non-linear interpolation in InfoNCE samples yields a continuum of lower bounds on MI. The new bound can be manually

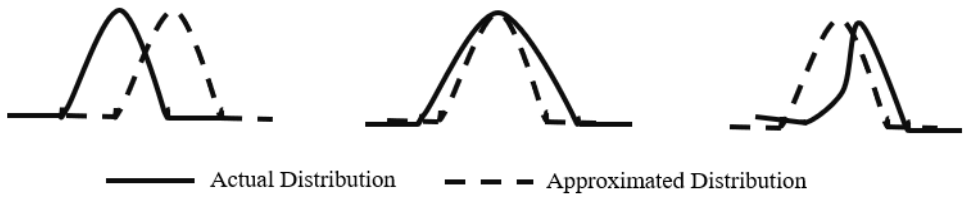

Figure 1: **Left** Conventional lose bounds suffer from high bias which hurts generalization of the learnt distribution, **Center** Learnt distribution is additionally hampered with noisy approximations, **Right** Biased estimates in conjunction with noisy dynamics hurt the completeness of learnt distribution.

tuned using γ which trades off bias with variance. Nonlinear interpolation bounds proposed in conjunction with MI saturate at $\log \frac{K}{\gamma}$ with K being the number of samples in the batch. Saturation of interpolation hurts the completeness of distribution and the bound fails to learn large MI estimates with inceasing batch sizes.

5 Variational Bounds for Generalization

This section provides insights into variational bounds as generelization measures of MI. Learning of variational bounds is discussed from the multi-sample and interpolation perspective. Nonlinear interpolation bounds give rise to the trade-off between bias and variance in estimates. Following generalization through this lens, we formulate an alternate approach with bias reduction as a contraction mapping. We extend our theoretical claims to previously discussed generalization bounds and discuss their formulations.

5.1 Learning Variational Bounds

InfoNCE: The InfoNCE objective is based on multi-sample unnormalized bounds. These bounds formulate multi-sample MI $I(X_1;Y)$ which is bounded by the optimal choice of critic $f(x,y)$. One such formulation is based on MINE [5] as presented in Equation 4.

$$I(X_1, Y) \geq 1 + \mathbb{E}_{p(x_{1:K})p(y|x_1)}[\log \frac{e^{f(x_1,y)}}{m(y;x_{1:K})}] - \mathbb{E}_{p(x_{1:K})p(y)}[\log \frac{e^{f(x_1,y)}}{m(y;x_{1:K})}] \tag{4}$$

Here $m(y;x_{1:K})$ is a Monte-Carlo estimate of the partition function $Z(y)$ and is mathematically expressed in Equation 5.

$$m(y;x_{1:K}) = \frac{1}{K}\sum_{k=1}^{K} e^{f(x_k,y)} \tag{5}$$

One can recover the InfoNCE bound (I_{NCE}) upon averaging over all K replicates in the last term of Equation 4 which yields 1. I_{NCE} can then be expressed as a lower bound on MI in Equation 6.

$$I(X;Y) \geq \mathbb{E}[\frac{1}{K}\sum_{k=1}^{K} \log \frac{e^{f(x_k,y_k)}}{\frac{1}{K}\sum_{j=1}^{K} e^{f(x_k,y_j)}}] \triangleq I_{NCE} \tag{6}$$

Nonlinear Interpolation: The multi-sample framework of MINE can be further extended using a simpler formulation. A nonlinear interpolation between MINE and I_{NCE} bridges the gap between low-bias and high-variance estimates of MINE with high-bias and low-variance estimates of I_{NCE}. The nonlinear interpolation bound (I_{IN}) is expressed in Equation 7.

$$I_{IN} \triangleq 1 + \mathbb{E}_{p(x_{1:K})p(y|x_1)}[\log \frac{e^{f(x_1,y)}}{\gamma m(y;x_{1:K}) + (1-\gamma)q(y)}] - \mathbb{E}_{p(x_{1:K})p(y)}[\log \frac{e^{f(x_1,y)}}{\gamma m(y;x_{1:K}) + (1-\gamma)q(y)}] \tag{7}$$

While I_{NCE} is upper bounded by $\log K$, I_{IN} is upper bounded by $\log \frac{K}{\gamma}$. The control in bias-variance trade-off improves accuracy of estimates. However, the significance of γ remains an open question in the case of higher-order divergence metrics and large value of MI in practical settings.

ϕ**-divergence**: Generalized divergence metrics facilitate tighter bounds by utilizing α-MI as the dependence measure. [6] presents a tight bound which is based on random variables with cumulant-generation functions. If $X_i - Y_i$ has a cumulant generation function $\leq \psi_i(\lambda)$ over domain $[0, b_i)$ where $0 \leq b_i \leq \infty$ and $\psi_i(\lambda)$ is convex and i denotes the iterates of the variables X and Y, one can define the expected cumulant-generation function $\bar{\psi}_i(\lambda)$ as in Equation 8 to obtain the bound expressed in Equation 9. Here, $\bar{\psi}^{*-1}$ denotes the inverse of the convex conjugate of $\bar{\psi}$.

$$\bar{\psi}(\lambda) = \mathbb{E}_i[\psi_i(\lambda)], \ \lambda \in [0, \min_i b_i) \tag{8}$$

$$\mathbb{E}[X_i - Y_i] \leq \bar{\psi}^{*-1}(I(X;Y)) \tag{9}$$

The bound of [6] generalizes the work of [1] as it is applicable to long-tailed distributions and variables which may not necessarily obey the sub-Gaussianity assumption. Based on Equation 9, [6] formulates the α-MI bound which improves the bound presented in [1]. Suppose $||X_i - Y_i||_\beta \leq \sigma_i$ where $1 \leq \beta \leq \infty$, if $\alpha \triangleq$ conjugate of β such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, then the improved α-MI bound can be expressed as in Equation 10.

$$|\mathbb{E}[X_i - Y_i]| \leq ||\sigma||_\beta I_\alpha(X;Y)^\alpha \tag{10}$$

While the bound in Equation 10 is generalizable to variables with no moment-generating functions, its tightness remains an open question. [6] prove the tightness of Equation 10 using extreme value theory. The bound is tight for $n^{\frac{1}{\beta}}$ with n being the number of data samples. However, tightness holds under the condition that β is bounded such that $2 \leq \beta \leq \infty$. For large values of β, $n^{\frac{1}{\beta}}$ tends to diminish which renders the estimation of $I_\alpha(X;Y)$ intractable. Moreover, ϕ-divergences are originally defined as power functions over β while the bound makes use of an affine transformation. The alternate formulation may not hold in the more generalized-case. This poses a hindrance for α-MI bounds to be used as substitutes to pre-existing methods.

5.2 Improving Bounds on MI

The bound expressed in Equation 9 can be improved by using the multi-sample InfoNCE bound. This arises as a direct consequence of the fact that $I(X;Y) \geq I_{NCE}$. We formalize this finding in Proposition 1 and defer all proofs to Appendix.

Proposition 1. *If $X_i - Y_i$ has the expected cumulant-generation function $\bar{\psi}_i(\lambda)$ and the inverse of the convex conjugate of $\bar{\psi}$ denoted by $\bar{\psi}^{*-1}$ exists, then the improved bound utilizing I_{NCE} can be expressed as follows,*

$$\mathbb{E}[X_i - Y_i] \leq \bar{\psi}^{*-1}(I_{NCE}) \tag{11}$$

While Proposition 1 presents an improvement over the original bound, the expression makes use of I_{NCE} which demonstrates high bias and saturation at $logK$. This motivates the need for a bound which can retain the tightness of I_{NCE} and at the same time trade off bias with variance.

To address the improvement of generlization bound, we turn our attention to non-linear interpolation bound I_{IN} introduced in the previous section. Interpolation aids in the trade off between bias and variance and improves the bound's stability at larger batch sizes. More specifically, I_{IN} saturates at $\log \frac{K}{\gamma}$. However, the tightness of I_{IN} with respect to I_{NCE} cannot be validated as a result of nonlinear iterates during interpolation. For $\gamma \neq 1$, $I_{IN} \neq I_{NCE}$ which poses a hindrance in the comparison of I_{IN} to I_{NCE}. However, a lower bound on I_{IN} can yield a lower bound on I_{NCE} with high probability since $I_{IN} = I_{NCE}$ at $\gamma = 1$. Thus, obtaining a tractable lower bound on I_{IN} would further improve generlization and strengthen the claim of Proposition 1.

One can leverage contraction theory [26, 27] to highlight mathematical properties which would aid in extracting a tractable bound. Such a mapping guarantees convergence in asymptotically-stable nonlinear systems [27] and provides a dynamical framework for assessing stability of bounds [28]. A contraction mapping between any two functions $f_1(w)$ and $f_2(w)$ implies that the norm distance between $f_1(w)$ and $f_2(w)$ decays at a constant (in some cases geometric [26]) rate. Given a contraction operator τ when iteratively applied on $f_1(w)$ and $f_2(w)$, the mapping $\tau f_1(w) - \tau f_2(w)$ is a contraction if the inequality in Equation 12 is satisfied.

$$\tau f_1(w) - \tau f_2(w) \leq ||f_1(w) - f_2(w)||, \; \forall w \tag{12}$$

Equation 12 is a generalization of the fixed-point theory in Banach metric spaces [27] which provides suitable conditions for assessing stability of nonlinear systems. The key component of evaluating a nonlinear system is motivated by its convergence towards a fixed point in the Banach metric space. Convergence towards a fixed point indicates stability of the overall mapping. We borrow from this insight in order to form a contraction on nonlinear interpolation bounds which can be interpreted as a continuum in a nonlinear space. Upon realizing input samples as points in this continuous nonlinear space, a simple yet elegant formulation of a contraction mapping can be achieved. To utilize a contraction mapping on I_{IN}, we seek a contraction operator $\mathcal{T}$ which is tractable. A suitable choice is the alternate Boltzmann (mellowmax) operator introduced in [29]. The Mellowmax operator $\mathcal{T}f(w) = \log \sum_w \exp f(w)$ which may be interpreted as an energy-based function. Retaining properties of the Boltzmann distribution, mellowmax is an asymptotically stable formulation of the Gibbs distribution. Mellowmax has been suitably adopted in control and learning settings [28, 30] wherein the probability distribution forms a continuum over the input space. Additionally, the exponent in $\mathcal{T}$ is tractable as it is followed by the log which prevents the arguments from exploding. Proposition 2 depicts the contraction property of $\mathcal{T}$.

Proposition 2. *Given a function $f(w) : w \to \mathbb{R}$, the operator $\mathcal{T}f(w) = \log \sum_w \exp f(w)$ forms a contraction on $f(w)$.*

We leverage the result of Proposition 2 to obtain a tractable lower bound on I_{IN}. This requires summing over interpolating samples $1 : K$ of the distribution. The novel bound I_N obtained by operating $\mathcal{T}$ on interpolation terms is expressed in Proposition 3. Note that the sum in the first interpolation term requires computation of the inner expectation which now depends on the distribution conditioned on each iterate. This facilitates tractability since the estimates obtained as a result of individual conditioning would be accurate in comparison to estimates based on a single sample.

Proposition 3. *If the operator $\mathcal{T}f(w) = \log \sum_w \exp f(w)$ is a contraction mapping $\forall w$, then the nonlinear interpolation bound I_{IN} can be further simplified as expressed in Equation 13.*

$$\begin{aligned} I_{IN} \geq I_N \triangleq 1 + \log \sum_{i=1}^{K} \exp \mathbb{E}_{p(x_{1:K})p(y|x_i)}[\log \frac{e^{f(x_i,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}] \\ \log \sum_{i=1}^{K} \exp \mathbb{E}_{p(x_{1:K})p(y)}[\log \frac{e^{f(x_i,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}] \end{aligned} \tag{13}$$

Equation 13 is a lower bound on I_{IN} which indicates that the novel bound I_N is also a lower bound on I_{NCE} with high probability since $I_{IN} = I_{NCE}$ at $\gamma = 1$. Thus, one can safely make the following claim,

$$I(X;Y) \geq I_{NCE} \geq I_N \tag{14}$$

Equation 14 can be further used to validate the suitability of the novel bound I_N as a replacement to I_{NCE} in Proposition 1. This leads us to formulate the new generalization bound presented in 4.

Proposition 4. *Since $I_{NCE} \geq I_N$ with high probability, the generalization bound in Proposition 1 can be simplified as follows,*

$$\mathbb{E}[X_i - Y_i] \leq \bar{\psi}^{*-1}(I_N) \leq \bar{\psi}^{*-1}(I_{NCE}) \tag{15}$$

Proposition 3 and Proposition 4 obtained as a result of Equation 12 highlight the main finding of our work. Theoretically, I_N retains the properties of I_{IN} and would depict convergence analogous to I_{IN} in settings with large batch sizes. Additionally, the bias-variance trade off arising as a virtue of nonlinear interpolation presents I_N as a suitable replacement to I_{NCE} in Proposition 1. On the other hand, an empirical comparison of stability between I_{NCE} and I_N may be difficult observe. This arises as a direct consequence of interpolation combined with dimensional contraction which may lead I_N to collapse for some iterates. Although theoretical in nature, our findings motivate empirical validation and application of the proposed bound.

6 Experiments

6.1 Setup

6.2 Unsupervised Instant Discrimination

7 Conclusion

References

[1] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240. PMLR, 2016.

[2] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. NIPS'17, 2017.

[3] Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In *Advances in Neural Information Processing Systems*, 2019.

[4] Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 2020.

[5] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: Mutual information neural estimation, 2018.

[6] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1475–1479. IEEE, 2017.

[7] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the royal statistical society series b-methodological*, 28:131–142, 1966.

[8] Jiantao Jiao, Thomas A. Courtade, Albert No, Kartik Venkat, and Tsachy Weissman. Information measures: The curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60(12):7616–7626, Dec 2014.

[9] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

[10] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[11] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[12] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2020.

[13] 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, 2019.

[14] MD Donsker and SRS Varadhan. Large deviations for markov processes and the asymptotic evaluation of certain markov process expectations for large times. In *Probabilistic Methods in Differential Equations*, pages 82–88. Springer, 1975.

[15] Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage, 2019.

[16] Ilja Kuzborskij, Nicolò Cesa-Bianchi, and Csaba Szepesvári. Distribution-dependent analysis of gibbs-erm principle. *arXiv preprint arXiv:1902.01846*, 2019.

[17] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[18] Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms, 2020.

[19] Thomas Steinke and Lydia Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. Proceedings of Machine Learning Research, pages 3437–3452. PMLR, 2020.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[21] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[22] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[23] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020.

[24] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[25] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[26] Krzysztof Ciesielski et al. On stefan banach and some of his results. *Banach Journal of Mathematical Analysis*, 1(1), 2007.

[27] Stefan Banach. On operations in abstract sets and their application to integral equations. 1922.

[28] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.

[29] Kavosh Asadi and Michael L Littman. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, 2017.

[30] Karush Suri, Xiao Qi Shi, Konstantinos Plataniotis, and Yuri Lawryshyn. Energy-based surprise minimization for multi-agent value factorization, 2020.

A Proofs

Proposition 1. *If $X_i - Y_i$ has the expected cumulant-generation function $\bar{\psi}_i(\lambda)$ and the inverse of the convex conjugate of $\bar{\psi}$ denoted by $\bar{\psi}^{*-1}$ exists, then the improved bound utilizing I_{NCE} can be expressed as follows,*

$$\mathbb{E}[X_i - Y_i] \leq \bar{\psi}^{*-1}(I_{NCE})$$

Proof. The proof is straightforward and based on original generalization bound in **??**. The result is obtained by following the steps of Theorem 2 [6] and replacing $I(X_i; Y_i)$ with I_{NCE} in the last step as follows,

$$\begin{aligned}
\mathbb{E}[X_i - Y_i] &\leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I(X_i; Y_i)}{\lambda} \\
\mathbb{E}[X_i - Y_i] &\leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I_{NCE}}{\lambda}; \quad since I_{NCE} \leq I(X_i; Y_i) \\
&= \bar{\psi}^{*-1}(I_{NCE})
\end{aligned}$$

This completes the proof. □

Proposition 2. *Given a function $f(w) : w \to \mathbb{R}$, the operator $\mathcal{T} f(w) = \log \sum_w \exp f(w)$ forms a contraction on $f(w)$.*

Proof. Let us first define a norm on the functions $||f_1(w) - f_2(w)|| \equiv \underset{w}{max}|f_1(w) - f_2(w)|$. Suppose $\epsilon = ||f_1(w) - f_2(w)||$,

$$\begin{aligned}
\log \sum_w \exp\left(f_1(w)\right) &\leq \log \sum_w \exp\left(f_2(w) + \epsilon\right) \\
= \log \sum_w \exp\left(f_1(w)\right) &\leq \log \exp\left(\epsilon\right) \sum_w \exp\left(f_2(w)\right) \\
= \log \sum_{w=1} \exp\left(f_1(w)\right) &\leq \epsilon + \log \sum_w \exp\left(f_2(w)\right) \\
= \log \sum_w \exp\left(f_1(w)\right) - \log \sum_w \exp\left(f_2(w)\right) &\leq ||f_1(w) - f_2(w)|| \qquad (16)
\end{aligned}$$

Similarly, using ϵ with $\log \sum_w \exp\left(f_1(w)\right)$,

$$\begin{aligned}
\log \sum_w \exp\left(f_1(w) + \epsilon\right) &\geq \log \sum_w \exp\left(f_2(w)\right) \\
= \log \exp\left(\epsilon\right) \sum_w \exp\left(f_1(w)\right) &\geq \log \sum_w \exp\left(f_2(w)\right) \\
= \epsilon + \log \sum_w \exp\left(f_1(w)\right) &\geq \log \sum_w \exp\left(f_2(w)\right) \\
= ||f_1(w) - f_2(w)|| &\geq \log \sum_w \exp\left(f_2(w)\right) - \log \sum_w \exp\left(f_1(w)\right) \qquad (17)
\end{aligned}$$

Results in Equation 16 and Equation 17 prove that $\mathcal{T}$ is a contraction. □

Proposition 3. *If the operator $\mathcal{T} f(w) = \log \sum_w \exp f(w)$ is a contraction mapping $\forall w$, then the nonlinear interpolation bound I_{IN} can be further simplified as expressed in Equation 13.*

$$\begin{aligned}
I_{IN} \geq I_N \triangleq 1 + \log \sum_{i=1}^{K} \exp \mathbb{E}_{p(x_{1:K})p(y|x_i)}[\log \frac{e^{f(x_i,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}] \\
\log \sum_{i=1}^{K} \exp \mathbb{E}_{p(x_{1:K})p(y)}[\log \frac{e^{f(x_i,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}]
\end{aligned}$$

Proof. From Equation 7, we have the following,

$$\begin{aligned}
I_{IN} \triangleq 1 + \mathbb{E}_{p(x_{1:K})p(y|x_1)}[\log \frac{e^{f(x_1,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}] \\
- \mathbb{E}_{p(x_{1:K})p(y)}[\log \frac{e^{f(x_1,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}]
\end{aligned}$$

For notational convenience, we relabel the two interpolation terms as follows,

$$\mathcal{P}(x) = \mathbb{E}_{p(x_{1:K})p(y|x_1)}[\log \frac{e^{f(x_1,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}]$$
$$\mathcal{Q}(x) = \mathbb{E}_{p(x_{1:K})p(y)}[\log \frac{e^{f(x_1,y)}}{\gamma m(y; x_{1:K}) + (1-\gamma)q(y)}]$$

Using the result from Proposition 2 and applying $\mathcal{T}$ to the two interpolation terms in the expressions gives us the following,

$$\mathcal{T}\mathcal{P}(x) - \mathcal{T}\mathcal{Q}(x) \leq \max_x |\mathcal{P}(x) - \mathcal{Q}(x)|$$
$$\mathcal{T}\mathcal{P}(x) - \mathcal{T}\mathcal{Q}(x) \leq \mathcal{P}(x) - \mathcal{Q}(x)$$

Using the above result in I_{IN} yields the lower bound I_N and completes the proof. □

Proposition 4. *Since $I_{NCE} \geq I_N$ with high probability, the generalization bound in Proposition 1 can be simplified as follows,*

$$\mathbb{E}[X_i - Y_i] \leq \bar{\psi}^{*-1}(I_N) \leq \bar{\psi}^{*-1}(I_{NCE})$$

Proof. We follow the steps of Proposition 1 to obtain the lower bound with I_{NCE} and replace it with I_N as follows,

$$\mathbb{E}[X_i - Y_i] \leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I(X_i; Y_i)}{\lambda}$$
$$\mathbb{E}[X_i - Y_i] \leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I_{NCE}}{\lambda}; \quad since I_{NCE} \leq I(X_i; Y_i)$$
$$\mathbb{E}[X_i - Y_i] \leq \inf_{\lambda \in [0, \min_i b_i)} \frac{\bar{\psi}(\lambda) + I_N}{\lambda}; \quad since I_N \leq I_{NCE}$$
$$= \bar{\psi}^{*-1}(I_N)$$

This completes the proof.

□

B Additional Results

C Implementation Details

C.1 Note on Experiment Setup

C.2 Hyperparameters