Neural Estimation of Information-Theoretic Generalization Bounds: Limitations and Guidelines

Nathália Barros Viana, Eduardo Nunes Velloso, Max Henrique Machado Costa, and José Cândido Silveira Santos Filho

***Abstract*— We investigate practical challenges of estimating information-theoretic generalization bounds using neural mutual information estimators. Focusing on a Gaussian mean estimation task, we compare input–output mutual information (MI), individual-sample mutual information (ISMI), and conditional mutual information (CMI) formulations under varying sample sizes and regularization strategies. Through empirical analysis, we identify underfitting regimes, characterize the bias–variance behavior across estimators, and highlight sample complexity ceilings that limit estimation accuracy. Our results provide practical guidelines for selecting estimators and tuning Monte Carlo parameters to achieve reliable generalization bounds in low-data settings.**

***Keywords*— mutual information estimation, generalization error bounds, neural estimation, bias–variance decomposition.**

I. Introduction

A major concern when analyzing algorithms for supervised learning problems is their generalization ability—that is, the model's capacity to transfer learned patterns to unseen data. Computing the expected generalization error exactly is infeasible, but it can be upper-bounded if the learning task satisfies certain conditions.

These conditions are typically divided into (i) complexity measures (e.g., Vapnik–Chervonenkis dimension and Rademacher complexity) and (ii) algorithmic stability guarantees. The former were foundational for the early studies of generalization, but scale poorly with increased model complexity and do not account for the internal dynamics of the learning process. For instance, they overlook implicit regularization effects introduced by optimization algorithms such as stochastic gradient descent (SGD). Although the latter focuses on such algorithmic properties, it generally neglects the data distribution, even though generalization is strongly influenced by that.

To better characterize the intrinsic nature of learning tasks, recent works have proposed information-theoretic conditions for generalization [1], [2]. Interpreting the learning process as a probabilistic communication channel where the input is a randomly sampled dataset and the output is the corresponding learned model, a bound on the input–output mutual information implies generalization guarantees, which depend on both the data distribution and the learning algorithm. Numerous variations of this concept have been developed in the form of tighter generalization bounds [3]–[6], usually related to two main aspects of the ordinary input–output dependence: pointwise stability (capturing how much the model changes with slight changes of individual data samples) and differential privacy (capturing the knowledge of which samples were used between a number of possible options), inspiring the design of algorithms with improved generalization. Other directions recently explored in the literature include the geometry of the hypothesis space [7], the optimal choice of a divergence metric [8], and the learning dynamics of iterative noisy algorithms [9].

In practice, the analytical computation of this mutual information is rarely tractable for real supervised learning problems. This means that, while the theoretical formulation of these bounds has been widely studied, their empirical behavior in practice remains relatively underexplored.

Historically, mutual information estimation for general problems has been addressed through either nonparametric approaches [10] or parametric neural networks like the Mutual Information Neural Estimator (MINE, [11]). The former are known not to scale well with sample size or dimension [12], which are of critical importance for most realistic learning tasks. The latter are scalable and flexible, but introduce a training procedure with many distinct hyperparameter knobs that influence the bias and variance characteristics of the final estimator.

In this paper, considering a simple toy learning problem — Gaussian mean estimation — we investigate the practical effects of using neural estimators of mutual information on the estimated generalization bounds for the mutual information (MI, [2]), individual-sample mutual information (ISMI, [3]), and conditional mutual information (CMI, [4]) approaches. We evaluate the convergence, bias–variance, and complexity characteristics of MINE and of some of its regularized variants under different learning algorithms. Our main contributions are guidelines on the use of MINE within the context of statistical learning problems, highlighting failure cases to avoid and the practical drawbacks of using theoretically tighter bounds.

II. Background

There exists some underlying data-generating distribution μ over a data space $\mathcal{Z} \subset \mathbb{R}^d$, from which N examples are sampled independently and identically distributed (i.i.d.) to form the training dataset $S = (Z_1, \ldots, Z_N)$. A learning algorithm is represented by a conditional distribution $P_{W|S}$, which samples a model W from the hypothesis space $\mathcal{W}$ based

N. B. Viana, E. N. Velloso, M. H. M. Costa, and J. C. S. Santos Filho are with the Department of Communications, School of Electrical and Computer Engineering, Universidade Estadual de Campinas (UNICAMP), Campinas, SP 13083-852, Brazil (e-mail: bviananathalia@gmail.com; e290885@dac.unicamp.br; max@fee.unicamp.br; jcssf@unicamp.br).

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

on the training dataset. A loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ defines the performance metric of the model at any point in the data space, and is used to determine the empirical and true risks, respectively:

$$\mathcal{L}_S(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(w, Z_i) \tag{1}$$

$$\mathcal{L}_\mu(w) = \mathbb{E}_Z \ell(w, z). \tag{2}$$

It is also helpful to fix the universe of possible data points by considering the existence of a larger supersample $S^{\pm} = (Z_1^+, Z_1^-, \ldots, Z_N^+, Z_N^-)$ of size $2N$, which is partitioned into the actual training sample S and a ghost sample $\bar{S}$ by a random binary selection vector $R \in \{+1, -1\}^N$, such that $S = (Z_1^{R_1}, \ldots, Z_N^{R_N})$.

A. Information-Theoretic Bounds

Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any integrable random function of the random variables X and Y. Specific choices of the variables X, Y and of the function f—usually the loss or the empirical risk—define different upper bounds on the expected generalization error $\text{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}_{W,S}(\mathcal{L}_\mu(w) - \mathcal{L}_S(w))$.

The Donsker–Varadhan (DV) variational representation of the Kullback-Leibler (KL) divergence [13] establishes an inequality of the form

$$I(X;Y) = D_{\text{KL}}(P_{X,Y} || P_X \otimes P_Y) \tag{3}$$

$$\geq \mathbb{E}_{X,Y} \lambda f - \log \mathbb{E}_X \mathbb{E}_Y \exp(\lambda f), \quad \forall \lambda \in \mathbb{R}, \tag{4}$$

where $P_X \otimes P_Y$ is the product of the marginal distributions and λ is a tunable parameter used to tighten the bound. Given the moment-generating function (MGF) of a centered random variable $\Omega_f(\lambda) = \mathbb{E}_X \mathbb{E}_Y \exp(\lambda(f - \mathbb{E}_X \mathbb{E}_Y f))$, if there exists a convex function $\psi : \mathbb{R} \to \mathbb{R}$ such that

$$\exp(\psi(\lambda)) \geq \Omega_f(\lambda), \tag{5}$$

then it can be shown that

$$\mathbb{E}_{X,Y} f - \mathbb{E}_X \mathbb{E}_Y f \leq \inf_\lambda \frac{I(X;Y) + \psi(\lambda)}{\lambda}. \tag{6}$$

The right-hand side of eq. (6) has the same form as the inverse of the convex conjugate of the function ψ, $\psi^{*-1}(y) \triangleq \inf_t (y + \psi(t))/t$. Besides the original input–output mutual information, we focus on the two main variations that better illustrate the concepts used for finding tighter bounds: pointwise stability and differential privacy.

1) Input–Output Mutual Information [2]:

$$\text{MI} = \psi^{*-1}\left(\frac{I(W;S)}{N}\right). \tag{7}$$

The original MI bound offers an interpretable connection between the independence of the learned hypothesis on the specific sampling of the data and generalization ability. In fact, it is possible to show that bounded mutual information implies that the algorithm satisfies an on-average stability condition. However, the MI bound suffers from practical limitations: it may become vacuous or infinite when the learning algorithm is highly sensitive to the input data, especially in settings involving continuous distributions or deterministic learners.

2) Individual Sample Mutual Information [3]:

$$\text{ISMI} = \frac{1}{N} \sum_{i=1}^{N} \psi^{*-1}(I(W;Z_i)). \tag{8}$$

The ISMI bound was proposed to overcome the shortcomings of the MI bound by decomposing the overall mutual information into contributions from each individual data point. Whenever it is bounded, it has a slightly stricter pointwise stability guarantee.

Considering the independence of each sampled Z_i, we can write $I(W; Z_i | Z_1, \ldots, Z_{i-1}) = I(W, Z_1, \ldots, Z_{i-1}; Z_i) = I(W; Z_i) + I(Z_1, \ldots, Z_{i-1}; Z_i | W) \geq I(W; Z_i)$, and thus we have, by the chain rule,

$$I(W;S) = \sum_{i=1}^{N} I(W; Z_i | Z_1, \ldots, Z_{i-1}) \geq \sum_{i=1}^{N} I(W; Z_i), \tag{9}$$

and then, because ψ^{*-1} is concave and nondecreasing, using Jensen's inequality, ISMI $\leq$ MI. In other words, the ISMI bound is always at least as tight as the MI bound.

3) Conditional Mutual Information [4]:

$$\text{CMI} = \psi^{*-1}\left(\frac{I(W;R|S^{\pm})}{N}\right). \tag{10}$$

The CMI bound addresses how much information the learned hypothesis reveals about the specific selection of training data, given a fixed supersample $S^{\pm}$ that restricts the possible sample choices. It is closely related not only with algorithmic stability but also with differential privacy.

It is possible to show that $I(W;R|S^{\pm}) = I(W;S|S^{\pm}) = I(W;S) - I(W;S^{\pm}) \leq I(W;S)$, which naturally means that CMI $\leq$ MI as well.

B. Mutual Information Estimation

Estimating mutual information from finite samples is generally challenging. A popular method uses k-nearest neighbor (kNN) distances in probability space to estimate the entropy of the underlying distributions [10]. While it is simple and effective for low-dimensional settings, it scales poorly with dimensionality, which led to the construction of estimators based on variational bounds of mutual information.

Based on the DV representation of the KL divergence [13], it has been shown that neural networks can be trained to approximate the optimal function that maximizes

$$I(X;Y) = \sup_T \mathbb{E}_{X,Y} T - \log \mathbb{E}_X \mathbb{E}_Y \exp(T), \tag{11}$$

that is, the log-odds $T^* = \log(P_{X,Y}/P_X \otimes P_Y) + C$. The MINE estimator [11] fixes a neural network architecture, introducing approximation error, and draws on a finite sample of the joint distribution and the product of the marginals, introducing estimation error. The MINE estimate $\hat{I}$ can be given by the following optimization problem:

$$\hat{I} = \sup_{\theta \in \Theta} \frac{1}{M} \sum_{m=1}^{M} T_\theta^{(m)} - \log \frac{1}{M} \sum_{m=1}^{M} \exp(\tilde{T}_\theta^{(m)}), \tag{12}$$

where $T_\theta^{(m)}$ is the statistics network parametrized by θ and evaluated on the joint distribution (correctly matched pairs

(x, y)), and $\tilde{T}_\theta^{(m)}$ is the same but evaluated on the product of the marginals (shuffled pairs $(\tilde{x}, \tilde{y})$, drawn independently).

Although MINE is a provably consistent estimator, it has known limitations regarding variance, its need for a large number M of samples from the distribution, and training difficulties. Some of these issues have been addressed in the form of small variations around its original concept. The Data-Efficient MINE (DEMINE, [14]) introduces the idea of using a separate validation set over which the estimation is computed, with the goal of avoiding overfitting. The Smoothed MI Lower-bound Estimator (SMILE, [15]) clips the log-partition term $\exp(\tilde{T})$ based on a parameter τ to control the variance caused by outliers. The Regularized MINE (ReMINE, [16]) controls the drift of the constant offset in the log-odds function by adding an L2 regularization term with strength λ. Alternatively, the Discriminative Estimator of MI (DEMI) reframes estimation as a binary classification problem between joint and marginal samples that directly approximates T^*, bypassing the log-partition altogether. Each of these methods trades off bias, variance, and training cost in different ways, offering practical tools for improving MI estimation in finite-data settings.

III. Experimental Framework

A. Benchmark Task: Gaussian Mean Estimation

To obtain tractable ground truths for assessing the estimators, we focus on the well-behaved task of learning an unknown mean parameter from a finite collection of N i.i.d. Gaussian samples $Z \sim \mathcal{N}(w^*, \sigma^2)$. We consider the mean squared error loss, $\ell(w, z) = (w - z)^2$.

The classical estimator of the mean of a Gaussian distribution from a finite number of samples is the sample average, which corresponds to the maximum likelihood estimator. To introduce randomness, we added a small Gaussian noise:

$$W = \frac{1}{N}\sum_{i=1}^{N} Z_i + \varepsilon, \tag{13}$$

with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. This means that $P_{W|S}$ is a Gaussian distribution with variance σ_ε^2 centered on the sample average. This setup reflects the common use of intentional noise in algorithms, either as a regularization mechanism (e.g., dropout) or as part of the optimization process (e.g., SGD).

Under the Gaussian data distribution, estimation noise affects both true and empirical risks equally, leading to cancellation, and the expected generalization error can be shown to be

$$\mathrm{gen}(\mu, P_{W|S}) = \frac{2\sigma^2}{N}, \tag{14}$$

and, taking the loss variance $\sigma_\ell^2 = (1 + 1/N)\sigma^2 + \sigma_\varepsilon^2$, the MGF of the centered loss is simply that of a $\sigma_\ell^2 \chi_1^2$ random variable, such that

$$\log \Omega_\ell(\lambda) = -\frac{1}{2}\log\left(1 - 2\lambda\sigma_\ell^2\right) - \lambda\sigma_\ell^2, \tag{15}$$

which is upper bounded by $\psi(\lambda) = \sigma_\ell^4\lambda^2$ for $\lambda < 0$. Therefore, the mutual information bounds can be found from the inverse convex conjugate $\psi^{*-1}(y) = 2\sigma_\ell^2\sqrt{y}$.

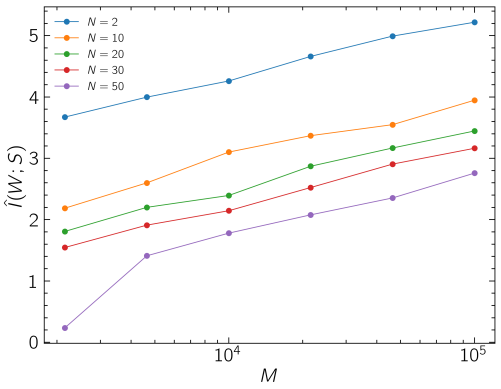


Fig. 1: Estimated MI vs. number of Monte Carlo samples M, for $\sigma_\varepsilon = 0$ (deterministic learner) and different training set sizes N.

For these distributions, analytical expressions for the relevant mutual information terms are available:

$$I(W;S) = \frac{1}{2}\log\left(1 + \frac{\sigma^2}{N\sigma_\varepsilon^2}\right) \tag{16}$$

$$I(W;Z_i) = \frac{1}{2}\log\left(\frac{\sigma^2/N + \sigma_\varepsilon^2}{(N-1)\sigma^2/N^2 + \sigma_\varepsilon^2}\right) \tag{17}$$

$$I(W;R|S^{\pm}) = h(W|S^{\pm}) - \frac{1}{2}\log\left(2\pi e\sigma_\varepsilon^2\right). \tag{18}$$

In the special case where $\sigma_\varepsilon^2 = 0$, the learner becomes a deterministic function of the training set S, with $P_{W|S}$ becoming a singular delta function, so that the mutual information terms reduce to the following (see [9]):

$$I(W;S) = \infty \tag{19}$$

$$I(W;Z_i) = \frac{1}{2}\log\left(1 + \frac{1}{N-1}\right) \tag{20}$$

$$I(W;R|S^{\pm}) = N\log 2. \tag{21}$$

B. Neural Estimator Implementation

To produce estimates using MINE, we simulated M Monte Carlo realizations of datasets S, sampling a corresponding value of W from $P_{W|S}$ each time, and we separate half of them into a holdout validation set. We implement MINE's statistics network with three fully connected layers with 100 units each, using the exponential linear unit (ELU) activation function to avoid dead units. We used the Adam optimizer for training with a batch size chosen as 10% of the total value of M, and an early stopping strategy triggered after 10 epochs without improvement.

Since the learning algorithm analyzed is not sensitive to data point ordering, all N values of $I(W;Z_i)$ are identical by symmetry. This significantly reduces the cost of estimating the ISMI bound, as it suffices to use only a single point for the estimation.

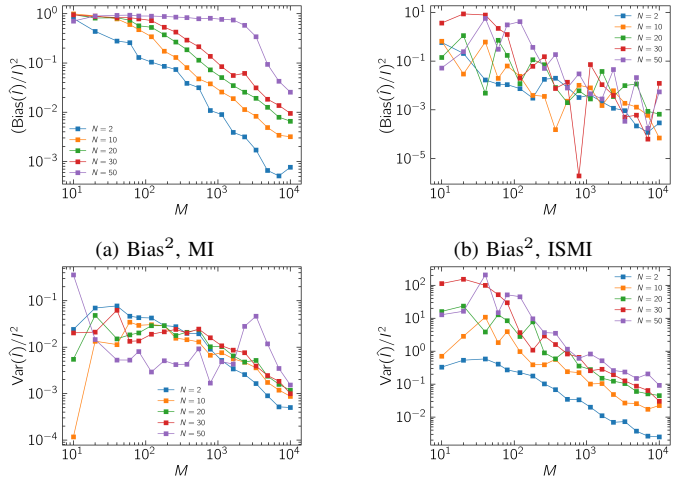
IV. Empirical Analysis and Discussion

A. Sample Complexity Ceiling

In the degenerate deterministic scenario with $\sigma_\varepsilon = 0$, although the true input–output MI is infinite, any approximation determined by an estimator from a finite number of

samples will be finite. Figure 1 indeed shows that the finite MINE estimate of $I(W;S) = \infty$ grows logarithmically as M increases. A similar result was argued by [17] in the context of an absolute ceiling of order $\mathcal{O}(\log M)$ for MI estimated with high confidence by any lower-bound distribution-free estimator. The actual ceiling may not be attained by any specific statistics network implementation, but the behavior observed for this infinite MI scenario empirically supports a not yet demonstrated logarithmic growth law of the saturation value for the DV family of lower-bound estimators [18]. We also observe that as the dataset size N increases, the relationship between W and S is harder to detect, implying an offset in M.

B. Bias and Variance



(a) Bias^2, MI (b) Bias^2, ISMI

(c) Variance, MI (d) Variance, ISMI

Fig. 2: Normalized bias and variance of estimated MI and ISMI vs. number of Monte Carlo samples M, for $\sigma_\varepsilon = 0.1$ (noisy learner) and different training set sizes N.

For the more general $\sigma_\varepsilon > 0$ case, we investigate the bias–variance decomposition of the estimation error:

$$\mathbb{E}(\hat{I} - I)^2 = (\mathbb{E}\hat{I} - I)^2 + \mathrm{Var}(\hat{I}), \tag{22}$$

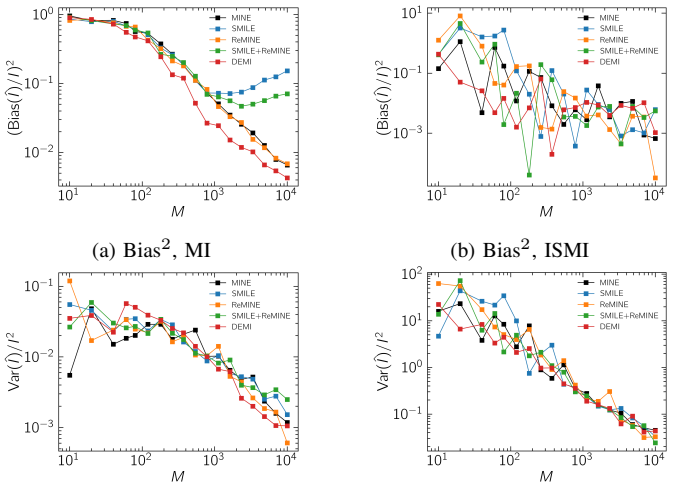
where we approximate the expectations and variance using a finite number $K = 50$ of realizations of the estimation process.

The resulting curves are shown in Figure 2 as functions of M for different values of N, for both the ordinary MI and the ISMI. The curves are normalized by the analytical value of I squared, as shown in eqs. (16) and (17). The main observation from these plots is the existence of a well-defined region of low number of samples M where the statistics network underfits. In fact, as can be seen by the normalized bias in this region approaching unity, all estimates stay close to zero within this region. This likely occurs because the empirical joint distribution becomes very similar to the product of marginals when the number of samples is small. Another reason is the fact that, as the number of samples is reduced, the loss landscape becomes dominated by noise, and trivial constant solutions tend to be favored to minimize risk.

In addition to that, we notice that this critical threshold of M scales with the input dimension N. Once M exceeds that threshold—say, $M_{\mathrm{crit}}(N)$—the statistical bias due to finite sampling becomes dominant and the squared bias term decays proportionally to $1/M$. This scaling is consistent with the finite-sample bias analysis of [11].

The equivalent change in behavior in the variance curves can be interpreted as a transition between a state with a high probability of underfitting when $M \ll M_{\mathrm{crit}}(N)$ to a regular $1/M$ scaling region when $M \gg M_{\mathrm{crit}}(N)$. In the transition region, some realizations of the randomly drawn input will underfit the statistics network with varying probabilities, and the resulting spread of estimates between zero and a tight lower bound leads to increased variance.

When contrasting the curves for MI and ISMI, we draw attention to two observations: firstly, it is significantly easier to avoid underfitting for the ISMI estimates, given the lower (constant with N) dimensionality of the required inputs; secondly, since the true MI values are much lower for ISMI and the estimator variance is expected to scale with $\exp(I)$, the scaling in the operational region seems to quickly reach a plateau due only to the capacity-bounded approximation error. The fluctuations in bias for the ISMI estimator possibly originate from the high sensitivity to the choice of individual data points.



(a) Bias^2, MI (b) Bias^2, ISMI

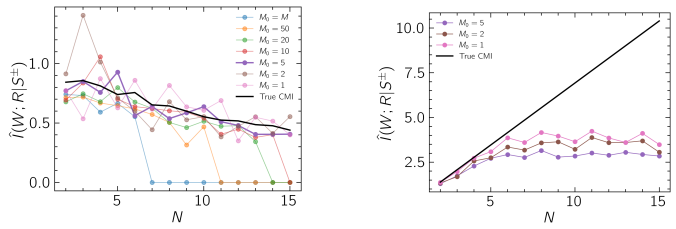
(c) Variance, MI (d) Variance, ISMI

Fig. 3: Normalized bias and variance of estimated MI and ISMI vs. number of Monte Carlo samples M, for $\sigma_\varepsilon = 0.1$ (noisy learner), $N = 20$, and different MINE variants.

Another aspect that influences the bias-variance behavior of the neural estimator is regularization, such as the techniques mentioned in Section II-B. Figure 3 shows the same curves as Figure 2, but now, instead of varying N, we keep it constant at an intermediate value of $N = 20$ and compare the effects of different MINE variants. We use $\tau = 2$ for SMILE and $\lambda = 0.25$ for ReMINE, based on the range used by the original papers and on the average mutual information values to be estimated; we also use the discriminative variant DEMI and a combined SMILE+ReMINE variant employing both log-partition clipping and L2 regularization simultaneously. The effect of low τ in both SMILE variants is particularly noticeable for higher M, where the introduced clipping forces a bias plateau. DEMI manages to present slightly lower

bias and variance, since it sidesteps completely the need to approximate the log-partition, but that comes with a slightly increased training cost. However, this behavior of DEMI as well as the regularization effect introduced by ReMINE appear to be quite negligible for this low-dimensionality problem with relatively low MI values. There seems to be no significant effect of these methods on $M_{\mathrm{crit}}(N)$, although this may not hold for tasks with higher dimensionality.

C. Conditional Mutual Information Estimation



(a) $\sigma_\varepsilon = 0.1$ (noisy learner) (b) $\sigma_\varepsilon = 0$ (deterministic learner)

Fig. 4: Estimated CMI vs. training set size N, for different numbers of conditional contexts M_0.

The CMI $I(W;R|S^{\pm}) = \mathbb{E}_{S^{\pm}} I(W;R|S^{\pm} = s^{\pm})$ requires, in addition to the Monte-Carlo budget M, a split into M_0 independent supersamples and L draws of the selection vector R per supersample, with $M = M_0 L$. The law of total variance gives $\mathrm{Var}(\hat{I}) = \mathbb{E}_{S^{\pm}}(\mathrm{Var}(\hat{I}|s^{\pm})) + \mathrm{Var}(\mathbb{E}_{S^{\pm}}(\hat{I}|s^{\pm}))$. Increasing L (while decreasing M_0) reduces within-context noise (first term) but inflates the between-context component (second term).

Figure 4a verifies this trade-off empirically for the noisy learner with $\sigma_\varepsilon = 0.1$. We approximate the ground truth by using rectangular numerical integration to compute the mixture of Gaussians differential entropy $h(W|S^{\pm})$ in eq. (18). The curve with $M_0 = 5$ (and therefore $L = 10^4$) tracks the ground truth almost exactly up to $N = 15$, whereas $M_0 = 1$ overfits (high between-context bias) and $M_0 = M$ underfits (high within-context variance). The observed behavior for the curves with few draws L converging to zero as N increases reflects the observed underfit regime of Section IV-B, suggesting the existence of an analogous critical value $L_{\mathrm{crit}}(N)$ at which the error is at its minimum.

For the deterministic learner, the true conditional MI equals $N \log 2$ nats. Figure 4b shows that all settings severely underestimate this linear trend; estimates saturate around 4 nats. This failure once again relates to the sample complexity ceiling of [17] illustrated for the infinite MI in Section IV-A.

V. Conclusions

We have provided empirical guidelines for estimating information-theoretic generalization bounds for supervised learning settings. In particular, our results bring empirical evidence for the logarithmic saturation of the estimated mutual information, which limits its ability to track increasing functions of N. The bias–variance analysis highlights the need for sufficient Monte Carlo sampling relative to N to avoid underfitting and being highly biased towards zero. More specifically, this has implications for the procedure used to estimate CMI bounds, as there appears to be an optimal balance between the number of conditional contexts M_0 and the number of draws per context L.

Higher-dimensional problems may be of particular interest in future work to better characterize the differences between variational estimators. Additionally, future work will compare how different bounds behave for algorithms with and without specific algorithmic stability guarantees, in order to assess the practical relevance of tightness conditions in each case.

References

[1] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Artificial Intelligence and Statistics*, pp. 1232–1240, PMLR, 2016.

[2] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[3] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 587–591, 2019.

[4] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Conference on Learning Theory*, pp. 3437–3452, PMLR, 2020.

[5] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite, "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9925–9935, 2020.

[6] R. Zhou, C. Tian, and T. Liu, "Individually conditional individual mutual information bound on generalization error," *IEEE Transactions on Information Theory*, vol. 68, no. 5, pp. 3304–3316, 2022.

[7] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[8] B. Rodríguez Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via wasserstein distance," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19109–19121, 2021.

[9] R. Zhou, C. Tian, and T. Liu, "Stochastic chaining and strengthened information-theoretic generalization bounds," *Journal of the Franklin Institute*, vol. 360, no. 6, pp. 4114–4134, 2023.

[10] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 6, p. 066138, 2004.

[11] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: Mutual information neural estimation," 2021.

[12] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators.," *CoRR*, vol. abs/1604.03006, 2016.

[13] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, iv," *Communications on pure and applied mathematics*, vol. 36, no. 2, pp. 183–212, 1983.

[14] X. Lin, I. Sur, S. A. Nastase, A. Divakaran, U. Hasson, and M. R. Amer, "Data-efficient mutual information neural estimator," *arXiv preprint arXiv:1905.03319*, 2019.

[15] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," in *International Conference on Learning Representations*, 2020.

[16] K. Choi and S. Lee, "Regularized mutual information neural estimation," 2021.

[17] D. McAllester and K. Stratos, "Formal limitations on the measurement of mutual information," in *International Conference on Artificial Intelligence and Statistics*, pp. 875–884, PMLR, 2020.

[18] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *International conference on machine learning*, pp. 5171–5180, PMLR, 2019.