Slicing Mutual Information Generalization Bounds for Neural Networks

Kimia Nadjahi [*1] **Kristjan Greenewald** [*2] **Rickard Brüel Gabrielsson** [1] **Justin Solomon** [1]

Abstract

The ability of machine learning (ML) algorithms to generalize well to unseen data has been studied through the lens of information theory, by bounding the generalization error with the input-output mutual information (MI), *i.e.* the MI between the training data and the learned hypothesis. These bounds have limited empirical use for modern ML applications (e.g. deep learning) since the evaluation of MI is difficult in high-dimensional settings. Motivated by recent reports of significant low-loss compressibility of neural networks, we study the generalization capacity of algorithms which *slice* the parameter space, *i.e.* train on a random lower-dimensional subspace. We derive information-theoretic bounds on the generalization error in this regime, and discuss an intriguing connection to the k-Sliced Mutual Information, an alternative measure of statistical dependence which scales well with dimension. The computational and statistical benefits of our approach allow us to empirically estimate the input-output information of these neural networks and compute their information-theoretic generalization bounds, a task which was previously out of reach.

1. Introduction

Generalization is a fundamental task of machine learning, where a model that is optimized to perform well on training data must also perform well on test data drawn from the same underlying data distribution. Neural networks (NNs), in particular, are well suited to both achieving high performance on training data and generalizing well to test data, allowing them to achieve excellent test performance on highly complex tasks. Despite this empirical success, the architectural factors influencing how well a neural network generalizes are still not fully understood theoretically, motivating a significant body of work utilizing a wide variety of tools (e.g. PAC-Bayes (Dziugaite & Roy, 2017), information theory (Xu & Raginsky, 2017)) to bound the generalization error of NNs (Jiang et al., 2020).

We formally describe the generalization problem as follows. Let Z be the input data space (e.g. the set of feature-label pairs $z = (x, y)$), μ a probability distribution on Z, $\mathrm{W} \subseteq \mathbb{R}^D$ the hypothesis space (e.g. weights of a NN), and $\ell : \mathrm{Z} \times \mathrm{W} \to \mathbb{R}_+$ a loss function (e.g. the classification error). The training procedure seeks to find a $w \in \mathrm{W}$ with low *population risk* given by $\mathcal{R}(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$. In practice, obtaining $\mathcal{R}(w)$ is difficult since μ is generally unknown: one only observes a dataset comprising a finite number of samples from μ. Instead, given a training dataset $S_n \triangleq \{z_i \in \mathrm{Z}, i = 1, \dots, n\}$, $(z_i)_{i=1}^n$ i.i.d. from μ, one uses $\widehat{\mathcal{R}}_n(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, z_i)$, called the *empirical risk*. A learning algorithm can then be described as a function $\mathcal{A} : \mathrm{Z}^n \to \mathrm{W}$ which returns the optimal hypothesis W learned from S_n. W is in general random, and we denote its probability distribution by $P_{W|S_n}$. The *generalization error* of $\mathcal{A}$ is $\mathrm{gen}(\mu, \mathcal{A}) = \mathbb{E}[\mathcal{R}(W) - \widehat{\mathcal{R}}_n(W)]$ where the expectation $\mathbb{E}$ is taken with respect to (w.r.t.) the joint distribution of (W, S_n), *i.e.* $P_{W|S_n} \otimes \mu^{\otimes n}$.

In recent years, there has been a flurry of interest in using theoretical approaches to bound $\mathrm{gen}(\mu, \mathcal{A})$ using mutual information. The most common approach, introduced in (Xu & Raginsky, 2017), considers mutual information measures between S_n and the optimal hypothesis W learned from S_n. We denote the *Shannon mutual information* (MI) between two random variables X and Y as $\mathsf{I}(X;Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y$, with $p(x, y)$ denoting the joint distribution of (X, Y) at (x, y), and $p(x), p(y)$ the marginals. Subsequently, (Bu et al., 2019) used the averaging structure of the empirical loss to obtain the following bound (which we have specialized to our setting).

Theorem 1.1 (Bu et al. (2019))**.** *Assume there exists $C > 0$ such that for any $(\tilde{W}, \tilde{Z}) \sim P_W \otimes \mu$, $\ell(\tilde{W}, \tilde{Z}) \leq C$ almost surely. Then,*

$$|gen(\mu, \mathcal{A})| \leq \frac{C}{n} \sum_{i=1}^{n} \sqrt{\frac{\mathsf{I}(W; Z_i)}{2}},$$

where $W = \mathcal{A}(S_n)$.

Note that in Theorem 1.1, the MI terms are between W

[*]Equal contribution [1]MIT [2]MIT-IBM Watson AI Lab; IBM Research. Correspondence to: Kimia Nadjahi <knadjahi@mit.edu>, Kristjan Greenewald <kristjan.h.greenewald@ibm.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

and *individual* points in the dataset Z_i rather than the entire training dataset S_n, making the bound tighter in certain problems (Bu et al., 2019, Section IV).

These approaches suffer from the fact that the dimension of W can be very large in modern ML models, e.g. NNs, and the sample complexity of MI estimation scales very poorly with dimension (Paninski, 2003). These samples can be very expensive to obtain, especially with NNs, as one realization of W requires one complete training run.

For space reasons, we omit a full literature review of other variants of information-theoretic generalization bounds (e.g. *conditional MI* approaches (Steinke & Zakynthinou, 2020) and followup works), and focus on (Bu et al., 2019).

Sliced neural networks. While modern neural networks utilize very large numbers of parameters, common neural network architectures can be highly compressible by random slicing: Li et al. (2018) found that restricting $W \in \mathbb{R}^D$ during training to lie in a random d-dimensional subspace (where $d \ll D$) not only provided computational advantages, but did not meaningfully damage the performance of the learned neural network, for appropriate choice of d often two orders of magnitude smaller than D. They interpreted this fact as indicating *compressibility* of the neural network architecture up to some *intrinsic dimension* below which performance degrades. Recently, this framework has been applied by (Lotfi et al., 2022) to significantly improve PAC-Bayes generalization bounds, to the point where they closely match empirically observed generalization error.

Sliced mutual information. It is a natural question whether we can leverage the compression created by this slicing to obtain tighter and computationally-friendly information-theoretic generalization bounds. Intriguingly, a recent parallel line of work has considered slicing mutual information itself, yielding significant sample complexity and computational advantages in high-dimensional regimes. Goldfeld & Greenewald (2021); Goldfeld et al. (2022) proposed to slice the arguments of MI via random k-dimensional projections and define the k-Sliced Mutual Information (SMI) between $X \in \mathbb{R}^{d_x}$ and $Y \in \mathbb{R}^{d_y}$ as

$$\mathsf{SI}_k(X;Y) = \iint \mathsf{I}(\mathrm{A}^T X; \mathrm{B}^T Y)\, \mathrm{d}\sigma_{k,d_x} \otimes \sigma_{k,d_y}(\mathrm{A},\mathrm{B})\,,$$

where $\sigma_{k,d}$ is the Haar measure on $\mathrm{St}(k,d)$, the Stiefel manifold of $d \times k$ matrices with orthonormal columns. SI_k has been shown to retain many important properties of MI, for instance X and Y are independent if and only if $\mathsf{SI}_k(X;Y) = 0$ (Goldfeld & Greenewald, 2021; Goldfeld et al., 2022). More importantly, the statistical convergence rate for estimating $\mathsf{SI}_k(X;Y)$ depends on k but not the ambient dimensions d_x, d_y, providing significant advantages over MI (which in general requires an exponential number of samples with $\max(d_x, d_y)$ (Paninski, 2003)).

Note that similar convergence rates can be achieved while slicing in only one dimension (e.g. X), if samples from the conditional distribution $X|Y=y$ are available (Goldfeld & Greenewald, 2021), yielding

$$\mathsf{SI}_k^{(1)}(X;Y) = \int_{\mathrm{St}(k,d_x)} \mathsf{I}(\mathrm{A}^T X; Y)\, \mathrm{d}\sigma_{k,d_x}(\mathrm{A})\,. \tag{1}$$

Recently (Wongso et al., 2023) empirically connected generalization to $\mathsf{SI}_k^{(1)}(T;Y)$ between the hidden representations T of neural networks and the true class labels Y.

Our contributions. Motivated by the above, we introduce information-theoretic bounds studying the generalization capacity of learning algorithms trained on random subspaces. Our bounds demonstrate that "compressible" neural networks (via random slicing) sense have significantly better generalization guarantees. We also find an intriguing connection to k-SMI, which we explore in learning problems where the information-theoretic generalization bounds are possible to analytically compute. We then leverage the computational and statistical benefits of our sliced approach to empirically compute nonvacuous information-theoretic generalization bounds for various neural networks.

2. Sliced Information-Theoretic Generalization Bounds

We establish information-theoretic generalization bounds for any model $\mathcal{A}'$ whose parameters lie on the constraint set $\mathrm{W}_{\Theta,d} = \{w \in \mathbb{R}^D : \exists w' \in \mathbb{R}^d \text{ s.t. } w = \Theta w'\}$, where Θ is a random projection matrix of size $D \times d$ with $d < D$ and $\Theta^\top \Theta = \mathbf{I}_D$. Training $\mathcal{A}'$ consists in choosing d and randomly sampling Θ, then optimizing $w \in \mathrm{W}_{\Theta,d}$, which, given Θ, boils down to optimizing the subspace coefficients $w' \in \mathbb{R}^d$. We denote by $W'_\Theta \sim P_{W'_\Theta|\Theta,S_n}$ the optimal subspace coefficients. The associated generalization error is denoted by $\mathrm{gen}_d(\mu, \mathcal{A}')$ to make explicit the intrinsic dimension d induced by $\mathrm{W}_{\Theta,d}$. Note that since Θ is random (with distribution P_Θ), $\mathrm{gen}_d(\mu, \mathcal{A}')$ is, by definition, computed as an expectation over $P_{W'_\Theta|\Theta,S_n} \otimes P_\Theta \otimes \mu^{\otimes n}$.

For clarity purposes, we will use the notation $\ell(w', z) = \ell(w, z)\, \forall w = \Theta w' \in \mathrm{W}_{\Theta,d}$. For bounded loss (e.g. classification error), we have the following.

Theorem 2.1. *Assume there exists $C > 0$ s.t. for any $(\tilde{W}'_\Theta, \Theta, \tilde{Z}) \sim P_{W'_\Theta|\Theta} \otimes P_\Theta \otimes \mu$, $\ell(\tilde{W}'_\Theta, \tilde{Z}) \leq C$ almost surely where $P_{W'_\Theta|\Theta}$ is the conditional distribution of W'_Θ given Θ. Then,*

$$|gen_d(\mu, \mathcal{A}')| \leq \frac{C}{n} \sum_{i=1}^{n} \mathbb{E}_{\Theta \sim P_\Theta} \left[\sqrt{\frac{\mathsf{I}(W'_\Theta; Z_i)}{2}} \right]. \tag{2}$$

While state-of-the-art MI-based bounds depend on $\mathsf{I}(W; Z_i)$ (e.g. Theorem 1.1), we leverage the constraint set $\mathrm{W}_{\Theta,d}$

to construct a bound in terms of $\mathsf{I}(W'_\Theta; Z_i)$. As a result, our bound can be estimated more easily in practice, since W'_Θ is lower-dimensional. Theorem 2.1 is a particular case of more general bounds stated hereafter, which we derive by adapting the proof techniques of (Bu et al., 2019, Theorem 2). These bounds hold under milder conditions on the *cumulant-generating function* (CGF)[1] of $\ell(\tilde{W}'_\Theta, \tilde{Z})$ for $(\tilde{W}'_\Theta, \tilde{Z}) \sim P_{W'_\Theta|\Theta} \otimes \mu$.

Theorem 2.2. *If there exists $C_- \in \mathbb{R}_+^* \cup \{+\infty\}$ s.t. for $t \in (C_-, 0]$, $K_{\ell(\tilde{W}'_\Theta, \tilde{Z})}(t) \leq \psi_-(-t, \Theta)$, where $\psi_-(\cdot, \Theta)$ is convex and $\psi_-(0, \Theta) = (\psi_-)'(0, \Theta) = 0$, then,*

$$gen_d(\mu, \mathcal{A}') \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\Theta \sim P_\Theta} \left[\inf_{t \in [0, -C_-)} \frac{\mathsf{I}(W'_\Theta; Z_i) + \psi_-(t, \Theta)}{t} \right]. \tag{3}$$

If there exists $C_+ \in \mathbb{R}_+^* \cup \{+\infty\}$ s.t. for $t \in [0, C_+)$, $K_{\ell(\tilde{W}'_\Theta, \tilde{Z})}(t) \leq \psi_+(t, \Theta)$, where $\psi_+(\cdot, \Theta)$ is convex and $\psi_+(0, \Theta) = (\psi_+)'(0, \Theta) = 0$, then,

$$gen_d(\mu, \mathcal{A}') \geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\Theta \sim P_\Theta} \left[\inf_{t \in [0, C_+)} \frac{\mathsf{I}(W'_\Theta; Z_i) + \psi_+(t, \Theta)}{t} \right]. \tag{4}$$

We illustrate Theorem 2.2 in the next section, by computing the generalization bounds of two specific models. This also allows us to draw an interesting connection with k-SMI.

2.1. Connection to k-Sliced Mutual Information

Denote by $\|\cdot\|$ the Euclidean norm, $\mathbf{I}_D$ the $D \times D$ identity matrix, and $\mathbf{0}_D$ the D-dimensional zero vector. We denote by $W^{(D)}$ the solution of the unconstrained problem, *i.e.* $W^{(D)} = \arg\min_{w \in \mathbb{R}^D} \widehat{\mathcal{R}}_n(w)$.

Gaussian mean estimation. We first study the problem of estimating the mean of $Z \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D)$ via empirical risk minimization. The training dataset $S_n = (Z_1, \ldots, Z_n)$ consists of n independently and identically distributed (i.i.d.) samples from $\mathcal{N}(\mathbf{0}_D, \sigma^2 \mathbf{I}_D)$. Our objective is,

$$\underset{w \in \mathrm{W}_{\Theta,d}}{\arg\min}\, \widehat{\mathcal{R}}_n(w) \triangleq \frac{1}{n} \sum_{i=1}^{n} \|w - Z_i\|^2 . \tag{5}$$

We prove in Appendix A.3 that $W'_\Theta = \Theta^\top \bar{Z}$ with $\bar{Z} \triangleq (1/n) \sum_{i=1}^n Z_i$, and $\mathrm{gen}_d(\mu, \mathcal{A}') = 2\sigma^2 d/n$. Applying Theorem 2.2 yields

$$\mathrm{gen}_d(\mu, \mathcal{A}') \leq \frac{2}{n} \sum_{i=1}^{n} \sqrt{\lambda \mathsf{SI}_d^{(1)}(W^{(D)}; Z_i)}, \tag{6}$$

where $\lambda = \mathbb{E}_\Theta[\|\lambda_\Theta\|^2]$, $\lambda_\Theta \in \mathbb{R}^D$ is the vector of eigenvalues of $\sigma^2(\mathbf{I}_D + \Theta\Theta^\top/n)$. Note that here, $W^{(D)} = \bar{Z}$ hence $W'_\Theta = \Theta^\top W^{(D)}$, which explains the SMI term in the upper-bound of $\mathrm{gen}_d(\mu, \mathcal{A}')$. In the limit case $d = D$, our bound (6) boils down to the one established in (Bu et al., 2019), since $\lambda = \sigma^4 D(n+1)^2/n^2$ and $\mathsf{SI}_D^{(1)}(W^{(D)}; Z_i) = \mathsf{I}(W^{(D)}; Z_i)$.

Linear regression. Consider n samples $(x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^D$ and a response variable $y = (y_1, \ldots, y_n)$, $y_i \in \mathbb{R}$. Denote by $X \in \mathbb{R}^{n \times D}$ the data matrix such that the i-th row is x_i. We assume $n \geq D$ and aim at solving

$$\underset{w \in \mathrm{W}_{\Theta,d}}{\arg\min}\, \widehat{\mathcal{R}}_n(w) \triangleq \frac{1}{n} \|y - Xw\|^2 . \tag{7}$$

We show that $W'_\Theta = (\Theta X^\top X \Theta^\top)^{-1} \Theta X^\top y$. Additionally, we consider the fixed-design setting: X is deterministic and there exists $w^\star$ s.t. $y_i = x_i^\top w^\star + \varepsilon_i$ where $(\varepsilon_i)_{i=1}^n$ are i.i.d. samples from $\mathcal{N}(0, \sigma^2)$. Then, using Theorem 2.2, we prove that $\mathrm{gen}_d(\mu, \mathcal{A}')$ is upper-bounded by a function of $\mathsf{I}(\Theta_X^\top W^{(D)}; y_i)$ with $\Theta_X^\top \triangleq (\Theta X^\top X \Theta^\top)^{-1} \Theta (X^\top X)$, which can be interpreted as a generalized SMI with a non-isotropic slicing distribution that depends on the fixed X. The corresponding derivations are detailed in Appendix A.4.

2.2. Rate-distortion generalization bounds

The above bounds require the learned weights W to exactly lie in the subspace spanned by Θ. While this does work well empirically, it can be a restrictive assumption when d is very small. Since our MI-based bounds generally scale with increasing d, it is important to keep d small. Motivated by recent work in applying rate-distortion theory to input-output MI generalization bounds (Sefidgaran et al., 2022), we have the following generalization result for approximately compressible weights and Lipschitz loss.

Theorem 2.3. *Assume there exists $C > 0$ s.t. for any $(\tilde{W}, \tilde{Z}) \sim P_W \otimes \mu$, $\ell(\tilde{W}, \tilde{Z}) \leq C$ almost surely. Assume for any $z \in \mathrm{Z}$, $\ell(\cdot, z) : \mathrm{W} \to \mathbb{R}_+$ is L-Lipschitz. Then,*

$$|gen(\mu, \mathcal{A})| \leq 2L\mathbb{E}_{W^{(D)}(\Theta), W'_\Theta} \|W^{(D)}(\Theta) - \Theta W'_\Theta\| + \frac{C}{n} \sum_{i=1}^{n} \mathbb{E}_{\Theta \sim P_\Theta} \left[\sqrt{\frac{\mathsf{I}(W'_\Theta; Z_i)}{2}} \right].$$

where here, $\mathcal{A}$ may take Θ into account to output $W^{(D)}(\Theta)$, and $W'_\Theta = \Theta^T W^{(D)}(\Theta)$.

If we instead apply (Xu & Raginsky, 2017), we can obtain a rate-distortion type bound based on quantization that does not require estimation of mutual information (Appendix B).

3. Empirical Analysis

Gausian mean estimation. We first study the mean estimation problem described in Section 2.1. We choose $n = 200$,

[1] The CGF of a random variable X is $K_X(t) = \log \mathbb{E}[e^{t(X - \mathbb{E}[X])}]$.

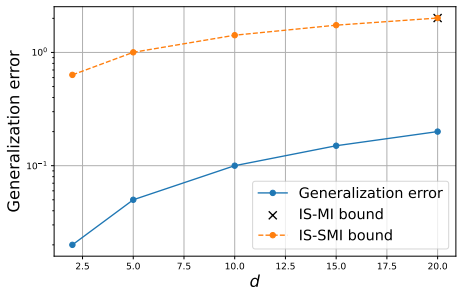

Figure 1. Empirical evaluation of IS-SMI bound (6) against d for Gaussian mean estimation. y-axis is in log scale. IS-MI is only evaluated for $D = 20$ since it cannot account for random slicing.

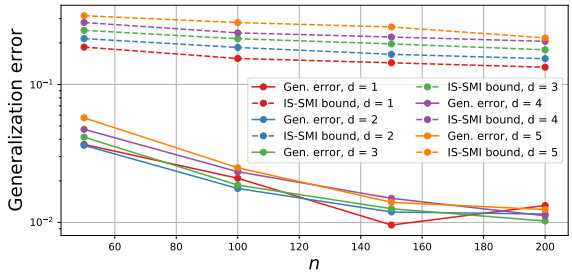

Figure 2. Empirical evaluation of (2) against n for logistic regression, with varying d. y-axis is in log scale.

$D = 20$ and $d \in \{2, 5, 10, 15, 20\}$, and compute the analytical generalization error for each d. We evaluate the individual sample MI (IS-MI) generalization bound (Bu et al., 2019) for D, which is available in closed form (Bu et al., 2019, Section IV.A). Then, for each d, we evaluate our bound (IS-SMI) (6), which requires approximating the SMI term. To this end, we use a Monte Carlo approximation based on 100 samples of Θ. Figure 1 confirms our bound, and more interestingly, shows that IS-SMI and the generalization error exhibit the same behavior, as they both increase with growing d. Besides, we observe that our bound boils down to IS-MI bound when $d = D$.

Logistic regression. We move on to empirical settings where $\mathsf{I}(W'_\Theta; Z_i)$ does not have an analytical solution, as opposed to the Gaussian mean estimation problem. We thus resort to MI estimators to evaluate our bound. We consider binary classification and for space constraints, refer to (Bu et al., 2019, Section VI) for full details on the empirical setting. Since the loss is bounded, we evaluate Theorem 2.1 using two types of MI estimators: k-nearest neighbor-based (k-NN-MI, (Kraskov et al., 2004)) and MINE (Belghazi et al., 2018). We observed in practice that kNN-MI returns NaN values as soon as $d > 2$, hence only report the bounds estimated with MINE. Figure 2 confirms our bound holds and accurately reflects the behavior of the generalization error as a function of d and n. We report the classification errors in Appendix C.2.

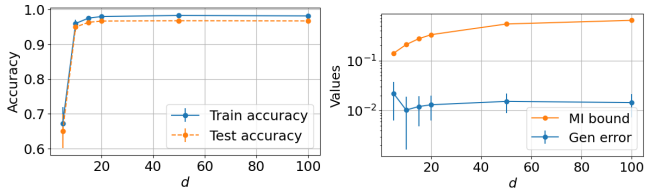

Figure 3. Generalization bounds of NNs trained on Iris dataset

Neural networks. We demonstrate that our derived generalization bounds on random subspaces allow us compute generalization bounds on a simple machine learning classification task involving neural networks, while also maintaining performance. We classify the Iris dataset (Fisher, 1936) with a two-hidden-layer NN with 10,903 parameters, and train 10,000 instantiations for 200 epochs. We estimate MI with MINE. We report results for $d \in \{5, 10, 15, 20, 50, 100\}$ in Figure 3, and refer to Appendix C.3 for further details. We obtain over 95% accuracy at $d = 10$ already, and both the best train and test accuracy is achieved for $d = 50$. As expected, our bound is an increasing function of d and all bounds are non-vacuous.

4. Conclusion

In this work, we combined recent empirical schemes for finding compressed models, including NNs, via random slicing with generalization bounds based on input-output MI. Our results indicate that architectures that are amenable to this compression scheme yield tighter information-theoretic generalization bounds. We also explore a notion of *approximate compressibility*, where the learned parameters are close to the compressed subspace but do not lie on it exactly. This framework provides more flexibility in the trained model, allowing it to maintain good training error for even smaller (approximate) projection dimension d, and ensuring that the resulting generalization bounds are as tight as possible.

Our contributions motivate further analyses which leverage compressibility to improve the tightness of information-theoretic generalization bounds. They can also help inform selection and design of NN architectures in practice. Future work include an empirical study of our rate-distortion type bound, and extension of the approach to other approximate compression schemes such as quantization. This will also be combined with an exploration of regularization approaches that encourage trained NNs to be as approximately compressible as possible to ensure that our bound is small in practice, while also potentially providing empirical benefits in observed test performance itself. Finally, we will explore the optimization of the rate-distortion tradeoff in order to obtain the best generalization bounds, potentially making use of analytical bounds on information that do not require estimating MI from multiple training runs of the network.

References

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, R. D. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Bu, Y., Zou, S., and Veeravalli, V. V. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 587–591, 2019. doi: 10.1109/ISIT.2019.8849590.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–188, 1936.

Goldfeld, Z. and Greenewald, K. Sliced mutual information: A scalable measure of statistical dependence. *Advances in Neural Information Processing Systems*, 34:17567–17578, 2021.

Goldfeld, Z., Greenewald, K., Nuradha, T., and Reeves, G. k-sliced mutual information: A quantitative study of scalability with dimension. *arXiv preprint arXiv:2206.08526*, 2022.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138.

Li, C., Farkhoor, H., Liu, R., and Yosinski, J. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Lotfi, S., Finzi, M., Kapoor, S., Potapczynski, A., Goldblum, M., and Wilson, A. G. PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

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

Sefidgaran, M., Gohari, A., Richard, G., and Simsekli, U. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. In *Conference on Learning Theory*, pp. 4416–4463. PMLR, 2022.

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

Wongso, S., Ghosh, R., and Motani, M. Using sliced mutual information to study memorization and generalization in deep neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11608–11629. PMLR, 2023.

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

A. Postponed Proofs for Section 2

A.1. Proof of Theorem 2.2

Consider a pair of random variables $(X, Y) \in \mathbb{R}^D \times \mathbb{R}^{D'}$, with joint distribution $P_{X,Y}$ and marginals P_X, P_Y. Let $\tilde{X}$ (resp., $\tilde{Y}$) be an independent copy of X (resp., Y) such that $P_{\tilde{X},\tilde{Y}} = P_X \otimes P_Y$.

Let (Θ, Γ) be a pair of independent random matrices of size $d \times D$ and $d' \times D'$ respectively, with $d < D$ and $d' < D'$. Denote by $P_{\Theta,\Gamma} = P_\Theta \otimes P_\Gamma$ the joint distribution of (Θ, Γ).

Let $f : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R}$ and given $(\Theta, \Gamma) \sim P_\Theta \otimes P_\Gamma$, denote by $K_{f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})}$ the cumulant generating function of the random variable $f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})$, *i.e.* for $t \in \mathbb{R}$,

$$K_{f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})}(t) = \log \mathbb{E}\left[e^{t(f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y}) - \mathbb{E}[f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})])}\right] \tag{8}$$

where the expectation is taken with respect to $P_{\Theta^\top X|\Theta} \otimes P_{\Gamma^\top Y|\Gamma}$.

Lemma A.1. *Suppose that for any (Θ, Γ), there exists $b_+ \in \mathbb{R}_+^* \cup \{+\infty\}$ and a convex function $\varphi_+(\cdot, \Theta, \Gamma) : [0, b_+) \to \mathbb{R}$ such that $\varphi_+(0, \Theta, \Gamma) = \varphi'_+(0, \Theta, \Gamma) = 0$ and for $t \in [0, b_+)$, $K_{f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})}(t) \leq \psi_+(t, \Theta, \Gamma)$. Then,*

$$\mathbb{E}[f(\Theta^\top X, \Gamma^\top Y)] - \mathbb{E}[f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})] \leq \mathbb{E}_{P_\Theta \otimes P_\Gamma}\left[\inf_{t \in [0, b_+)} \frac{I(\Theta^\top X; \Gamma^\top Y) + \psi_+(t, \Theta, \Gamma)]}{t}\right]. \tag{9}$$

Suppose that for any (Θ, Γ), there exists $b_- \in \mathbb{R}_+^* \cup \{+\infty\}$ and a convex function $\varphi_-(\cdot, \Theta, \Gamma) : [0, b_-) \to \mathbb{R}$ such that $\varphi_-(0, \Theta, \Gamma) = \varphi'_-(0, \Theta, \Gamma) = 0$ and for $t \in (b_-, 0]$, $K_{f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})}(t) \leq \psi_-(-t, \Theta, \Gamma)$. Then,

$$\mathbb{E}[f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})] - \mathbb{E}[f(\Theta^\top X, \Gamma^\top Y)] \leq \mathbb{E}_{P_\Theta \otimes P_\Gamma}\left[\inf_{t \in [0, -b_-)} \frac{I(\Theta^\top X; \Gamma^\top Y) + \psi_-(t, \Theta, \Gamma)}{t}\right]. \tag{10}$$

Proof. Fix Θ, Γ. By Donsker-Varadhan variational representation,

$$\mathbf{KL}(P_{(\Theta^\top X, \Gamma^\top Y)|\Theta,\Gamma} \| P_{\Theta^\top X|\Theta} \otimes P_{\Gamma^\top Y|\Gamma}) = \sup_{g \in \mathcal{G}} \mathbb{E}[g(\Theta^\top X, \Gamma^\top Y)] - \log \mathbb{E}[\exp(g(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y}))] \tag{11}$$

where $\mathcal{G} = \{g : \mathbb{R}^d \times \mathbb{R}^{d'} \to \mathbb{R} \ : \ \mathbb{E}[e^{g(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})}] < \infty\}$. Therefore, for any $t \in [0, b_+)$,

$$\begin{aligned}
\mathbf{KL}(P_{(\Theta^\top X, \Gamma^\top Y)|\Theta,\Gamma} \| P_{\Theta^\top X|\Theta} \otimes P_{\Gamma^\top Y|\Gamma}) &\geq t\mathbb{E}[f(\Theta^\top X, \Gamma^\top Y)] - \log \mathbb{E}[\exp(tf(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y}))] &(12)\\
&\geq t\left(\mathbb{E}[f(\Theta^\top X, \Gamma^\top Y)] - \mathbb{E}[f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})]\right) - \psi_+(t, \Theta, \Gamma) &(13)
\end{aligned}$$

where (13) follows from the assumption that for $t \in [0, b_+)$, $K_{f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})}(t) \leq \psi_+(t, \Theta, \Gamma)$. Hence,

$$\mathbb{E}[f(\Theta^\top X, \Gamma^\top Y)] - \log \mathbb{E}[\exp(\lambda f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y}))] \leq \inf_{t \in [0, b_+)} \frac{I(\Theta^\top X; \Gamma^\top Y) + \psi_+(t, \Theta, \Gamma)}{t}. \tag{14}$$

Our final result (9) follows from taking the expectation of (14) over $(\Theta, \Gamma) \sim P_\Theta \otimes P_\Gamma$.

We can prove analogously that (10) holds, assuming that for $t \in [0, b_-)$, $K_{f(\Theta^\top \tilde{X}, \Gamma^\top \tilde{Y})}(t) \leq \psi_-(-t, \Theta, \Gamma)$. □

Proof of Theorem 2.2. By definition, the generalization error is

$$\text{gen}_d(\mu, \mathcal{A}') = \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}_{P_{W'_\Theta|\Theta} \otimes P_\Theta \otimes \mu}[\ell(\Theta W'_\Theta, Z)] - \mathbb{E}_{P_{W'_\Theta, Z_i|\Theta} \otimes P_\Theta}[\ell(\Theta W'_\Theta, Z_i)]\right\}, \tag{15}$$

where given $\Theta \sim P_\Theta$, $W'_\Theta \in \mathbb{R}^d$ is such that there exists $W \in \mathbb{R}^D$, $W'_\Theta = \Theta^\top W$. Assume $\mathrm{Z} \subset \mathbb{R}^s$ and define $\ell' : \mathbb{R}^d \times \mathrm{Z} \to \mathbb{R}$ s.t. given $\Theta \sim P_\Theta$, $\ell'(w', z) = \ell(\Theta w', z)$. We can then reformulate (15) as,

$$\text{gen}_d(\mu, \mathcal{A}') = \mathbb{E}_{P_{W'_\Theta|\Theta} \otimes P_\Theta \otimes \mu \otimes P_\Gamma}[\ell'(W'_\Theta, \Gamma^\top S_n)] - \mathbb{E}_{P_{W'_\Theta, \Gamma^\top S_n|\Theta,\Gamma} \otimes P_\Theta \otimes P_\Gamma}[\ell'(W'_\Theta, \Gamma^\top S_n)], \tag{16}$$

where P_Γ is the uniform distribution over $\{e_1, \ldots, e_n\}$, with for $i = 1, \ldots, n$, $e_i \in \mathbb{R}^n$ s.t. $e_{i,i} = 1$ and $e_{i,j} = 0$ for $j \in \{1, \ldots, s\}, j \neq i$. Our final bounds (4) and (3) result from applying Lemma A.1 on (15). □

A.2. Applications of Theorem 2.2 to sub-Gaussian or bounded loss

Corollary A.2 (Sub-Gaussian loss)**.** *Suppose that for all $\Theta \sim P_\Theta$, there exists C_Θ^2 such that for $t \in \mathbb{R}$,*

$$\mathbb{E}_{P_{W'_\Theta|\Theta}\otimes\mu}\left[e^{t\{\ell(\Theta W'_\Theta,Z)-\mathbb{E}_{P_{W'_\Theta|\Theta}\otimes\mu}[\ell(\Theta W'_\Theta,Z)]\}}\right] \leq e^{C_\Theta^2 t^2/2} \tag{17}$$

Then,

$$|gen_d(\mu,\mathcal{A}')| \leq \frac{\sqrt{2}}{n}\sum_{i=1}^{n}\mathbb{E}_{\Theta\sim P_\Theta}\left[\sqrt{C_\Theta^2 \mathsf{I}(W'_\Theta;Z_i)}\right]. \tag{18}$$

Proof. Let $\Theta \sim P_\Theta$. Assuming (17), implies that for any $t \in \mathbb{R}$,

$$K_{\ell(\Theta\tilde{W}'_\Theta,\tilde{Z})}(t) \leq \frac{C_\Theta^2 t^2}{2}. \tag{19}$$

We conclude by applying Theorem 2.2 and the fact that for $i \in \{1,\dots,n\}$,

$$\inf_{\lambda>0}\frac{\mathsf{I}(W'_\Theta;Z_i)+C_\Theta^2 t^2/2}{t} = \sqrt{2C_\Theta^2\mathsf{I}(W'_\Theta;Z_i)}. \tag{20}$$

□

Proof of Theorem 2.1. Let $\Theta \sim P_\Theta$. By Hoeffding's lemma, for all $t \in \mathbb{R}$,

$$\mathbb{E}_{P_{W'_\Theta|\Theta}\otimes\mu}\left[e^{t\{\ell(\Theta W'_\Theta,Z)-\mathbb{E}_{P_{W'_\Theta|\Theta}\otimes\mu}[\ell(\Theta W'_\Theta,Z)]\}}\right] \leq e^{M^2\lambda^2/8}. \tag{21}$$

Therefore, (17) is satisfied with $C_\Theta^2 = \frac{M^2}{4}$ for all $\Theta \sim P_\Theta$. Applying Corollary A.2 along with the linearity of the expectation completes the proof. □

Remark A.3. By applying Jensen's inequality on the right-hand side term of (18), we obtain

$$|\mathrm{gen}_d(\mu,\mathcal{A}')| \leq \frac{\sqrt{2C_\Theta^2}}{n}\sum_{i=1}^{n}\sqrt{I(W'_\Theta;Z_i|\Theta)} \tag{22}$$

A.3. Detailed derivations for Gaussian mean estimation problem

First, we justify why, given $\Theta \sim P_\Theta$ s.t. $\Theta^\top\Theta = \mathbf{I}_d$ and $W'_\Theta = \Theta^\top\bar{Z}$, $W = \Theta W'_\Theta$ is the solution of $\arg\min_{w\in \mathrm{W}_{\Theta,d}}\widehat{\mathcal{R}}_n(w)$, with

$$\forall w \in \mathrm{W}_{\Theta,d},\quad \widehat{\mathcal{R}}_n(w) = \frac{1}{n}\sum_{i=1}^{n}\|w-Z_i\|^2. \tag{23}$$

By writing $w = \Theta w'$ and deriving the gradient of (23) with respect to w', we obtain

$$\nabla_{w'}\widehat{\mathcal{R}}_n(\Theta w') = \frac{2}{n}\sum_{i=1}^{n}\Theta^\top(\Theta w'-Z_i). \tag{24}$$

Solving $\nabla_{w'}\widehat{\mathcal{R}}_n(w') = 0$ yields $(\Theta^\top\Theta)w' = \Theta^\top\bar{Z}$. We conclude by using $\Theta^\top\Theta = \mathbf{I}_d$.

Generalization error. We recall that the generalization error is defined as

$$\mathrm{gen}_d(\mu,\mathcal{A}') = \mathbb{E}[\mathcal{R}(W)-\widehat{\mathcal{R}}_n(W)], \tag{25}$$

where the expectation is computed with respect to $P_{W,S_n} = P_{W'_\Theta|S_n}\otimes P_\Theta\otimes\mu^{\otimes n}$.

We prove that the expectation of $\widehat{\mathcal{R}}_n(W)$ over $P_{W'_\Theta|S_n} \otimes P_\Theta \otimes \mu^{\otimes n}$ is $\mathbb{E}[\widehat{\mathcal{R}}_n(W)] = \sigma^2(D - d/n)$.

$$\mathbb{E}[\widehat{\mathcal{R}}_n(W)] = \mathbb{E}[\widehat{\mathcal{R}}_n(\Theta W'_\Theta)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\Theta W'_\Theta - Z_i\|^2] \tag{26}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left\{\mathbb{E}[\|\Theta W'_\Theta\|^2] - 2\mathbb{E}[(W'_\Theta)^\top \Theta^\top Z_i] + \mathbb{E}[\|Z_i\|^2]\right\} \tag{27}$$

Since $W'_\Theta = \Theta^\top \bar{Z}$, $\Theta^\top\Theta = \mathbf{I}_d$, $Z_1, \dots, Z_n$ are n i.i.d. samples from $\mathcal{N}(\mathbf{0}_D, \mathbf{I}_D)$, we deduce $W'_\Theta \sim \mathcal{N}(\mathbf{0}, (\sigma^2/n)\mathbf{I}_d)$, and

$$\mathbb{E}[\|\Theta W'_\Theta\|^2] = \text{Tr}(\mathbb{E}[(\Theta W'_\Theta)^\top (\Theta W'_\Theta)]) \tag{28}$$

$$= \text{Tr}(\mathbb{E}[(W'_\Theta)^\top W']) \tag{29}$$

$$= \frac{\sigma^2 d}{n}\,. \tag{30}$$

Besides, for $i \in \{1, \dots, n\}$,

$$\mathbb{E}[\|Z_i\|^2] = \text{Tr}(\mathbb{E}[Z_i^\top Z_i]) = \sigma^2 D\,, \tag{31}$$

and

$$\mathbb{E}[(W'_\Theta)^\top \Theta^\top Z_i] = \mathbb{E}[\bar{Z}^\top \Theta\Theta^\top Z_i] \tag{32}$$

$$= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[Z_j^\top \Theta\Theta^\top Z_i] \tag{33}$$

$$= \frac{1}{n}\sum_{j=1}^{n} \text{Tr}(\mathbb{E}[Z_j^\top \Theta\Theta^\top Z_i]) \tag{34}$$

$$= \frac{1}{n}\text{Tr}(\mathbb{E}[Z_i^\top \Theta\Theta^\top Z_i]) \tag{35}$$

$$= \frac{\sigma^2 d}{n}\,. \tag{36}$$

We conclude that,

$$\mathbb{E}[\widehat{\mathcal{R}}_n(W)] = \frac{1}{n}\sum_{i=1}^{n} \left\{\frac{\sigma^2 d}{n} - \frac{2\sigma^2 d}{n} + \sigma^2 D\right\} \tag{37}$$

$$= \sigma^2\left(D - \frac{d}{n}\right)\,. \tag{38}$$

The true risk is defined for $w \in \mathrm{W}_{\Theta,d}$ as $\mathcal{R}(w) = \mathbb{E}[\|w - \tilde{Z}\|^2]$, where the expectation is computed over $\tilde{Z} \sim \mathcal{N}(\mathbf{0}_D, \sigma^2\mathbf{I}_D)$. We have,

$$\mathcal{R}(w) = \mathcal{R}(\Theta w') = \mathbb{E}[\text{Tr}((\Theta w' - \tilde{Z})^\top (\Theta w' - \tilde{Z}))] \tag{39}$$

$$= \text{Tr}(w'^\top \Theta^\top \Theta w') - 2\text{Tr}(w'^\top \Theta^\top \mathbb{E}[\tilde{Z}]) + \text{Tr}(\mathbb{E}[\tilde{Z}^\top \tilde{Z}]) \tag{40}$$

$$= \text{Tr}(w'^\top w') + \sigma^2 D \tag{41}$$

where (41) results from $\Theta^\top\Theta = \mathbf{I}_d$ and $\tilde{Z} \sim \mathcal{N}(\mathbf{0}_D, \sigma^2\mathbf{I}_D)$.

Taking the expectation of $\mathcal{R}(W)$ over $P_{W'_\Theta|S_n}$ yields,

$$\mathbb{E}[\mathcal{R}(W)] = \mathbb{E}[\mathcal{R}(\Theta W'_\Theta)] = \mathbb{E}[\text{Tr}((W'_\Theta)^\top W'_\Theta) + \sigma^2 D] \tag{42}$$

$$= \text{Tr}(\mathbb{E}[(W'_\Theta)^\top W'_\Theta]) + \sigma^2 D \tag{43}$$

$$= \sigma^2\left(D + \frac{d}{n}\right)\,. \tag{44}$$

where (44) follows from (30).

We can now compute the generalization error using (38) and (44): we obtain,

$$\text{gen}_d(\mu, \mathcal{A}') = \mathbb{E}[\mathcal{R}(W) - \widehat{\mathcal{R}}_n(W)] = \frac{2\sigma^2 d}{n}\,. \tag{45}$$

Generalization error upper-bound. We study the cumulant generating function of

$$\ell(W, \tilde{Z}) = \ell(\Theta W'_\Theta, \tilde{Z}) = \|\Theta W'_\Theta - \tilde{Z}\|^2\,, \tag{46}$$

where $\tilde{Z}, W'_\Theta$ are independent. Since $W'_\Theta \sim \mathcal{N}(\mathbf{0}_d, (\sigma^2/n)\mathbf{I}_d)$ and $\tilde{Z} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$, then

$$(\Theta W'_\Theta - \tilde{Z}) \sim \mathcal{N}(\mathbf{0}_D, \sigma^2(\Theta\Theta^\top/n\mathbf{I}_D))\,. \tag{47}$$

Therefore, $\ell(W, \tilde{Z})$ is the sum of squares of D dependent Gaussian random variables, which can equivalently be written as,

$$\ell(W, \tilde{Z}) = \sum_{k=1}^{D} \lambda_{\Theta,k} U_{\Theta,k}^2\,, \tag{48}$$

$$U_\Theta = P\Sigma_\Theta^{-1/2}(\Theta W'_\Theta - \tilde{Z}) \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D) \tag{49}$$

$$\Sigma_\Theta = \sigma^2(\Theta\Theta^\top/n + \mathbf{I}_D) \tag{50}$$

where $P \in \mathbb{R}^{D\times D}$ and $\lambda_\Theta = (\lambda_{\Theta,1}, \ldots, \lambda_{\Theta,D}) \in \mathbb{R}^D$ come from the eigendecomposition of Σ_Θ, *i.e.* $\Sigma_\Theta = P\Lambda P^\top$, with Λ the diagonal matrix such that $\Lambda_{k,k} = \lambda_{\Theta,k}$. Note that, since Σ_Θ is positive definite, P is orthogonal and for any $k \in \{1, \ldots, D\}$, $\lambda_{\Theta,k} > 0$.

By (48), $\ell(W, \tilde{Z})$ is a linear combination of independent chi-square variables (each with 1 degree of freedom), and is thus distributed from a generalized chi-square distribution. We deduce that the CGF of $\ell(W, \tilde{Z})$ is given for $s \leq \frac{1}{2}\min_{k\in\{1,\ldots,D\}} \lambda_{\Theta,k}$ as,

$$K_{\ell(W,\tilde{Z})}(s) = -s\sum_{k=1}^{D}\lambda_{\Theta,k} - \frac{1}{2}\sum_{k=1}^{D}\log(1 - 2\lambda_{\Theta,k}s) \tag{51}$$

$$= \frac{1}{2}\sum_{k=1}^{D}[-2\lambda_{\Theta,k}s - \log(1 - 2\lambda_{\Theta,k}s)]\,. \tag{52}$$

Since for $t < 0$, $-t - \log(1-t) \leq t^2/2$, we can bound $K_{\ell(W,\tilde{Z})}(s)$ for $s < 0$ as follows.

$$K_{\ell(W',\tilde{Z})}(s) \leq \frac{1}{2}\sum_{k=1}^{D}\frac{(2\lambda_{\Theta,k}s)^2}{2} = \|\lambda_\Theta\|^2 s^2\,. \tag{53}$$

We then apply Theorem 2.2 along with (53) and Jensen's inequality to prove (6).

A.4. Detailed derivations for linear regression

Problem statement. Consider n samples of a D-dimensional random variable $(x_1, \ldots, x_n)$ and a response variable $y = (y_1, \ldots, y_n)$ where $y_i \in \mathbb{R}$. Denote by $X \in \mathbb{R}^{n\times D}$ the data matrix such that the i-th row is x_i. The empirical risk is defined for any $w \in \mathbb{R}^D$ as

$$\widehat{\mathcal{R}}_n(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top w)^2 = \frac{1}{n}\|y - Xw\|^2 \tag{54}$$

Let $d \in \mathbb{N}^*$, $d < D$ and $\Theta \sim P_\Theta$, $\Theta^\top\Theta = \mathbf{I}_d$. Our objective is,

$$\underset{w\in \mathrm{W}_{\Theta,d}}{\arg\min}\, \widehat{\mathcal{R}}_n(w) \tag{55}$$

We consider that the problem is *over-determined*, the rank of X is D, thus $D \leq n$. $X^\top X$ is then invertible, and since $X^\top X$ is always positive semi-definite, this implies that $X^\top X$ is positive definite. The solution of (55) is unique and given by,

$$W'_\Theta = (\Theta X^\top X \Theta^\top)^{-1} \Theta X^\top y \tag{56}$$

Note that the solution of $\arg\min_{w \in \mathbb{R}^D} \widehat{\mathcal{R}}_n(w)$ in the over-determined setting yields the ordinary least squares (OLS) estimator, given by

$$W^{(D)} = (X^\top X)^{-1} X^\top y \,. \tag{57}$$

Hence, by comparing (57) and (56), we have

$$W'_\Theta = (\Theta X^\top X \Theta^\top)^{-1} \Theta (X^\top X) W^{(D)} \tag{58}$$

Generalization error bounds. We consider the fixed-design setting, *i.e.* for $i = 1, \dots, n$, $y_i = x_i^\top W^\star + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0,1)$. First, by using analogous derivations as in Appendix A.3, we can show that

$$\mathrm{gen}_d(\mu, \mathcal{A}') = \frac{2\sigma^2 d}{n} \,. \tag{59}$$

Since $y_i \sim \mathcal{N}(x_i^\top W^\star, \sigma^2)$, then by using (56),

$$x_i^\top \Theta^\top W' \sim \mathcal{N}(x_i^\top \Theta_X W^\star, \sigma^2 x_i^\top \Theta^\top [\Theta X^\top X \Theta^\top]^{-1} \Theta x_i) \tag{60}$$

where $\Theta_X = \Theta^\top (\Theta X^\top X \Theta^\top)^{-1} \Theta (X^\top X) \in \mathbb{R}^{D \times D}$. Therefore, $(\tilde{y}_i - x_i^\top \Theta^\top W') \sim \mathcal{N}(x_i^\top (I - \Theta_X) W^\star, \sigma^2 (1 + x_i^\top \Theta^\top [\Theta X^\top X \Theta^\top]^{-1} \Theta x_i))$, and for $W = \Theta W'_\Theta$ and $\tilde{y}_i$ s.t. W and $\tilde{y}_i$ are independent,

$$\ell(W, \tilde{y}_i) \sim \sigma_i^2 \, \chi^2(1, \lambda_i) \tag{61}$$

where $\sigma_i^2 = \sigma^2 (1 + x_i^\top \Theta^\top [\Theta X^\top X \Theta^\top]^{-1} \Theta x_i)$, $\lambda_i = (x_i^\top (I - \Theta_X) W^\star)^2$ and $\chi^2(k, \lambda)$ denotes the noncentral chi-squared distribution with k degrees of freedom and noncentrality parameter λ. We deduce that the moment-generating function of $\ell(W, \tilde{y}_i)$ is given for $s < 1/(2\sigma_i^2)$ by,

$$\mathbb{E}[\exp(s\, \ell(W, \tilde{y}_i))] = \frac{e^{(\lambda_i \sigma_i^2 s)/(1 - 2\sigma_i^2 s)}}{(1 - 2\sigma_i^2 s)^{1/2}} \tag{62}$$

and its expectation is $\mathbb{E}[\ell(W, \tilde{y}_i)] = \sigma_i^2 (1 + \lambda_i)$. Therefore, for $s < 1/(2\sigma_i^2)$,

$$K_{\ell(W, \tilde{y}_i)}(s) = \frac{\lambda_i u_i}{2(1 - u_i)} - \frac{1}{2} \log(1 - u_i) - \frac{1}{2}(1 + \lambda_i) u_i \tag{63}$$

$$= \frac{1}{2} \{ - \log(1 - u_i) - u_i \} + \frac{\lambda_i u_i^2}{2(1 - u_i)} \tag{64}$$

with $u_i = 2\sigma_i^2 s$. Since $-\log(1 - x) - x \leq \frac{x^2}{2}$ for $x < 0$, we deduce that for $s < 0$,

$$K_{\ell(W', \tilde{y}_i)}(s) \leq \frac{u_i^2}{4} + \frac{\lambda_i u_i^2}{2(1 - u_i)} \tag{65}$$

$$= \sigma_i^4 s^2 + \frac{2\lambda_i \sigma_i^4 s^2}{1 - 2\sigma_i^2 s} \,. \tag{66}$$

Then, by applying Theorem 2.2,

$$\mathrm{gen}_d(\mu, \mathcal{A}') \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\Theta \left[\inf_{s > 0} \frac{\mathsf{I}(W'_\Theta; y_i) + \sigma_i^4 s^2 \big(1 + 2\lambda_i (1 + 2\sigma_i^2 s)^{-1}\big)}{s} \right] \tag{67}$$

Discussion. By (58), W'_Θ is equal to a projection of the unconstrained problem's solution $W^{(D)}$, hence the right-hand side term in (67) can be interpreted as a generalized SMI.

As d gets closer to D, $\lambda = (\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n$ decreases to $\mathbf{0}_n$. Indeed, consider the compact singular value decomposition (SVD) of $X\Theta^T = USV^T$, where $S \in \mathbb{R}^{d\times d}$ is diagonal, $U \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{d\times m}$, such that $U^\top U = V^\top V = \mathbf{I}_d$. Then, using the pseudo-inverse expression of SVD,

$$X\Theta_X = X\Theta^\top (VS^{-1}U^\top)X \tag{68}$$
$$= USV^\top VS^{-1}U^\top X \tag{69}$$
$$= UU^\top X \tag{70}$$

Therefore,

$$\sqrt{\lambda} = (\mathbf{I}_n - UU^\top)XW^\star \tag{71}$$

Since $U^\top U = \mathbf{I}_n$ and U is of size $n \times d$, we can consider $\bar{U} \in \mathbb{R}^{n\times(n-d)}$ such that $[U, \bar{U}]$ is an orthogonal $n \times n$ matrix. Then, $\mathbf{I}_n = [U, \bar{U}][U, \bar{U}]^\top = UU^T + \bar{U}\bar{U}^T$, so $\mathbf{I}_n - UU^T = \bar{U}\bar{U}^T$. We deduce that $\mathbf{I}_n - UU^\top$ is a matrix with $(n-d)$ eigenvalues equal to 1, and the d remaining eigenvalues are zero. Hence, increasing d corresponds to increasing the number of null eigenvalues, which implies that λ converges to $\mathbf{0}_n$.

A.5. Proof of Theorem 2.3

Proof of Theorem 2.3. By the triangle inequality,

$$|\text{gen}(\mu, \mathcal{A})| \le |\text{gen}(\mu, \mathcal{A}) - \text{gen}_d(\mu, \mathcal{A}')| + |\text{gen}_d(\mu, \mathcal{A}')| \tag{72}$$

The final result follows from bounding (72) from above, by applying Theorem 2.1 to bound $\text{gen}_d(\mu, \mathcal{A}')$, and using the L-Lipschitz assumption on the loss to show

$$|\text{gen}(\mu, \mathcal{A}) - \text{gen}_d(\mu, \mathcal{A}')| \le 2L\mathbb{E}_{W^{(D)}(\Theta), W'_\Theta}\|W^{(D)}(\Theta) - \Theta W'_\Theta\|\,. \tag{73}$$

□

B. Generalization Bounds with Quantization

By using analogous arguments as in (Xu & Raginsky, 2017, Section 4.1), we refine Theorem 2.1 when W'_Θ is assumed to lie on a countable space. We state the formal result below.

Corollary B.1. *Assume there exists $C > 0$ s.t. for any Θ and $(\tilde{W}'_\Theta, \tilde{Z}) \sim P_{W'_\Theta|\Theta} \otimes \mu$, $\ell(\tilde{W}'_\Theta, Z) \le C$ almost surely. Additionally, assume that for $\Theta \sim P_\Theta$, $W'_\Theta \in \mathrm{W}'$ s.t. the cardinality of W' is k. Then,*

$$|gen_d(\mu, \mathcal{A}')| \le C\sqrt{\frac{\log(k)}{2}}\,. \tag{74}$$

Theorem B.2. *Assume the conditions of Theorem 2.3 hold. Furthermore, suppose that $\|W'_\Theta\| \le M$ (e.g. as a result of enforcing the Lipschitz constant L), and consider a function $q(W'_\Theta)$ quantizing the coordinates of W'_Θ with stepsize ϵ. Then,*

$$|gen(\mu, \mathcal{A})| \le 2L\left(\mathbb{E}_{W^{(D)}(\Theta), W'_\Theta}[\|W^{(D)}(\Theta) - \Theta W'_\Theta\|] + \epsilon\sqrt{d}\right) + C\mathbb{E}_{\Theta\sim P_\Theta}\left[\sqrt{\frac{\mathrm{I}(q(W'_\Theta); Z_i)}{2n}}\right] \tag{75}$$
$$\le 2L\left(\mathbb{E}_{W^{(D)}(\Theta), W'_\Theta}[\|W^{(D)}(\Theta) - \Theta W'_\Theta\|] + \epsilon\sqrt{d}\right) + C\sqrt{\frac{d\log(2M/\epsilon)}{2n}}\,. \tag{76}$$

Proof of Theorem B.2. The proof follows by slicing the bound in (Xu & Raginsky, 2017) and applying the rate-distortion argument from Theorem 2.3. Finally, we upper-bound the mutual information of a discrete random variable $q(W')$ by its discrete entropy, which is in turn upper bounded by the logarithm of the number of states it can take. □

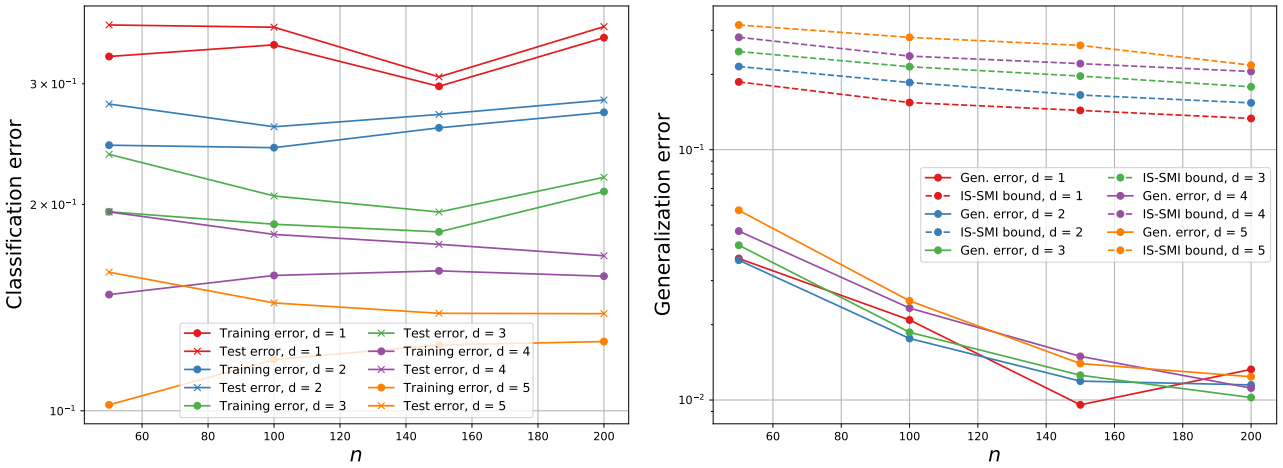

Figure 4. Empirical evaluation of *(left)* training/test classification errors, and *(right)* IS-SMI bound (6) against n for binary classification with logistic regression, with varying d. y-axis is in log scale.

C. Additional Experimental Details for Section 3

C.1. Gaussian mean estimation

Evaluating (6) involves estimating $\mathsf{SI}_k^{(1)}(W; Z_i)$. The expectation defining SMI (1) is in general intractable (Goldfeld et al., 2022), thus we approximate it with a Monte Carlo estimate. This amounts to evaluating $\mathsf{I}(\Theta^\top W; Z_i)$ for each sampled Θ: to this end, we use the analytical formula of MI between Gaussian random variables (Bu et al., 2019, Appendix B), since as shown in Appendix A.3, Z_i and $\Theta^\top W$ are both Gaussian.

C.2. Logistic regression

We use the binary classification problem as described in (Bu et al., 2019, Section VI). For each value of (n, d), we approximate the generalization error and its bound over 500 runs, using 50 random projections for each run.

Regarding the evaluation of our generalization bounds, MI is estimated via MINE (Belghazi et al., 2018) based on the following neural network architecture: we implement a fully-connected neural network with a single hidden layer with dimension 100. We train for 200 epochs using the Adam optimizer with a batch-size of 64 and learning rate of 0.001.

We report additional errors for this experiment on Figure 4.

C.3. Neural networks

For our neural network experiment, we use the Iris dataset (Fisher, 1936), which contains measured properties from three species of Iris flowers with 50 samples from each species. Each data point has four features (sepal length in cm, sepal width in cm, petal length in cm, and petal width in cm) and the task consists of predicting the correct species. We use a two-hidden-layer NN with $10,903$ parameters; specifically, each inner layer has a hidden dimension of 100. For each d-value $(5, 10, 15, 20, 50, 100)$ we sample 20 projection matrices Θ, and for each such Θ, we train 500 randomly initialized NNs for 200 epochs. This amounts to training $10,000$ NNs. In addition, we use the Adam optimizer (Kingma & Ba, 2017) with a batch size of 64 and a learning rate of 0.1.

We estimate MI using MINE (Belghazi et al., 2018). For MI estimation via MINE, we use a fully-connected neural network with one hidden layer of dimension 100. We train for 200 epochs using the Adam optimizer with a batch-size of 64 and learning rate of 0.001.