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

Preprint.

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:

3 Preliminaries

4 When Do Bounds Hurt Learning?

High Variance:

High Bias:

A Failure to Learn:

5 Variational Bounds for Generalization

5.1 Learning Variational Bounds

5.2 Bias Reduction as a Contraction

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.