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. Yet, these bounds have limited practicality for modern ML applications (e.g., deep learning), due to the difficulty of evaluating MI in high dimensions. Motivated by recent findings on the compressibility of neural networks, we consider algorithms that operate by *slicing* the parameter space, i.e., trained on random lower-dimensional subspaces. We introduce new, tighter information-theoretic generalization bounds tailored for such algorithms, demonstrating that slicing improves generalization. Our bounds offer significant computational and statistical advantages over standard MI bounds, as they rely on scalable alternative measures of dependence, i.e., disintegrated mutual information and k-sliced mutual information. Then, we extend our analysis to algorithms whose parameters do not need to exactly lie on random subspaces, by leveraging rate-distortion theory. This strategy yields generalization bounds that incorporate a distortion term measuring model compressibility under slicing, thereby tightening existing bounds without compromising performance or requiring model compression. Building on this, we propose a regularization scheme enabling practitioners to control generalization through compressibility. Finally, we empirically validate our results and achieve the computation of non-vacuous information-theoretic generalization bounds for neural networks, a task that was previously out of reach.

[1]MIT [2]MIT-IBM Watson AI Lab; IBM Research. Correspondence to: Kimia Nadjahi <kimia.nadjahi@ens.fr>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Generalization is a fundamental aspect of machine learning, where models optimized on training data are expected to perform similarly on test data. Neural networks (NNs), in particular, are able to both perform and generalize well, allowing them to achieve excellent test performance on complex tasks. Despite this empirical success, the architectural factors influencing how well NNs generalize are not fully understood theoretically. This has motivated a substantial body of work using a variety of tools to bound their *generalization error* (Jiang et al., 2020b), i.e.., the gap between the average loss on training data (*empirical risk*) and its expected loss on a new data *from the same distribution* (*population risk*). The goal is to identify when and why a given model yields a low generalization error, and ultimately, design architectures or training algorithms that guarantee good generalization. Common approaches include PAC-Bayes analysis (Dziugaite & Roy, 2017) and information theory (Xu & Raginsky, 2017).

Compression is another topic which has provided a fertile ground for machine learning research. As model architectures have become more and more complex, their evaluation, training and fine-tuning become even more challenging. For instance, large language models are parameterized with billions of parameters. Compressed models, which reduce the number of trainable parameters without significantly deteriorating the performance, have been growing in practical relevance, for example LoRA finetuning of large language models (Hu et al., 2021). One compression scheme which has found success consists in training NNs on random, lower-dimensional subspaces. NNs compressed that way have been shown to yield satisfying test performances in various tasks while being faster to train (Li et al., 2018). This framework has recently 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.

A recent line of work argues that there is actually an interplay between *compressible* models and their ability to generalize well. The main conclusion is that one can construct tighter generalization bounds by leveraging compression schemes (Arora et al., 2018; Hsu et al., 2021; Kuhn et al., 2021; Sefidgaran et al., 2022).

In this paper, we seek further understanding on the generalization ability of learning algorithms trained on random subspaces. We introduce new information-theoretic generalization bounds tailored for such algorithms, which are tighter than existing ones. Our bounds demonstrate that algorithms that are "compressible" via random slicing have significantly better information-theoretic generalization guarantees. We also find an intriguing connection to Sliced Mutual Information (Goldfeld & Greenewald, 2021; Goldfeld et al., 2022), which we explore in learning problems where the information-theoretic generalization bounds are analytically computable. We then leverage the computational and statistical benefits of our sliced approach to empirically compute nonvacuous information-theoretic generalization bounds for various neural networks.

We further increase the practicality of our approach by using the *rate-distortion*-based framework introduced by Sefidgaran et al. (2022) to extend our bounds to the setting where the weight vector W only approximately lies on random subspace. This extension applies when the loss is Lipschitz w.r.t. the weights, which we can promote using techniques by Béthune et al. (2024). As Sefidgaran et al. (2022) did for quantization, this allows us to apply generalization bounds based on projection and quantization to networks whose weights are unrestricted. We tighten the bound by using regularization in training to encourage the weights to be close to the random subspace.

2. Related Work

Compression of neural networks. Our work focuses on random projection and quantization (Hubara et al., 2016) as tools for compressing neural networks. Many other compression approaches exist (Cheng et al., 2017), e.g., pruning (Dong et al., 2017; Blalock et al., 2020), low-rank compression (Wen et al., 2017), and optimizing architectures via neural architecture search and meta-learning (Pham et al., 2018; Cai et al., 2020; Finn et al., 2017). Further exploring alternative compression approaches from an information-theoretic generalization bound perspective is an interesting avenue for future work.

Compressibility and generalization. A body of work has emerged leveraging various notions of compressibility to adequately explain why neural networks can generalize (Arora et al., 2018; Suzuki et al., 2020; Simsekli et al., 2020; Bu et al., 2021; Hsu et al., 2021; Kuhn et al., 2021; Sefidgaran et al., 2022). In particular, Bu et al. (2021) and Sefidgaran et al. (2022) connected compressibility and generalization using the rate-distortion theory, which inspired our analysis. Sefidgaran et al. (2022) derived a set of theoretical generalization bounds, but their applicability to compressible neural networks is unclear. Bu et al. (2021) established a generalization bound for a learning model whose weights W are optimized and then compressed into $\hat{W}$. They consider compression as a post-processing technique, while we propose to take compressibility into account during training. Furthermore, Bu et al. (2021) applied the rate-distortion theory for a slightly different purpose than ours: to compare the population risk of the compressed model with that of the original model. In contrast, we use it to bound the generalization error of the original model and show that if it is almost compressible on a random subspace, one can obtain significantly tighter bounds than existing information-theoretic ones.

Conditional MI generalization bounds. Following (Xu & Raginsky, 2017) and (Bu et al., 2019), which treat the training data as random, (Steinke & Zakynthinou, 2020) instead obtain a bound where the dataset is fixed (i.e. *conditioned* on a dataset). This framework assumes that two independent datasets are available, and random Bernoulli indicator variables create a random training set by randomly selecting which of the two datasets to use for the ith training point. This approach has the advantage of creating a generalization bound involving the mutual information between the learned weights and a set of *discrete* random variables, in which case the mutual information is always finite. Connections to other generalization bound strategies and to data privacy are established by (Steinke & Zakynthinou, 2020). Followup works tightened these bounds by considering the conditional mutual information between the indicator variables and either the *predictions* (Harutyunyan et al., 2021; Haghifam et al., 2022) or *loss* (Wang & Mao, 2023) of the learned model rather than the weights. A practical limitation of this general approach is that it requires a second dataset (or *supersample*) to compute the conditional mutual information, whereas this extra data could be used to get a better estimate of the test error (hence, the generalization error) directly. Additionally, some of these bounds depend on a mutual information term between low-dimensional variables (e.g., functional CMI-based bounds (Harutyunyan et al., 2021)), which can be evaluated efficiently but does not inform practitioners for selecting model architectures. Exploring slicing for the conditional MI framework is beyond the scope of our paper and is a promising direction for future work.

Other generalization bounds for neural networks. Beyond the information-theoretic frameworks above, many methods bound the generalization of neural networks. Classic approaches in learning theory bound generalization error with complexity of the hypothesis class (Bartlett & Mendelson, 2002; Vapnik & Chervonenkis, 2015), but these fail to explain the generalization ability of highly flexible deep neural network models (Zhang et al., 2017). More successful approaches include the PAC-Bayes framework (including Lotfi et al., whose use of slicing inspired our work), margin-based approaches (Koltchinskii et al., 2002; Kuznetsov et al., 2015; Chuang et al., 2021), flatness of the loss curve (Petzka et al., 2021), and even empirically-trained prediction not

based on theoretical guarantees (Jiang et al., 2020a; Lassance et al., 2020; Natekar & Sharma, 2020; Schiff et al., 2021). Each approach has its own benefits and drawbacks; for instance, many of the tightest predictions are highly data-driven and as a result may provide limited insight into the underlying sources of generalization and how to design networks to promote it.

Our work. Our approach dramatically improves the tightness of *input-output information-theoretic generalization bounds* for neural networks, which up to this point have not seen practical use. That said, our bounds (unsurprisingly) are still looser than generalization bounds available through some other frameworks, particularly those employing additional data (e.g., data-driven PAC-Bayes priors (Lotfi et al., 2022) or the super-sample of conditional MI bounds (Wang & Mao, 2023)) or involving some kind of trained or ad hoc prediction function. Regardless, continuing to improve information-theoretic bounds is a fruitful endeavor that improves our understanding of the connection between machine learning and information theory, and gives insights that can drive algorithmic and architectural innovation.

3. Preliminaries

Let Z be the input data space, $\mathrm{W} \subseteq \mathbb{R}^D$ the hypothesis space, and $\ell : \mathrm{W} \times \mathrm{Z} \to \mathbb{R}_+$ a loss function. For instance, in supervised learning, $\mathrm{Z} = \{(x, y) \in \mathrm{X} \times \{-1, 1\}\}$ is the set of feature-label pairs, $w \in \mathrm{W}$ is the parameter vector of a classifier $f_w : \mathbb{R}^D \to \{-1, 1\}$ (e.g., the weights of a neural network), and $\ell(w, (x, y)) = \mathbf{1}_{y \neq f_w(x)}$ is the error made by predicting y as $f_w(x)$.

Consider a training dataset $S_n \triangleq (Z_1, \dots, Z_n) \in \mathrm{Z}^n$ consisting of n i.i.d. samples from μ. For any $w \in \mathrm{W}$, let $\mathcal{R}(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$ denote the *population risk*, and $\widehat{\mathcal{R}}_n(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, z_i)$ the *empirical risk*. Training a machine learning algorithm aims at minimizing the population risk, i.e., solving $\min_{w\in\mathrm{W}} \mathcal{R}(w)$. However, computing $\mathcal{R}(w)$ is difficult in practice, since the data distribution μ is generally unknown: one would only observe a finite number of samples from μ. Therefore, a common workaround is *empirical risk minimization*, i.e., $\min_{w\in\mathrm{W}} \widehat{\mathcal{R}}_n(w)$. A learning algorithm can then be described as the mapping $\mathcal{A} : \mathrm{Z}^n \to \mathrm{W}$, where $\mathcal{A}(S_n)$ is a hypothesis learned from S_n. We assume that $\mathcal{A}$ is *randomized*: its output $W \triangleq \mathcal{A}(S_n)$ is a random variable distributed from $P_{W|S_n}$.

The *generalization error* of $\mathcal{A}$ is defined as $\mathrm{gen}(\mu, \mathcal{A}) \triangleq \mathbb{E}_{P_{W|S_n}\otimes\mu^{\otimes n}}[\mathcal{R}(W) - \widehat{\mathcal{R}}_n(W)]$, where the expectation $\mathbb{E}$ is taken with respect to (w.r.t.) the joint distribution of (W, S_n). The higher $\mathrm{gen}(\mu, \mathcal{A})$, the more $\mathcal{A}$ overfits when trained on $S_n \sim \mu^{\otimes n}$.

3.1. Information-theoretic generalization bounds

In recent years, there has been a flurry of interest in using theoretical approaches to bound $\mathrm{gen}(\mu, \mathcal{A})$ using *mutual information* (MI). The most common information-theoretic bound on generalization error was introduced by Xu & Raginsky (2017) and depends on the mutual information between the training data S_n and the hypothesis W learned from S_n. We recall the formal statement below.

Theorem 3.1 (Xu & Raginsky, 2017)**.** *Assume that $\ell(w, Z)$ is σ-sub-Gaussian[1] under $Z \sim \mu$ for all $w \in \mathrm{W}$. Then, $|\mathrm{gen}(\mu, \mathcal{A})| \leq \sqrt{2\sigma^2\, \mathsf{I}(W; S_n)/n}$, where $\mathsf{I}(W; S_n)$ is the mutual information between $W = \mathcal{A}(S_n)$ and S_n.*

The class of σ-sub-Gaussian losses includes Gaussian-distributed losses $\ell(w, Z) \sim \mathcal{N}(0, \sigma^2)$ and bounded losses satisfying $0 \leq \ell(w, Z) \leq 2\sigma$. For example, the 0-1 correct classification loss satisfies this with $\sigma = 0.5$. Subsequently, Bu et al. (2019) used the averaging structure of the empirical loss to derive a bound that depends on $\mathsf{I}(W; Z_i)$. Evaluating MI on each *individual* data point Z_i rather than the entire training dataset S_n has been shown to produce tighter bounds than Xu & Raginsky (2017) (Bu et al., 2019, §IV).

Theorem 3.2 (Bu et al., 2019)**.** *Assume that $\ell(\tilde{W}, \tilde{Z})$ is σ-sub-Gaussian under $(\tilde{W}, \tilde{Z}) \sim P_W \otimes \mu$. Then, $|\mathrm{gen}(\mu, \mathcal{A})| \leq (1/n)\sum_{i=1}^n \sqrt{2\sigma^2\, \mathsf{I}(W; Z_i)}$, where $\mathsf{I}(W; Z_i)$ is the mutual information between $W = \mathcal{A}(S_n)$ and Z_i.*

These and other information-theoretic bounds, however, suffer from the fact that the dimension of W can be large when using modern ML models, e.g. NNs. Indeed, the sample complexity of MI estimation scales poorly with dimension (Paninski, 2003). Collecting more samples of (W, Z_i) can be expensive, especially with NNs, as one realization of $W \sim P_{W|S_n}$ requires one complete training run. Moreover, McAllester & Stratos (2020) recently proved that estimating MI from finite data have important statistical limitations when the underlying MI is large, e.g., hundreds of bits.

3.2. Random subspace training and sliced mutual information

While modern neural networks use large numbers of parameters, common 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 d-dimensional subspace spanned by a random matrix (with $d \ll D$) not only provides computational advantages, but does not meaningfully damage the performance of the neural network, for appropriate choice of d (often two orders of magnitude smaller than D). They interpreted this fact as indicating *compressibil-*

[1] A random variable X is σ-sub-Gaussian ($\sigma > 0$) under μ if for $t \in \mathbb{R}$, $\mathbb{E}_\mu[e^{t(X-\mathbb{E}_\mu[X])}] \leq e^{\sigma^2 t^2/2}$.

ity of the neural network architecture up to some *intrinsic dimension* d, below which performance degrades.

Denote by $\mathrm{St}(d, D) = \{\Theta \in \mathbb{R}^{D\times d} : \Theta^\top\Theta = \boldsymbol{I}_d\}$ the Stiefel manifold, equipped with the uniform distribution P_Θ. We consider a learning algorithm $\mathcal{A}^{(d)}$ whose hypothesis space is restricted to $\mathrm{W}_{\Theta,d} \triangleq \{w \in \mathbb{R}^D : \exists w' \in \mathbb{R}^d \text{ s.t. } w = \Theta w'\}$, where $\Theta \sim P_\Theta$. Note that Θ is *not* trained: it is randomly generated from P_Θ at the beginning of training, and frozen during training. In other words, $\mathcal{A}^{(d)}$ trains the parameters on a random d-dimensional subspace of $\mathbb{R}^D$ characterized by Θ.

Sliced mutual information. The random d-dimensional weight subspace is closely related to *slicing*, which projects a high dimensional quantity to a random lower dimensional subspace. Intriguingly, a recent line of work has considered slicing mutual information, yielding significant sample complexity and computational advantages in high-dimensional regimes. Goldfeld & Greenewald (2021) and Goldfeld et al. (2022) slice the arguments of MI via random k-dimensional projections, thus defining the *k-Sliced Mutual Information* (SMI). SI_k has been shown to retain many important properties of MI (Goldfeld et al., 2022), and 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. This provides significant advantages over MI, whose computation generally requires an exponential number of samples in $\max(d_x, d_y)$ (Paninski, 2003). Similar convergence rates can be achieved while slicing in only one dimension. Recently, Wongso et al. (2023) empirically connected generalization to SMI between the true class labels Y and the hidden representations T of NNs.

4. Information-Theoretic Generalization Bounds for Compressed Models

Motivated by the advantageous properties of sliced mutual information and the practical success of training neural networks in random subspaces, we seek an input-output information-theoretic generalization bound for learning algorithms trained on the random subspace $\mathrm{W}_{\Theta,d}$. We will see that improved tightness and statistical properties of such bounds allow these to scale to larger models than possible for the traditional bounds in Theorems 3.1 and 3.2, without significantly damaging the test-time performance of the resulting learned models. To this end, we will derive in this section new information-theoretic bounds on the generalization error for these random subspace algorithms. Using the terminology in Section 3, the generalization error of $\mathcal{A}^{(d)}$ is

$$\mathrm{gen}(\mu, \mathcal{A}^{(d)}) = \mathbb{E}_{P_{W'|\Theta,S_n}\otimes P_\Theta\otimes\mu^{\otimes n}}[\mathcal{R}(\Theta W') - \widehat{\mathcal{R}}_n(\Theta W')]\,. \tag{1}$$

Note that the expectation is taken w.r.t. to P_Θ, so the error does not depend on Θ.

A natural strategy to bound the generalization error of $\mathcal{A}^{(d)}$ is by applying Xu & Raginsky (2017): if $\ell(w, Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all $(\Theta, w) \in \mathrm{St}(d, D) \times \mathrm{W}_{\Theta,d}$, then by Theorem 3.1,

$$|\mathrm{gen}(\mu, \mathcal{A}^{(d)})| \le \sqrt{\frac{2\sigma^2}{n}\mathsf{I}(\Theta W'; S_n)}\,, \tag{2}$$

where $\Theta W' = \mathcal{A}^{(d)}(S_n)$, $\Theta \sim P_\Theta$. However, this bound does not clearly explain the impact of the intrinsic dimension d on generalization. In addition, the MI term $\mathsf{I}(\Theta W'; S_n)$ is hard to estimate in modern machine learning applications since $\Theta W'$ is high-dimensional.

We derive new upper-bounds on $\mathrm{gen}(\mu, \mathcal{A}^{(d)})$ to mitigate these issues. Our strategy consists in applying the *disintegration technique* on the hypothesis space $\mathrm{W}_{\Theta,d}$. In our setting, disintegration boils down to deriving a bound for a fixed Θ, then taking the expectation over P_Θ. This yields information-theoretic bounds which are tighter than existing ones and rely on mild assumptions. Moreover, our bounds exhibit an explicit dependence on d, which helps capture the impact of compressing the hypothesis space on generalization. Finally, their evaluation is computationally more friendly as it requires estimating MI between lower-dimensional variables. Specifically, our bounds depend on the alternative information theory measure called *disintegrated mutual information* (Negrea et al., 2019, Definition 1.1). The disintegrated MI between X and Y given U is defined as $\mathsf{I}^U(X;Y) = \mathbf{KL}(P_{X,Y|U}\|P_{X|U}\otimes P_{Y|U})$, where $\mathbf{KL}$ denotes the Kullback-Leibler divergence and $P_{X,Y|U}$ the conditional distribution of (X, Y) given U, $P_{X|U}$ (respectively, $P_{Y|U}$) the conditional distribution of X (resp., Y) given U.

4.1. A first bound on $\mathrm{gen}(\mu, \mathcal{A}^{(d)})$

We first bound $\mathrm{gen}(\mu, \mathcal{A}^{(d)})$ by *disintegrating* the proof of Theorem 3.1 (Xu & Raginsky, 2017).

Theorem 4.1. *Assume for all $w' \in \mathbb{R}^d$ and $\Theta \in \mathrm{St}(d, D)$, $\ell(\Theta w', Z)$ is σ_Θ-sub-Gaussian under $Z \sim \mu$, where σ_Θ is a positive constant which may depend on Θ. Then,*

$$|\mathrm{gen}(\mu, \mathcal{A}^{(d)})| \le \sqrt{\frac{2}{n}}\,\mathbb{E}_{P_\Theta}\left[\sqrt{\sigma_\Theta^2\mathsf{I}^\Theta(W'; S_n)}\right]. \tag{3}$$

Note that since $W = \Theta W'$ implies $W' = \Theta^T W$, $\mathsf{I}^\Theta(W'; S_n)$ has significant parallels with the sliced mutual information (Goldfeld et al., 2022) with slicing in the first argument only, denoted $\mathsf{SI}_k^{(1)}(W; S_n)$. Sliced mutual information, however, is formulated with W being independent of Θ, which is not generally true in our setting (except in specific regimes such as the Gaussian mean estimation example below).

Theorem 4.1 holds under a sub-Gaussianity assumption, which is slightly different than the one in Theorem 3.1.

It is immediate that the assumption in Theorem 3.1 implies the assumption in Theorem 4.1 with $\sigma_\Theta = \sigma$. For instance, consider the supervised learning setting, where for all $w = \Theta w' \in \mathrm{W}_{\Theta,d}$, $\ell(\Theta w', z) = \mathbf{1}_{f_{\Theta w'}(x) \neq y} \leq 1$. Then, by Hoeffding's lemma, (2) and (3) both hold, with $\sigma_\Theta = \sigma = 2$. Conversely, if the assumption in Theorem 4.1 holds, then the assumption in Theorem 3.1 is met with $\sigma = \sup_{\Theta \in \mathrm{St}(d,D)} \sigma_\Theta$.

Our bound entails notable advantages over Xu & Raginsky (2017). First, (3) is tighter than (2), since $\mathbb{E}_{P_\Theta}\left[\sqrt{\mathsf{I}^\Theta(W';S_n)}\right] \leq \sqrt{\mathsf{I}(\Theta W';S_n)}$ (see Appendix A.3). This is a natural consequence of the proof technique of Theorem 4.1, which leverages disintegration, a strategy known to yield tighter characterizations of concave generalization bounds by Jensen's inequality (Hellström et al., 2023, §4.3).

Then, our bound is more tractable in regimes where (2) is fundamentally intractable, e.g., when D is very large. Indeed, common estimators for $\mathsf{I}^\Theta(W';S_n)$ exhibit faster convergence rates than $\mathsf{I}(\Theta W';S_n)$ because W' has a lower dimension than $\Theta W'$ ($d \ll D$). For instance, the theoretical guarantees of MINE (Belghazi et al., 2018) clearly support the favorable computational and statistical properties of our bounds. By (Goldfeld et al., 2022, Theorem 2), the approximation error induced by MINE decays rapidly as the dimensionality decreases. Additionally, some of the assumptions of MINE can be relaxed, allowing optimization over a larger class of distribution in lower-dimensional spaces (Goldfeld et al., 2022, Remark 6). One may argue that our bound requires computing the expectation of $\mathsf{I}^\Theta(W';S_n)$ over $\Theta \sim P_\Theta$, which makes its evaluation expensive. However, in our experiments, we estimated this expectation with a Monte Carlo approximation and found that increasing the number of samples of Θ had little practical impact on our bounds. This is consistent with prior work (Li et al., 2018), which showed that test performance remains relatively stable across multiple values of Θ for a fixed d, while the choice of d has a greater impact on the quality of solutions.

4.2. A tighter bound via individual samples

One limitation of Theorem 4.1 is that $\mathsf{I}^\Theta(W';S_n) = +\infty$ if $P_{W',S_n|\Theta}$ is not absolutely continuous w.r.t $P_{W'|\Theta} \otimes \mu^{\otimes n}$, therefore the bound is vacuous. For instance, this happens when $W' = g(S_n)$ where g is a smooth, non-constant deterministic function that may depend on Θ. To overcome this issue, we combine disintegration with the *individual-sample* technique introduced by Bu et al. (2019). This allows us to construct a bound in terms of the *individual-sample disintegrated MI*, $\mathsf{I}^\Theta(W';Z_i)$.

Theorem 4.2. *Assume that* (i) *for all* $w' \in \mathbb{R}^d$ *and* $\Theta \in \mathrm{St}(d,D)$, $\ell(\Theta w', Z)$ *is* σ_Θ*-sub-Gaussian under* $Z \sim \mu$, *where* σ_Θ *is a positive constant which may depend on* Θ*; or* (ii) *for all* $\Theta \in \mathrm{St}(d,D)$, $\ell(\Theta \tilde{W}', \tilde{Z})$ *is* σ_Θ*-sub-Gaussian under* $(\tilde{W}', \tilde{Z}) \sim P_{W'|\Theta} \otimes \mu$. *Then,*

$$|\mathrm{gen}(\mu, \mathcal{A}^{(d)})| \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_\Theta}\left[\sqrt{2\sigma_\Theta^2 \mathsf{I}^\Theta(W';Z_i)}\right]. \quad (4)$$

Assumption *(i)* is not stronger than *(ii)* (e.g., one can adapt (Bu et al., 2019, Remark 1)), and conversely (e.g., see Gaussian mean estimation in Section 4.3). Note that Theorem 4.2 under assumption *(ii)* is a particular case of a more general theorem, which we present in Appendix A.1 for readability purposes. A key advantage of Theorem A.2 is its broader applicability as compared to Theorems 4.1 and 4.2. We will illustrate this in Section 4.3 on linear regression, where the loss is not sub-Gaussian.

Thanks to the individual-sample technique, our bound in (4) is no worse than the one in Theorem 4.1. Indeed, assumption *(i)* in Theorem 4.2 is the same as the one in Theorem 4.1, and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_\Theta}\left[\sqrt{2\sigma_\Theta^2 \mathsf{I}^\Theta(W';Z_i)}\right] \leq \sqrt{\frac{2}{n}}\mathbb{E}_{P_\Theta}\left[\sqrt{\sigma_\Theta^2 \mathsf{I}^\Theta(W';S_n)}\right]$ (Proposition A.6). In particular, if W' is a deterministic function of S_n, the bound in (4) can be non-vacuous as opposed to (3).

Thanks to disintegration, the bound in Theorem 4.2 is no worse than the one in Theorem 3.2, i.e., $|\mathrm{gen}(\mu, \mathcal{A}^{(d)})| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 \mathsf{I}(\Theta W';Z_i)}$. This is justified using similar arguments as in Section 4.1 to compare the sub-Gaussian conditions and MI terms: see Proposition A.5. Interestingly, we show in Section 4.3 that Theorem 3.2 cannot be applied in a simple learning problem, because its assumptions do not trivially hold. In contrast, assumption *(ii)* of Theorem 4.1 is easily verified because we condition on Θ. This illustrates that, in addition to providing tighter bounds, disintegration helps formulate alternative sub-Gaussianity assumptions that can be milder.

4.3. Applications

We illustrate more concretely the benefits of our findings over Xu & Raginsky (2017) and Bu et al. (2019) in terms of tightness and applicability. Moreover, we uncover an interesting connection with the Sliced MI evaluated on $W = \mathcal{A}(S_n)$ and Z_i where only W is projected, i.e., $\mathsf{SI}_d^{(1)}(W;Z_i) = \mathbb{E}_{P_\Theta}\left[\mathsf{I}^\Theta(\Theta^\top W;Z_i)\right]$.

Countable hypothesis space. Our generalization bounds provide a clear explanation on why algorithms with low intrinsic dimension d are likely to generalize well. Indeed, suppose that for any $\Theta \in \mathrm{St}(d,D)$ and $w = \Theta w' \in \mathrm{W}_{\Theta,d}$, we have $\|w'\| \leq b_\Theta$, where $\|\cdot\|$ is the Euclidean norm. Then, using the same argumentation as in (Xu & Raginsky, 2017, §4.1) in Theorem 4.1,

$$|\mathrm{gen}(\mu, \mathcal{A}^{(d)})| \leq \sqrt{\frac{2d}{n}}\mathbb{E}_{P_\Theta}\left[\sqrt{\sigma_\Theta^2 \log(2b_\Theta\sqrt{dn})}\right]. \quad (5)$$

We make two key observations based on (5). First, the right-hand side (RHS) term decreases as d shrinks. This confirms that algorithms trained on lower-dimensional random subspace tend to generalize better, as observed in practice (Li et al., 2018). Then, our bound can help guide architecture choices for practitioners who wish to control the generalization error, thanks to the explicit dependence on d, n and b_Θ. Note that given Θ, achieving $\|w'\| \leq b_\Theta$ can easily be achieved in practice by quantizing w'.

Gaussian mean estimation. We now study the following problem inspired by (Bu et al., 2019, Section IV.A). The training dataset $S_n = (Z_1, \dots, Z_n)$ consists of n i.i.d. samples from $\mathcal{N}(\mathbf{0}_D, \boldsymbol{I}_D)$, where $\mathbf{0}_D \in \mathbb{R}^D$ is the zero vector. The objective function is $\widehat{\mathcal{R}}_n(w) \triangleq \frac{1}{n}\sum_{i=1}^n \|w - Z_i\|^2$. Consider the models $\mathcal{A}$ and $\mathcal{A}^{(d)}$ which minimize $\widehat{\mathcal{R}}_n(w)$ on $\mathrm{W} = \mathbb{R}^D$ and $\mathrm{W}_{\Theta,d}$ respectively. Then, $\mathcal{A}(S_n) = \bar{Z}$ and $\mathcal{A}^{(d)}(S_n) = \Theta\Theta^\top \bar{Z}$, where $\bar{Z} \triangleq \frac{1}{n}\sum_{i=1}^n Z_i$. Since W' is a deterministic function of S_n given Θ, applying Theorem 4.1 would give a vacuous bound. Instead, we apply Theorem 4.2 and obtain

$$\mathrm{gen}(\mu, \mathcal{A}^{(d)}) \leq C_{D,d,n} \sum_{i=1}^n \mathbb{E}_{P_\Theta}\left[\sqrt{\mathsf{I}^\Theta(\Theta^\top \bar{Z}; Z_i)}\right], \quad (6)$$

with $C_{D,d,n} \triangleq \frac{2}{n}\sqrt{d\left(1+\frac{1}{n}\right)^2 + (D-d)}$. The detailed derivations are in Appendix A.4. For a fixed pair (D, n), $C_{D,d,n}$ increases as d goes to D. By adapting the proof of (Goldfeld et al., 2022, Proposition 2.2), one can show that $\mathbb{E}_{P_\Theta}\left[\sqrt{\mathsf{I}^\Theta(\Theta^\top \bar{Z}; Z_i)}\right]$ decreases as $d \to 0$, and by the data-processing inequality, for any $d \leq D$ and $\Theta \sim P_\Theta$, $\mathsf{I}^\Theta(\Theta^\top \bar{Z}; Z_i) \leq \mathsf{I}(\bar{Z}; Z_i)$. Therefore, the RHS term in (6) accurately captures that compressing the hypothesis space improves generalization. Note that here, $\mathsf{I}^\Theta(\Theta^\top \bar{Z}; Z_i)$ can actually be computed in closed form: since Z_i and $\Theta^T \bar{Z}$ (given Θ) are Gaussian random variables, and $\Theta^\top\Theta = \boldsymbol{I}_d$, we have $\mathsf{I}^\Theta(\Theta^\top \bar{Z}; Z_i) = \frac{d}{2}\log(\frac{n}{n-1})$. We also show that the bound is sub-optimal, since it scales in $\mathcal{O}(1/\sqrt{n})$ as $n \to +\infty$, and $\mathrm{gen}(\mu, \mathcal{A}^{(d)}) = 2d/n$.

This example shows that our findings allow us to derive generalization bounds where prior work does not apply. Indeed, for $d = D$, our bound boils down to the one derived by Bu et al. (2019); but when $d < D$, their strategy cannot be applied: the distribution of $\ell(\Theta W', Z) = ||\Theta W' - Z||^2$ is unknown, making it highly non-trivial to verify the sub-Gaussian condition by Bu et al. (2019) (in particular, $\Theta W'$ is not Gaussian). We overcome this issue via disintegration, i.e., by conditioning on Θ: we prove in Appendix A.4 that $\ell(\Theta\tilde{W}', \tilde{Z}) = \|\Theta\tilde{W}' - \tilde{Z}\|^2$ is sub-Gaussian *given* Θ (i.e., under $(\tilde{W}', \tilde{Z}) \sim P_{W'|\Theta} \otimes \mu$), which allows the application of Theorem 4.2. Finally, by Jensen's inequality and $\bar{Z} = W$, $\mathbb{E}_{P_\Theta}\left[\sqrt{\mathsf{I}^\Theta(\Theta^\top \bar{Z}; Z_i)}\right] \leq \sqrt{\mathsf{SI}_d^{(1)}(W; Z_i)}$, thus our bound is controlled by SMI. We report $\mathrm{gen}(\mu, \mathcal{A}^{(d)})$ and this bound in Figure 1. In Appendix A.5, we apply our theorem on linear regression and obtain a bound that also relies on SMI.

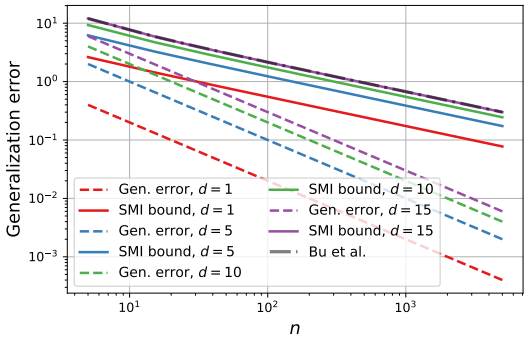


Figure 1: Gaussian mean estimation: generalization error and bound against n, for $D = 15$, $d \in \{1, 5, 10, 15\}$. Errors and bounds decrease as $d \to 1$. The bound in Bu et al. (2019) can only be applied for $d = D$. The scale is log-log.

5. Information-Theoretic Generalization Bounds for Compressible Models

The above bounds require the learned weights W to lie in $\mathrm{W}_{\Theta,d}$. When d is very small, this constraint can be restrictive and significantly deteriorate the performance of the model, as we illustrate in Section 6. That said, since our MI-based bounds generally increase with increasing d, it is important to keep d small. Motivated by recent work applying rate-distortion theory to input-output MI generalization bounds (Sefidgaran et al., 2022), we establish the following result for *approximately compressible* weights and Lipschitz loss.

Theorem 5.1. *Consider $\mathcal{A} : \mathrm{Z}^n \to \mathrm{W}$ s.t. $\mathcal{A}$ may take $\Theta \sim P_\Theta$ into account to output W. Assume there exists $C > 0$ s.t. $\ell(\tilde{W}, \tilde{Z}) \leq C$ almost surely. Assume for any $z \in \mathrm{Z}$, $\ell(\cdot, z) : \mathrm{W} \to \mathbb{R}_+$ is L-Lipschitz, i.e., $\forall (w_1, w_2) \in \mathrm{W} \times \mathrm{W}, |\ell(w_1, z) - \ell(w_2, z)| \leq L\rho(w_1, w_2)$, where ρ is a metric on W. Then,*

$$|\mathrm{gen}(\mu, \mathcal{A})| \leq 2L\mathbb{E}_{P_{W|\Theta}\otimes P_\Theta}\left[\rho(W, \Theta\Theta^\top W)\right] + \frac{C}{n}\sum_{i=1}^n \mathbb{E}_{P_\Theta}\left[\sqrt{\frac{\mathsf{I}^\Theta(\Theta^\top W; Z_i)}{2}}\right]. \quad (7)$$

This result shows a trade-off between distortion and generalization, aligning with prior work on generalization through the rate-distortion theory (e.g., Bu et al. (2021), Sefidgaran et al. (2022); see Section 2 for a detailed comparison). In the limit case where $d = D$, we retrieve the bound by Xu & Raginsky (2017).

The proof of Theorem 5.1 consists in considering two models $\mathcal{A} : \mathrm{Z}^n \to \mathbb{R}^D$ and $\mathcal{A}' : \mathrm{Z}^n \to \mathrm{W}_{\Theta,d}$ such that $\mathcal{A}(S_n) = W$ may depend on $\Theta \sim P_\Theta$, and $\mathcal{A}'(S_n) =$

$\Theta\Theta^\top W$. We then use the triangle inequality to obtain $|\mathrm{gen}(\mu, \mathcal{A})| \leq |\mathrm{gen}(\mu, \mathcal{A}) - \mathrm{gen}(\mu, \mathcal{A}')| + |\mathrm{gen}(\mu, \mathcal{A}')|$. Finally, we bound the first term (the *distortion* term) using the Lipschitz condition and the second term (the *rate* term) using Theorem 4.2.

Using similar arguments and Theorem 4.1, we derive a second rate-distortion bound based on quantization, which does not require estimating MI.

Theorem 5.2. *Assume the conditions of Theorem 5.1 hold. Furthermore, suppose that* $\|\Theta^\top W\| \leq M$ *for* $(W, \Theta) \sim P_{W|\Theta} \otimes P_\Theta$. *Consider a function* $\mathcal{Q}$ *quantizing* $\Theta^\top W$ *such that* $\rho\big(\Theta^\top W, \mathcal{Q}(\Theta^\top W)\big) \leq \delta$. *Then,*

$$|\mathrm{gen}(\mu, \mathcal{A})| \leq 2L\mathbb{E}_{P_{W|\Theta}\otimes P_\Theta}\big[\rho\big(W, \Theta\mathcal{Q}(\Theta^\top W)\big)\big] + C\mathbb{E}_{P_\Theta}\left[\sqrt{\frac{\mathsf{I}^\Theta(\mathcal{Q}(\Theta^\top W); S_n)}{2n}}\right] \tag{8}$$

$$\leq 2L\left(\mathbb{E}_{P_{W|\Theta}\otimes P_\Theta}\big[\rho(W, \Theta\Theta^\top W)\big] + \delta\right) \tag{9}$$

$$+ C\sqrt{\frac{d\log(2M\sqrt{d}/\delta)}{2n}}\,. \tag{10}$$

Note that $\|\Theta^\top W\| \leq M$ is a mild assumption, since in general, this is a result of enforcing Lipschitz continuity (e.g., the Lipschitz neural networks studied by Béthune et al. (2024) require weights with bounded norms). We will set $\delta = 1/\sqrt{n}$ to reflect the fact that training on more samples reduces the generalization error.

The MI term in (7) or (8) is evaluated between the training data S_n and a low-dimensional, potentially quantized projection of W. This makes our rate-distortion bounds simpler to estimate than standard information-theoretic ones that rely on $\mathsf{I}(S_n; W)$. Our bound in (9)-(10) further reduces the computational complexity by bounding the MI term in (8) using similar arguments to those in (Xu & Raginsky, 2017, §4.1). It should be viewed as an interpretable and easily computable alternative bound that supports our main message: almost-compressibility on random subspaces implies better generalization. Indeed, given that the term in (10) increases with increasing d, a tighter generalization bound can be achieved for $d < D$, provided that the corresponding distortion in (9) (which measures the rate of compressibility) is sufficiently small. Practitioners also do not need to quantize the weights to evaluate (9)-(10), making the process computationally more efficient: assuming the existence of a quantizer $\mathcal{Q}$ as described in Theorem 5.2 is sufficient.

Our theoretical findings provide concrete guidelines on how to tighten the generalization error bounds in practice. First, the value of the Lipschitz constant L can be directly controlled through the design of the neural network, as we explain in Section 6 and Appendix B.2. The term $\mathbb{E}_{P_{W|\Theta}\otimes P_\Theta}\big[\rho(W, \Theta\Theta^\top W)\big]$ can be regularized by simply adding a penalty term $\lambda\mathbb{E}_{P_{W|\Theta}\otimes P_\Theta}\big[\rho(W, \Theta\Theta^\top W)\big]$ to the training objective. Depending on the value of the hyperparameter λ, this regularization can encourage solutions to be close to the subspace $\mathrm{W}_{\Theta,d}$, i.e., having low distortion from the compressed weights. The choice of d is also important and can be tuned to balance the MI term with the distortion required (how small λ needs to be) to achieve low training error. Indeed, choosing a higher λ increases the importance of the regularization term, effectively reducing the importance of the empirical risk. Hence, the empirical risk may rise, which in most cases will increase the training error.

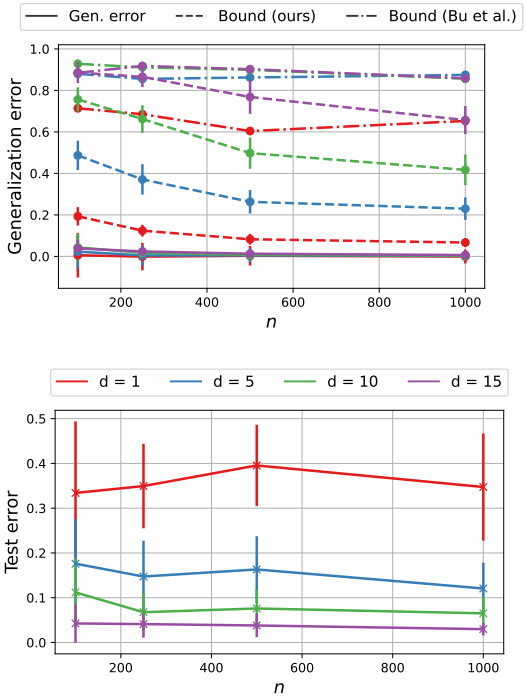


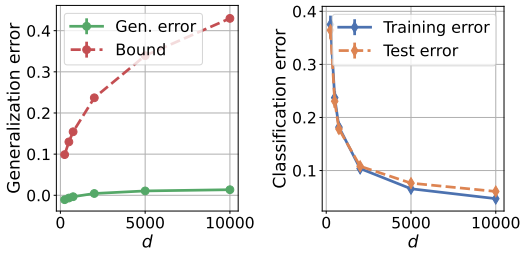
Figure 2: Illustration of our bound (4) and (Bu et al., 2019) on binary classification of Gaussian data of dimension 20 with logistic regression trained on $\mathrm{W}_{\Theta,d}$

6. Empirical Analysis

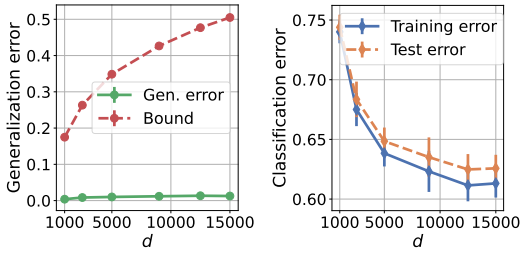
To illustrate our findings and their practical impact, we train several neural networks for classification, and evaluate their generalization error and our bounds. This requires compressing NNs (via random slicing and quantization) and estimating MI. We explain our methodology below, and refer to Appendix C.1 for more details and results. We provide the code to reproduce the experiments[2].

Random projections. To sample $\Theta \in \mathbb{R}^{D\times d}$ such that $\Theta^\top\Theta = \boldsymbol{I}_d$, we construct an orthonormal basis using the singular value decomposition of a random matrix $\Gamma \in \mathbb{R}^{D\times d}$

[2]Code is available here: https://github.com/kimiandj/slicing_mi_generalization



(a) MNIST



(b) CIFAR-10

Figure 3: Generalization bounds with NNs for image classification. The weights are projected and quantized.

whose entries are i.i.d. from $\mathcal{N}(0,1)$. Since the produced matrix Θ is dense, the projection $\Theta^\top w$ induces a runtime of $\mathcal{O}(dD)$. To improve scalability, we use the sparse projector by Li et al. (2018) and the Kronecker product projector by Lotfi et al. (2022), which compute $\Theta^\top w$ in $\mathcal{O}(d\sqrt{D})$ and $\mathcal{O}(\sqrt{dD})$ operations respectively, and require storing only $\mathcal{O}(d\sqrt{D})$ and $\mathcal{O}(\sqrt{dD})$ matrix elements respectively.

Quantization. We use the quantizer by Lotfi et al. (2022), which simultaneously learns quantized weights W' and quantized levels $(c_1, \cdots, c_L)$. This allows us to highly compress NNs and bypass the estimation of MI: for any $\Theta \sim P_\Theta$, $\mathsf{I}^\Theta(W';S_n) \leq H^\Theta(W') \leq \lceil d \times H(p) \rceil + L \times (16 + \lceil \log_2 d \rceil) + 2$, where $H^\Theta(W')$ denotes the entropy of W' given Θ, and $H(p) \triangleq -\sum_{l=1}^{L} p_l \log(p_l)$ is the entropy of the quantized level (p_l is the empirical probability of c_l).

Estimating MI. In our practical settings, the MI terms arising in the generalization bounds cannot be computed exactly, so we resort to two popular estimators: the k-nearest neighbor estimator (kNN-MI, Kraskov et al., 2004) and MINE (Belghazi et al., 2018). We obtain NaN values with kNN-MI for $d > 2$ thus only report the bounds estimated with MINE. In our experiments, the use of MINE was not a practical issue because d had low to relatively high values.

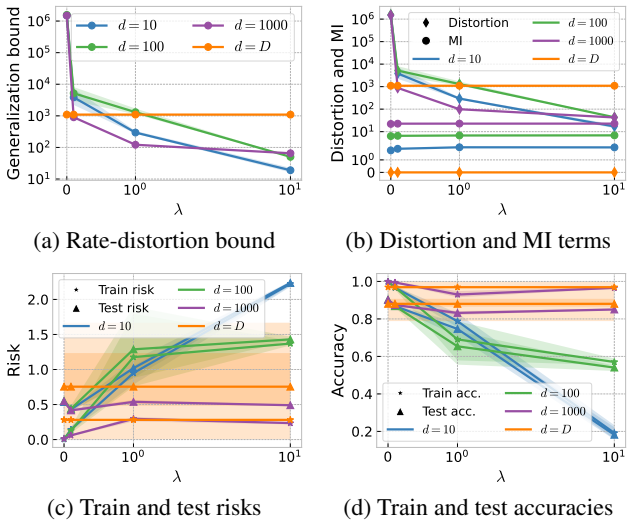
6.1. Generalization bounds for models trained on $\mathrm{W}_{\Theta,d}$

Binary classification with logistic regression. We consider the same setting as Bu et al. (2019, §VI): each data point $Z = (X, Y)$ consist of features $X \in \mathbb{R}^s$ and labels $Y \in \{0,1\}$, Y is uniformly distributed in $\{0,1\}$, and $X|Y \sim \mathcal{N}(\mu_Y, 4\boldsymbol{I}_s)$ with $\mu_0 = (-1,\ldots,-1)$ and $\mu_1 = (1,\ldots,1)$. We use a linear classifier and evaluate the generalization error based on the loss function $\ell(w,z) = \mathbf{1}_{\hat{y}\neq y}$, where $\hat{y}$ is the prediction of input x defined as $\hat{y} \triangleq \mathbf{1}_{\bar{w}^T x + w_0 \geq 0}$, $\forall w = (\bar{w}, w_0) \in \mathbb{R}^{s+1}$. We train a logistic regression on $\mathrm{W}_{\Theta,d}$ and estimate the generalization error. Since ℓ is bounded by $C = 1$, we approximate the generalization error bound from Theorem 4.2 for $d < D$, and Bu et al. (2019, Prop. 1) for $d = D$. Figure 2 reports the results for $s = 20$ and different values of n and d: we observe that our bound holds and accurately reflects the behavior of the generalization error against (n, d). Our methodology also provides tighter bounds than Bu et al. (2019), and the difference increases with decreasing d. On the other hand, the lower d, the lower generalization error and its bound, but the higher the test risk (Figure 2). This is consistent with prior empirical studies (Li et al., 2018) and explained by the fact that lower values of d induce a more restrictive hypothesis space, thus make the model less expressive.

Multiclass classification with NNs. Next, we evaluate our generalization error bounds for neural networks trained on image classification. Denote by $f(w,x) \in \mathbb{R}^K$ the output of the NN parameterized by w given an input image x, with $K > 1$ the number of classes. The loss is $\ell(w,z) = \mathbf{1}_{\hat{y}\neq y}$, with $\hat{y} = \arg\max_{i\in\{1,\ldots,K\}} [f(w,x)]_i$. We train fully-connected NNs to classify MNIST and CIFAR-10 datasets, with $D = 199\,210$ and $D = 656\,810$ respectively: implementation details are given in Appendix C.2. Even though we significantly decrease the dimension of the parameters by slicing, d can still be quite high thus obtaining an accurate estimation of $\mathsf{I}^\Theta(W';S_n)$ remains costly. To mitigate this issue, in addition to slicing, we discretize W' with the quantizer by Lotfi et al. (2022) and evaluate Theorem 4.1 with $\mathsf{I}^\Theta(W';S_n)$ replaced by $\lceil d \times H(p) \rceil + L \times (16 + \lceil \log_2 d \rceil) + 2$, as discussed at the beginning of Section 6. Our results in Figure 3 illustrate that employing slicing followed by quantization enables the computation of non-vacuous generalization bounds for NNs, while still maintaining test performance for adequate values of d (which is consistent with Li et al. (2018)). We point out that in these high-dimensional problems, slicing is unequivocally the key to making information-theoretic bounds possible to estimate: quantization alone is far from sufficient in this regime. For example, even binary quantization of D weights would yield 2^D states, requiring an unimaginably large number of samples to accurately estimate the mutual information term. Additional results on MNIST and Iris datasets are given in Appendix C.2.

6.2. Rate-distortion bounds

We evaluate our rate-distortion generalization bounds for neural networks trained on image classification. For $q \in \mathbb{N}^*$, let $f : \mathrm{W} \times \mathrm{X} \to \mathbb{R}^K$ be a q-layer feedforward network with ReLU activations. We train f using a slightly modified cross-entropy loss defined for $w \in \mathrm{W}$ and $z = (x,y) \in \mathrm{X} \times \{1,\ldots,K\}$ as $\ell(w,z) =$



(a) Rate-distortion bound (b) Distortion and MI terms

(c) Train and test risks (d) Train and test accuracies

Figure 4: Generalization errors and rate-distortion bounds for feedforward NNs trained on MNIST. Results are averaged over 5 runs. Shaded areas represent the 2.5% and 97.5% percentiles. For each run, expectations are computed with Monte Carlo estimates over 5 samples of Θ.

$-\log(\hat{p}(w,x)_y)$, where $\hat{p}(w,x) = \max(p(w,x), p_{\min})$, $p(w,x) = e^{f(w,x)}/\mathbf{1}^\top e^{f(w,x)} \in (0,1]^K$ and $p_{\min} > 0$. This loss is bounded from above by $-\log(p_{\min})$ and for any z, $\ell(\cdot, z)$ is Lipschitz-continuous with constant $\sqrt{2}$. Assuming that the weight matrix of each layer has bounded spectral norm ($\forall i \in \{1,\dots,q\}$, $\|W^{(i)}\|_2 \leq M$), we show that Theorem 5.1 or 5.2 applies with the distortion $\rho(W, \Theta\Theta^\top W) = \sum_{i=1}^q \|W^{(i)} - (\Theta\Theta^\top W)^{(i)}\|_2$ and the constants C and L specified in Theorem B.2.

We train a 3-layer feedforward NN f to classify MNIST. The first two layers each contain 1000 neurons and the final layer has $K = 10$ neurons, thus the total number of parameters is $D = 1000 \cdot (784 + 1000 + 10) = 1\,794\,000$. Our goal is to evaluate the generalization error and its rate-distortion bound in Theorem 5.2 for different values of subspace dimension d and regularization coefficient λ. To this end, we parameterize f with $w = \Theta(\Theta^\top w_1) + \bar{\Theta}(\bar{\Theta}^\top w_2) \in \mathbb{R}^D$, where $\Theta \in \mathrm{St}(d, D)$ is randomly generated at initialization and $\bar{\Theta} \in \mathbb{R}^{D\times(D-d)}$ is such that $[\Theta, \bar{\Theta}] \in \mathbb{R}^{D\times D}$ forms an orthogonal basis of $\mathbb{R}^D$. At each run, we train f on a random subset of MNIST with $n = 1000$ samples for 5 different samples of Θ. We set $p_{\min} = 1\text{e-}4$. To estimate the generalization error, we approximate the population risk on a test dataset of 10 000 samples. Our results clearly depict the interplay between λ and d and their impact on the generalization error, bound and risk. Specifically, a higher λ makes our model more compressible by encouraging its parameters to lie on the d-dimensional subspace characterized by Θ (see our discussion in Section 5). This has two main consequences, as predicted by our theory and illustrated in our plots. First, a higher λ leads to a lower generalization error (Figure 4c) and a tighter rate-distortion bound, the distortion term being smaller (Figures 4a and 4b). Second, as λ increases and d decreases, the train/test risk become higher (Figure 4c). This is consistent with Li et al. (2018), as this regime effectively reduces to training on a low-dimensional random subspace. To further demonstrate the trade-off between high compressibility/low generalization error and high train/test error, we also plot the distortion (i.e., $2L(\mathbb{E}[\rho(W, \Theta\Theta^T W)] + 1/\sqrt{n})$) and MI term ($C\sqrt{d\log(2M\sqrt{dn})/(2n)}$) against λ, for different values of d (Figure 4b). Additionally, we plot the test and train accuracies vs. λ and d (Figure 4d). We observe there exist combinations of λ and d that yield tighter generalization error bound while inducing satisfactory training and test errors, e.g., $(\lambda, d) = (10, 1000)$. This suggests that with carefully chosen λ and d, our methodology can tighten generalization bounds while preserving model performance. This is a significant step towards practical relevance of information-theoretic bounds: to our knowledge, such bounds applied to neural networks have been fundamentally intractable and pessimistic, thus lacking practical use beyond very small toy examples (we refer to Section 2 for an expanded discussion). In contrast, by taking into account almost-compressibility via random slicing and quantization, we can derive bounds that are much easier to compute, and develop a theoretically-grounded regularization scheme to effectively control the generalization error in practice.

7. Conclusion

In this work, we combined recent empirical compression methods for learning models, such as NNs, with generalization bounds based on input-output MI. Our results indicate that random slicing is a very interesting scheme, as it is easy to implement, performs well, and is highly suitable for practically computable and tighter information-theoretic bounds. We also explore a notion of approximate compressibility, i.e., rate-distortion, where the learned parameters are close to a quantization of the compressed subspace but do not lie on it exactly. This framework provides more flexibility, enabling the model to maintain good training error even with a smaller subspace dimension d, while ensuring that the resulting generalization bounds are as tight as possible and permitting the use of analytical bounds on the MI instead of difficult-to-compute MI estimates. Our rate-distortion approach also motivated a weight regularization technique to make trained NNs as approximately compressible as possible and ensure that our bound is small in practice. Future work includes a more detailed exploration of strategies for using our bounds to help inform selection and design of NN architectures, and exploring bounds and regularizers based on other successful compression methods.

Acknowledgements

Kimia Nadjahi acknowledges the generous support from the MIT Postdoctoral Fellowship Program for Engineering Excellence. The MIT Geometric Data Processing Group acknowledges the generous support of Army Research Office grants W911NF2010168 and W911NF2110293, from the CSAIL Systems that Learn program, from the MIT–IBM Watson AI Laboratory, from the Toyota–CSAIL Joint Research Center, and from an Amazon Research Award. Kimia Nadjahi expresses gratitude to the organizers of the ICML 2023 workshop "Neural Compression: From Information Theory to Applications", where an early version of this work was accepted for an oral presentation and sparked fruitful discussions with the attendees.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

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.

Béthune, L., Massena, T., Boissin, T., Bellet, A., Mamalet, F., Prudent, Y., Friedrich, C., Serrurier, M., and Vigouroux, D. DP-SGD Without Clipping: The Lipschitz Neural Network Way. In *The Twelfth International Conference on Learning Representations*, 2024.

Blalock, D., Gonzalez Ortiz, J. J., Frankle, J., and Guttag, J. What is the state of neural network pruning? In Dhillon, I., Papailiopoulos, D., and Sze, V. (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 129–146, 2020.

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.

Bu, Y., Gao, W., Zou, S., and Veeravalli, V. V. Population risk improvement with model compression: An information-theoretic approach. *Entropy*, 23(10), 2021. ISSN 1099-4300. doi: 10.3390/e23101255.

Cai, H., Gan, C., Wang, T., Zhang, Z., and Han, S. Once for all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations*, 2020.

Cheng, Y., Wang, D., Zhou, P., and Zhang, T. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

Chuang, C.-Y., Mroueh, Y., Greenewald, K., Torralba, A., and Jegelka, S. Measuring generalization with optimal transport. *Advances in Neural Information Processing Systems*, 34:8294–8306, 2021.

Dong, X., Huang, J., Yang, Y., and Yan, S. More is less: A more complicated network with less inference complexity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5840–5848, 2017.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2017.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1126–1135. JMLR, 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. In *Advances in Neural Information Processing Systems*, 2022.

Haghifam, M., Moran, S., Roy, D. M., and Karolina Dziugiate, G. Understanding generalization via leave-one-out conditional mutual information. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2487–2492, 2022. doi: 10.1109/ISIT50566.2022.9834400.

Harutyunyan, H., Raginsky, M., Ver Steeg, G., and Galstyan, A. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems*, 34:24670–24682, 2021.

Hellström, F., Durisi, G., Guedj, B., and Raginsky, M. Generalization bounds: Perspectives from information theory and pac-bayes, 09 2023.

Hsu, D., Ji, Z., Telgarsky, M., and Wang, L. Generalization bounds via distillation. *arXiv preprint arXiv:2104.05641*, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., and Bengio, Y. Binarized neural networks. In *Advances in neural information processing systems*, pp. 4107–4115, 2016.

Jiang, Y., Foret, P., Yak, S., Roy, D. M., Mobahi, H., Dziugaite, G. K., Bengio, S., Gunasekar, S., Guyon, I., and Neyshabur, B. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*, 2020a.

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*, 2020b.

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

Koltchinskii, V., Panchenko, D., et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of statistics*, 30(1):1–50, 2002.

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

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kuhn, L., Lyle, C., Gomez, A. N., Rothfuss, J., and Gal, Y. Robustness to pruning predicts generalization in deep neural networks. *arXiv preprint arXiv:2103.06002*, 2021.

Kuznetsov, V., Mohri, M., and Syed, U. Rademacher complexity margin bounds for learning with a large number of classes. In *ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels*, 2015.

Lassance, C., Béthune, L., Bontonou, M., Hamidouche, M., and Gripon, V. Ranking Deep Learning Generalization using Label Variation in Latent Geometry Graphs. *arXiv preprint arXiv:2011.12737*, 2020.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

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.

McAllester, D. and Stratos, K. Formal limitations on the measurement of mutual information. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 26–28 Aug 2020.

Natekar, P. and Sharma, M. Representation based complexity measures for predicting generalization in deep learning. *arXiv preprint arXiv:2012.02775*, 2020.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

Petzka, H., Kamp, M., Adilova, L., Sminchisescu, C., and Boley, M. Relative flatness and generalization. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Pham, H., Guan, M., Zoph, B., Le, Q., and Dean, J. Efficient neural architecture search via parameters sharing. In *International Conference on Machine Learning*, pp. 4095–4104, 2018.

Schiff, Y., Quanz, B., Das, P., and Chen, P.-Y. Gi and pal scores: Deep neural network generalization statistics. *arXiv preprint arXiv:2104.03469*, 2021.

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.

Simsekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

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

Suzuki, T., Abe, H., and Nishimura, T. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer, 2015.

Wang, Z. and Mao, Y. Tighter Information-Theoretic Generalization Bounds from Supersamples. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Wen, W., Xu, C., Wu, C., Wang, Y., Chen, Y., and Li, H. Coordinating filters for faster deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 658–666, 2017.

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.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

A. Postponed Proofs for Section 4

Notation. $\mathcal{R}^{\Theta}(w') = \mathbb{E}_{Z\sim\mu}[\ell^{\Theta}(w', Z)]$ and $\widehat{\mathcal{R}}_n^{\Theta}(w') \triangleq \frac{1}{n}\sum_{i=1}^n \ell^{\Theta}(w', z_i)$, $\forall w = \Theta w' \in \mathrm{W}_{\Theta,d}$ and $\ell^{\Theta}(w', z) \triangleq \ell(\Theta w', z)$. The generalization error of $\mathcal{A}^{(d)}$ is $\mathrm{gen}(\mu, \mathcal{A}^{(d)}) = \mathbb{E}[\mathcal{R}^{\Theta}(W') - \widehat{\mathcal{R}}_n^{\Theta}(W')]$ with the expectation taken over $P_{W'|\Theta,S_n} \otimes P_{\Theta} \otimes \mu^{\otimes n}$.

A.1. Proof of Theorem A.2

Consider three random variables $X \in \mathrm{X}$, $Y \in \mathrm{Y}$ and $U \in \mathrm{U}$. Denote by $P_{X,Y,U}$ their joint distribution and by P_X, P_Y, P_U the marginals. Let $\tilde{X}$ (respectively, $\tilde{Y}$) be an independent copy of X (resp., Y) with joint distribution $P_{\tilde{X},\tilde{Y}} = P_{\tilde{X}} \otimes P_{\tilde{Y}}$. Given U, let $f^U : \mathrm{X} \times \mathrm{Y} \to \mathbb{R}$ be a mapping parameterized by U, and denote by $K_{f^U(\tilde{X},\tilde{Y})}$ the cumulant generating function of $f^U(\tilde{X}, \tilde{Y})$, i.e. for $t \in \mathbb{R}$,

$$K_{f^U(\tilde{X},\tilde{Y})}(t) = \log \mathbb{E}\big[e^{t(f^U(\tilde{X},\tilde{Y}) - \mathbb{E}[f^U(\tilde{X},\tilde{Y})])}\big] \tag{11}$$

where the expectations are taken w.r.t. $P_{X|U} \otimes P_{Y|U}$.

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

$$\mathbb{E}_{P_{X,Y,U}}[f^U(X,Y)] - \mathbb{E}_{P_{\tilde{X},\tilde{Y},U}}[f^U(\tilde{X},\tilde{Y})] \le \mathbb{E}_{P_U}\left[\inf_{t\in[0,b_+)} \frac{\mathsf{I}^U(X;Y) + \psi_+(t,U)}{t}\right]. \tag{12}$$

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

$$\mathbb{E}_{P_{\tilde{X},\tilde{Y},U}}[f^U(\tilde{X},\tilde{Y})] - \mathbb{E}_{P_{X,Y,U}}[f^U(X,Y)] \le \mathbb{E}_{P_U}\left[\inf_{t\in[0,-b_-)} \frac{\mathsf{I}^U(X;Y) + \psi_-(t,U)}{t}\right]. \tag{13}$$

Proof. Let $U \sim P_U$. By Donsker-Varadhan variational representation,

$$\mathsf{I}^U(X;Y) = \mathbf{KL}(P_{(X,Y)|U} \| P_{X|U} \otimes P_{Y|U}) \tag{14}$$

$$= \sup_{g\in\mathcal{G}^U} \mathbb{E}_{P_{(X,Y)|U}}[g^U(X,Y)] - \log \mathbb{E}_{P_{X|U}\otimes P_{Y|U}}[e^{g^U(\tilde{X},\tilde{Y})}] \tag{15}$$

where $\mathcal{G}^U \triangleq \{g^U : \mathrm{X} \times \mathrm{Y} \to \mathbb{R} \text{ s.t. } \mathbb{E}_{P_{X|U}\otimes P_{Y|U}}[e^{g^U(\tilde{X},\tilde{Y})}] < \infty\}$. Therefore, for any $t \in [0, b_+)$,

$$\mathbf{KL}(P_{(X,Y)|U} \| P_{X|U} \otimes P_{Y|U}) \ge t\mathbb{E}[f^U(X,Y)] - \log \mathbb{E}[e^{tf^U(\tilde{X},\tilde{Y})}] \tag{16}$$

$$\ge t\left(\mathbb{E}[f^U(X,Y)] - \mathbb{E}[f^U(\tilde{X},\tilde{Y})]\right) - \psi_+(t,U) \tag{17}$$

where (17) follows from assuming that for $t \in [0, b_+)$, $K_{f^U(\tilde{X},\tilde{Y})}(t) \le \psi_+(t, U)$. Hence,

$$\mathbb{E}[f^U(X,Y)] - \mathbb{E}[f^U(\tilde{X},\tilde{Y})] \le \inf_{t\in[0,b_+)} \frac{\mathsf{I}^U(X;Y) + \psi_+(t,U)}{t}. \tag{18}$$

We obtain the final result (12) by taking the expectation of (18) over P_U.

We can prove analogously that (13) holds, assuming for $t \in [0, b_-)$, $K_{f^U(\tilde{X},\tilde{Y})}(t) \le \psi_-(-t, U)$.

□

Theorem A.2. *Assume that for $\Theta \sim P_\Theta$, there exists $C_- \in \mathbb{R}_+^* \cup \{+\infty\}$ s.t. for $t \in (C_-, 0]$, $K_{\ell^\Theta(\tilde{W}',\tilde{Z})}(t) \le \psi_-(-t, \Theta)$, where $\psi_-(\cdot, \Theta)$ is convex and $\psi_-(0, \Theta) = \psi_-'(0, \Theta) = 0$. Then,*

$$\mathrm{gen}(\mu, \mathcal{A}^{(d)}) \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{P_\Theta}\left[\inf_{t\in[0,-C_-)} \frac{\mathsf{I}^\Theta(W';Z_i) + \psi_-(t,\Theta)}{t}\right]. \tag{19}$$

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

$$\operatorname{gen}(\mu, \mathcal{A}^{(d)}) \geq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{P_\Theta} \left[\inf_{t \in [0, C_+)} \frac{\mathsf{I}^\Theta(W'; Z_i) + \psi_+(t, \Theta)}{t} \right] . \tag{20}$$

Proof of Theorem A.2. The generalization error of $\mathcal{A}^{(d)}$ can be written as

$$\operatorname{gen}(\mu, \mathcal{A}^{(d)}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}_{P_{W'|\Theta} \otimes P_\Theta \otimes \mu}[\ell^\Theta(\tilde{W}', \tilde{Z}_i)] - \mathbb{E}_{P_{W'|\Theta, Z_i} \otimes P_\Theta \otimes \mu}[\ell^\Theta(W', Z_i)] \right\} . \tag{21}$$

Our final bounds (20) and (19) result from applying Lemma A.1 on each term of the sum in (21), i.e. with $X = W'$, $Y = Z_i$ and $f^U(X, Y) = \ell^\Theta(W', Z_i)$.

□

A.2. Applications of Theorem A.2

We specify Theorem A.2 under different sub-Gaussian conditions on the loss. A random variable X is said to be σ-sub-Gaussian (with $\sigma > 0$) if for any $t \in \mathbb{R}$,

$$\mathbb{E}[e^{t(X - \mathbb{E}[X])}] \leq e^{\sigma^2 t^2 / 2} . \tag{22}$$

Proof of Theorem 4.1. Define $h^\Theta(w', s) = (1/n) \sum_{i=1}^n \ell^\Theta(w', z_i)$ for $w' \in \mathbb{R}^d$, $s = (z_1, \ldots, z_n) \in \mathrm{Z}^n$ and $\Theta \in \mathbb{R}^{D \times d}$ s.t. $\Theta^\top \Theta = \boldsymbol{I}_d$. The generalization error of $\mathcal{A}^{(d)}$ can be written as,

$$\operatorname{gen}(\mu, \mathcal{A}^{(d)}) = \mathbb{E}_{P_{W'|\Theta} \otimes P_\Theta \otimes \mu^{\otimes n}} \left[h^\Theta(\tilde{W}', \tilde{S}_n) \right] - \mathbb{E}_{P_{W'|Z_i, \Theta} \otimes P_\Theta \otimes \mu^{\otimes n}} \left[h^\Theta(W', S_n) \right] . \tag{23}$$

Since we assume that $\ell^\Theta(w', Z)$ is σ-sub-Gaussian under $Z \sim \mu$ for all w' and Θ, and $Z_1, \ldots, Z_n$ are i.i.d, then $h^\Theta(w', S_n)$ is $\sigma/\sqrt{n}$-sub-Gaussian under $S_n \sim \mu^{\otimes n}$ for all w' and Θ. Therefore, $h^\Theta(\tilde{W}', \tilde{S}_n)$ is $\sigma/\sqrt{n}$-sub-Gaussian under $(\tilde{W}', S_n) \sim P_{W'|\Theta} \otimes \mu^{\otimes n}$ for all Θ, and for $t \in \mathbb{R}$,

$$K_{h^\Theta(\tilde{W}', \tilde{S}_n)}(t) \leq \frac{\sigma^2 t^2}{2n} . \tag{24}$$

We conclude by applying Lemma A.1 with $X = W'$, $Y = S_n$, $U = \Theta$ and $f^U(X, Y) = h^\Theta(W', S_n)$, and the fact that,

$$\inf_{t > 0} \frac{\mathsf{I}^\Theta(W'; S_n) + \sigma^2 t^2 / (2n)}{t} = \sqrt{\frac{2\sigma^2}{n} \mathsf{I}^\Theta(W'; S_n)} . \tag{25}$$

□

Proof of Theorem 4.2. Let $\Theta \in \mathbb{R}^{D \times d}$ s.t. $\Theta^\top \Theta = \boldsymbol{I}_d$. Since $\ell^\Theta(\tilde{W}', \tilde{Z})$ is σ_Θ-sub-Gaussian under $(\tilde{W}', \tilde{Z}) \sim P_{W'} \otimes \mu$, then for any $t \in \mathbb{R}$, $K_{\ell^\Theta(\tilde{W}', \tilde{Z})}(t) \leq \sigma_\Theta^2 t^2 / 2$. We conclude by applying Theorem A.2 and the fact that for $i \in \{1, \ldots, n\}$,

$$\inf_{t > 0} \frac{\mathsf{I}^\Theta(W'; Z_i) + \sigma_\Theta^2 t^2 / 2}{t} = \sqrt{2\sigma_\Theta^2 \mathsf{I}^\Theta(W'; Z_i)} . \tag{26}$$

□

Corollary A.3. *Assume that for any $\Theta \sim P_\Theta$, $\ell^\Theta(\tilde{W}', \tilde{Z}) \leq C$ almost surely. Then,*

$$|\operatorname{gen}(\mu, \mathcal{A}^{(d)})| \leq \frac{C}{n} \sum_{i=1}^{n} \mathbb{E}_{P_\Theta} \left[\sqrt{\frac{\mathsf{I}^\Theta(W'; Z_i)}{2}} \right] . \tag{27}$$

Proof of Corollary A.3. Since for any $\Theta \sim P_\Theta$, $\ell^\Theta(\tilde{W}', \tilde{Z}) \leq C$ almost surely, then by Hoeffding's lemma, we have for all $t \in \mathbb{R}$,

$$\mathbb{E}_{P_{W'|\Theta} \otimes \mu}\left[e^{t\{\ell^\Theta(\tilde{W}', \tilde{Z}) - \mathbb{E}_{P_{W'|\Theta} \otimes \mu}[\ell^\Theta(\tilde{W}', \tilde{Z})]\}}\right] \leq e^{C^2 t^2/8} . \tag{28}$$

Therefore, $K_{\ell^\Theta(\tilde{W}', \tilde{Z})}(t) \leq C^2 t^2/8$. We conclude by applying Lemma A.1 and the fact that for $i \in \{1, ..., n\}$,

$$\inf_{t>0} \frac{\mathsf{I}^\Theta(W'; Z_i) + C^2 t^2/8}{t} = C\sqrt{\frac{\mathsf{I}^\Theta(W'; Z_i)}{2}} . \tag{29}$$

□

A.3. Tightness of our generalization bounds

Proposition A.4. *For any concave and non-decreasing function $\phi : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}_{P_\Theta}\left[\phi\big(\mathsf{I}^\Theta(W'; S_n)\big)\right] \leq \phi\big(\mathsf{I}(W; S_n)\big) . \tag{30}$$

Proof of Proposition A.4. Let $W \in \mathrm{W}_{\Theta,d}$. Then, $S_n \to (W', \Theta) \to W$ and $S_n \to W \to (W', \Theta)$ form two Markov chains, so equality holds in the data-processing inequality, leading to $\mathsf{I}(W; S_n) = \mathsf{I}(W', \Theta; S_n)$.

By the chain rule of mutual information, and since Θ and S_n are independent,

$$\mathsf{I}(W', \Theta; S_n) = \mathsf{I}(\Theta; S_n) + \mathsf{I}(W'; S_n|\Theta) = \mathsf{I}(W'; S_n|\Theta) . \tag{31}$$

Since ϕ is non-decreasing,

$$\phi\big(\mathsf{I}(W', \Theta; S_n)\big) \geq \phi\big(\mathsf{I}(W'; S_n|\Theta)\big) \tag{32}$$

Applying the definition of conditional mutual information and Jensen's inequality yields,

$$\phi(\mathsf{I}(W'; S_n|\Theta)) = \phi(\mathbb{E}_{P_\Theta}\left[\mathsf{I}^\Theta(W'; S_n)\right]) \geq \mathbb{E}_{P_\Theta}\left[\phi(\mathsf{I}^\Theta(W'; S_n))\right] , \tag{33}$$

which concludes the proof.

□

Proposition A.5. *For any concave and non-decreasing function $\phi : \mathbb{R} \to \mathbb{R}$,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_\Theta}\left[\phi\big(\mathsf{I}^\Theta(W'; Z_i)\big)\right] \leq \frac{1}{n}\sum_{i=1}^{n}\left[\phi\big(\mathsf{I}(W; Z_i)\big)\right] . \tag{34}$$

Proof of Proposition A.5. Let $i \in \{1, \ldots, n\}$. By applying the same proof techniques of Proposition A.4 with Z_i instead of S_n, one has

$$\mathbb{E}_{P_\Theta}\left[\phi\big(\mathsf{I}^\Theta(W'; Z_i)\big)\right] \leq \phi\big(\mathsf{I}(W; Z_i)\big) . \tag{35}$$

The final result follows immediately.

□

Proposition A.6. *For any concave and non-decreasing function $\phi : \mathbb{R} \to \mathbb{R}$,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_\Theta}\left[\phi\big(\mathsf{I}^\Theta(W'; Z_i)\big)\right] \leq \mathbb{E}_{P_\Theta}\left[\phi\left(\frac{\mathsf{I}^\Theta(W'; S_n)}{n}\right)\right] . \tag{36}$$

Proof of Proposition A.6. By adapting the proof of (Bu et al., 2019, Proposition 2), one has

$$\frac{1}{n}\sum_{i=1}^{n} \phi\big(\mathsf{I}^\Theta(W'; Z_i)\big) \leq \phi\left(\frac{\mathsf{I}^\Theta(W'; S_n)}{n}\right) \tag{37}$$

The result follows immediately by taking the expectation of (37) and applying the linearity of the expectation on the left-hand side term.

□

A.4. Detailed derivations for Gaussian mean estimation

Problem statement. The loss function is defined for any $(w, z) \in \mathbb{R}^D \times \mathbb{R}^D$ as $\ell(w, x) = \|w - z\|^2$. Let $Z_1, \dots, Z_n$ be n random variables i.i.d. from $\mathcal{N}(\mathbf{0}_D, \boldsymbol{I}_D)$. Let $d \leq D$ and $\Theta \sim P_\Theta$ s.t. $\Theta^\top \Theta = \boldsymbol{I}_d$. Consider a model $\mathcal{A}^{(d)}$ whose objective is $\arg\min_{w \in \mathrm{W}_{\Theta,d}} \widehat{\mathcal{R}}_n(w)$ where the empirical risk is defined for $w \in \mathbb{R}^D$ as $\widehat{\mathcal{R}}_n(w) = \frac{1}{n}\sum_{i=1}^n \|w - Z_i\|^2$. This is equivalent to solving $\arg\min_{w' \in \mathbb{R}^d} \widehat{\mathcal{R}}_n^\Theta(w')$, where

$$\forall w' \in \mathbb{R}^d, \;\; \widehat{\mathcal{R}}_n^\Theta(w') = \frac{1}{n}\sum_{i=1}^n \|\Theta w' - Z_i\|^2 \,. \tag{38}$$

The gradient of (68) with respect to w' is,

$$\nabla_{w'} \widehat{\mathcal{R}}_n^\Theta(w) = \frac{2}{n}\sum_{i=1}^n \Theta^\top(\Theta w' - Z_i) \,, \tag{39}$$

and solving $\nabla_{w'} \widehat{\mathcal{R}}_n^\Theta(w) = 0$ yields $(\Theta^\top\Theta)w' = \Theta^\top \bar{Z}$ where $\bar{Z} \triangleq (1/n)\sum_{i=1}^n Z_i$. Since $\Theta^\top\Theta = \boldsymbol{I}_d$, we conclude that the minimizer of (68) is $W' = \Theta^\top \bar{Z}$.

Generalization error. We recall that the generalization error of $\mathcal{A}^{(d)}$ is defined as,

$$\mathrm{gen}(\mu, \mathcal{A}^{(d)}) = \mathbb{E}[\mathcal{R}^\Theta(W') - \widehat{\mathcal{R}}_n^\Theta(W')] \tag{40}$$

where the expectation is computed with respect to $P_{W'|\Theta,S_n} \otimes P_\Theta \otimes \mu^{\otimes n}$. Since $W' = \Theta^\top \bar{Z}$, $\mathrm{gen}(\mu, \mathcal{A}^{(d)})$ can be written as

$$\mathrm{gen}(\mu, \mathcal{A}^{(d)}) = \mathbb{E}_{(S_n,\Theta)\sim\mu^{\otimes n}\otimes P_\Theta}\left[\mathbb{E}_{\tilde{Z}\sim\mu}[\|\Theta\Theta^\top\bar{Z} - \tilde{Z}\|^2] - \frac{1}{n}\sum_{i=1}^n \|\Theta\Theta^\top\bar{Z} - Z_i\|^2\right] \tag{41}$$

Since $Z_1, \dots, Z_n$ are n i.i.d. samples from $\mathcal{N}(\mathbf{0}_D, \boldsymbol{I}_D)$ and $\Theta^\top\Theta = \boldsymbol{I}_d$, then $P_{\Theta^\top\bar{Z}|\Theta} = \mathcal{N}(\mathbf{0}_d, (1/n)\boldsymbol{I}_d)$ and we have

$$\begin{aligned}
\mathbb{E}_{\mu^{\otimes n}\otimes P_\Theta}[\|\Theta\Theta^\top\bar{Z}\|^2] &= \mathbb{E}_{\mu^{\otimes n}\otimes P_\Theta}[\mathrm{Tr}((\Theta\Theta^\top\bar{Z})^\top(\Theta\Theta^\top\bar{Z}))] &(42)\\
&= \mathbb{E}_{\mu^{\otimes n}\otimes P_\Theta}[\mathrm{Tr}(\bar{Z}^\top\Theta\Theta^\top\Theta\Theta^\top\bar{Z})] &(43)\\
&= \mathrm{Tr}(\mathbb{E}_{\mu^{\otimes n}\otimes P_\Theta}[\Theta^\top\bar{Z}(\Theta^\top\bar{Z})^\top]) &(44)\\
&= \frac{d}{n} \,. &(45)
\end{aligned}$$

For $i \in \{1, \dots, n\}$, $\mathbb{E}[\|Z_i\|^2] = \mathrm{Tr}(\mathbb{E}[Z_i Z_i^\top]) = D$, and

$$\begin{aligned}
\mathbb{E}[(\Theta\Theta^\top\bar{Z})^\top Z_i] &= \frac{1}{n}\sum_{j=1}^n \mathbb{E}[Z_j^\top\Theta\Theta^\top Z_i] &(46)\\
&= \frac{1}{n}\sum_{j=1}^n \mathrm{Tr}(\mathbb{E}[\Theta^\top Z_i(\Theta^\top Z_j)^\top]) &(47)\\
&= \frac{1}{n}\,\mathrm{Tr}(\mathbb{E}[\Theta^\top Z_i(\Theta^\top Z_i)^\top]) &(48)\\
&= \frac{d}{n} \,. &(49)
\end{aligned}$$

Equations (48) to (49) can be justified as follows. Since $Z_i \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D)$, the conditional distribution of $\Theta^\top Z_i$ given Θ is $\mathcal{N}(\mathbf{0}_d, \Theta^\top\Theta)$, and $\Theta^\top\Theta = \mathbf{I}_d$ by definition. Therefore, $\mathbb{E}[\Theta^\top Z_i(\Theta^\top Z_i)^\top] = \mathbb{E}[\mathbb{E}[\Theta^\top Z_i(\Theta^\top Z_i)^\top|\Theta]] = \mathbf{I}_d$. We conclude that $\mathrm{Tr}(\mathbb{E}[\Theta^\top Z_i(\Theta^\top Z_i)^\top]) = d$.

We thus obtain,

$$\mathbb{E}[\widehat{\mathcal{R}}_n^{\Theta}(W')] = \mathbb{E}_{(S_n,\Theta)\sim\mu^{\otimes n}\otimes P_\Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\|\Theta\Theta^\top\bar{Z} - Z_i\|^2\right] \tag{50}$$

$$= \mathbb{E}_{(S_n,\Theta)\sim\mu^{\otimes n}\otimes P_\Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\|\Theta\Theta^\top\bar{Z}\|^2 - 2(\Theta\Theta^\top\bar{Z})^\top Z_i + \|Z_i\|^2\right] \tag{51}$$

$$= D - \frac{d}{n}\,. \tag{52}$$

Indeed, by the linearity of expectation, (51) simplifies as

$$\mathbb{E}[\widehat{\mathcal{R}}_n^{\Theta}(W')] = \mathbb{E}_{\mu^{\otimes n}\otimes P_\Theta}[\|\Theta\Theta^\top\bar{Z}\|^2] - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{\mu^{\otimes n}\otimes P_\Theta}[(\Theta\Theta^\top\bar{Z})^\top Z_i] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mu}[\|Z_i\|^2] \tag{53}$$

Since $(Z_i)_{i=1}^n$ are i.i.d. from $\mathcal{N}(\mathbf{0}_D, \mathbf{I}_D)$, we proved that $\mathbb{E}_{\mu^{\otimes n}\otimes P_\Theta}[\|\Theta\Theta^\top\bar{Z}\|^2] = \frac{d}{n}$ (eq. (45)) and $\mathbb{E}[(\Theta\Theta^\top\bar{Z})^\top Z_i] = \frac{d}{n}$ (eq. (49)). Additionally,

$$\mathbb{E}_\mu[\|Z_i\|^2] = \mathbb{E}_\mu[\mathrm{Tr}(\|Z_i\|^2)] = \mathbb{E}_\mu[\mathrm{Tr}(Z_iZ_i^\top)] = \mathrm{Tr}(\mathbb{E}_\mu[Z_iZ_i^\top]) = \mathrm{Tr}(\mathbf{I}_D) = D \tag{54}$$

Plugging these identities in (53) yields (52).

On the other hand,

$$\mathbb{E}_{(S_n,\Theta,\tilde{Z})\sim\mu^{\otimes n}\otimes P_\Theta\otimes\mu}[(\Theta\Theta^\top\bar{Z})^\top\tilde{Z}] = \mathbb{E}[\Theta\Theta^\top\bar{Z}]^\top\mathbb{E}[\tilde{Z}] = 0\,, \tag{55}$$

therefore,

$$\mathbb{E}[\mathcal{R}^{\Theta}(W')] = \mathbb{E}_{(S_n,\Theta)\sim\mu^{\otimes n}\otimes P_\Theta}\mathbb{E}_{\tilde{Z}\sim\mu}[\|\Theta\Theta^\top\bar{Z} - \tilde{Z}\|^2] \tag{56}$$

$$= \mathbb{E}_{(S_n,\Theta)\sim\mu^{\otimes n}\otimes P_\Theta}\mathbb{E}_{\tilde{Z}\sim\mu}[\|\Theta\Theta^\top\bar{Z}\|^2 - 2(\Theta\Theta^\top\bar{Z})^\top\tilde{Z} + \|\tilde{Z}\|^2] \tag{57}$$

$$= D + \frac{d}{n}\,. \tag{58}$$

By plugging (52) and (58) in (41), we conclude that $\mathrm{gen}(\mu, \mathcal{A}^{(d)}) = 2d/n$.

Generalization error bound. We apply Theorem A.2 to bound the generalization error. To this end, we need to bound the cumulant generating function of $\ell^\Theta(\tilde{W}', \tilde{Z}) = \|\Theta\Theta^\top\bar{Z} - \tilde{Z}\|^2$ given Θ.

Since $(Z_1, \ldots, Z_n, \tilde{Z}) \sim \mu^{\otimes n} \otimes \mu$ with $\mu = \mathcal{N}(\mathbf{0}_D, \boldsymbol{I}_D)$, then, given Θ, one has $\Theta^\top\bar{Z} \sim \mathcal{N}(\mathbf{0}_d, (1/n)\boldsymbol{I}_d)$ and $(\Theta\Theta^\top\bar{Z} - \tilde{Z}) \sim \mathcal{N}(\mathbf{0}_D, \Sigma_\Theta)$ with $\Sigma_\Theta = \Theta\Theta^\top/n + \boldsymbol{I}_D$. Therefore, for $d < D$, $\ell^\Theta(\tilde{W}', \tilde{Z}) = \|\Theta\Theta^\top\bar{Z} - \tilde{Z}\|^2$ is the sum of squares of D dependent Gaussian random variables, which can equivalently be written as

$$\ell^\Theta(\tilde{W}', \tilde{Z}) = \sum_{k=1}^{D}\lambda_{\Theta,k}U_{\Theta,k}^2\,, \tag{59}$$

$$U_\Theta = P\Sigma_\Theta^{-1/2}(\Theta W' - \tilde{Z}) \tag{60}$$

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 $\Lambda = \mathrm{diag}(\lambda_\Theta)$. As a consequence, $U_\Theta \sim \mathcal{N}(\mathbf{0}_D, \boldsymbol{I}_D)$. Note that, since Σ_Θ is positive definite, P is orthogonal and for any $k \in \{1, \ldots, D\}$, $\lambda_{\Theta,k} > 0$.

By (59), $\ell^\Theta(\tilde{W}', \tilde{Z})$ is a linear combination of independent chi-square variables, each with 1 degree of freedom. Therefore, $\ell^\Theta(\tilde{W}', \tilde{Z})$ is distributed from a generalized chi-square distribution, and its CGF is given by,

$$\forall t \le \frac{1}{2}\min_{k\in\{1,\ldots,D\}}\lambda_{\Theta,k}, \quad K_{\ell^\Theta(\tilde{W}',\tilde{Z})}(t) = -t\sum_{k=1}^{D}\lambda_{\Theta,k} - \frac{1}{2}\sum_{k=1}^{D}\log(1 - 2\lambda_{\Theta,k}t) \tag{61}$$

$$= \frac{1}{2}\sum_{k=1}^{D}[-2\lambda_{\Theta,k}t - \log(1 - 2\lambda_{\Theta,k}t)]\,. \tag{62}$$

Since for any $s < 0$, $-s - \log(1-s) \leq s^2/2$, we deduce that

$$\forall t < 0, \quad K_{\ell^\Theta(\tilde{W}', \tilde{Z})}(t) \leq \frac{1}{2} \sum_{k=1}^{D} \frac{(2\lambda_{\Theta,k} t)^2}{2} = \|\lambda_\Theta\|^2 t^2 \,. \tag{63}$$

Since $\mathrm{rank}(\Theta\Theta^\top) = \mathrm{rank}(\Theta^\top\Theta)$ and $\Theta^\top\Theta = \boldsymbol{I}_d$, then $\mathrm{rank}(\Theta\Theta^\top) = d$. Moreover, $\Theta\Theta^\top$ and $\Theta^\top\Theta$ share the same non-zero eigenvalues. Therefore, $\Theta\Theta^\top$ has d eigenvalues equal to 1, and $(D-d)$ eigenvalues equal to 0, thus $\Theta^\top\Theta/n + \boldsymbol{I}_d$ has d eigenvalues equal to $1 + 1/n$ and and $(D-d)$ eigenvalues equal to 1, and

$$\|\lambda_\Theta\|^2 = d\left(1 + \frac{1}{n}\right)^2 + (D-d) \,. \tag{64}$$

By combining Theorem A.2 with (63) and (64), we obtain

$$\mathrm{gen}(\mu, \mathcal{A}^{(d)}) \leq \frac{2}{n} \sqrt{d\left(1 + \frac{1}{n}\right)^2 + (D-d)} \sum_{i=1}^{n} \mathbb{E}_{P_\Theta}\left[\sqrt{\mathsf{I}^\Theta(W'; Z_i)}\right] \tag{65}$$

Applying Jensen's inequality on (65) and the fact that $W' = \Theta^\top W$ with $W = \arg\min_{w \in \mathbb{R}^D} \widehat{\mathcal{R}}_n(w) = \bar{Z}$ finally yields,

$$\mathrm{gen}(\mu, \mathcal{A}^{(d)}) \leq \frac{2}{n} \sqrt{d\left(1 + \frac{1}{n}\right)^2 + (D-d)} \sum_{i=1}^{n} \sqrt{\mathsf{SI}_d^{(1)}(W; Z_i)} \,. \tag{66}$$

A.5. Detailed derivations for linear regression

Summary. Consider n i.i.d. samples $(x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^D$ and a response variable $y = (y_1, \ldots, y_n)$, $y_i \in \mathbb{R}$. The goal of $\mathcal{A}^{(d)}$ is $\min_{w \in \mathrm{W}_{\Theta,d}} \widehat{\mathcal{R}}_n(w) \triangleq (1/n)\|y - Xw\|^2$, where $X \in \mathbb{R}^{n \times D}$ is the design matrix. We show that if $n \geq D$, then $W' = (\Theta X^\top X \Theta^\top)^{-1} \Theta X^\top y$. Moreover, assume that X is deterministic and $y_i = x_i^\top W^\star + \varepsilon_i$ where $W^\star \in \mathbb{R}^D$ and $(\varepsilon_i)_{i=1}^n$ i.i.d. from $\mathcal{N}(0, \sigma^2)$. Then, by applying Theorem A.2, we bound $\mathrm{gen}(\mu, \mathcal{A}^{(d)})$ by a function of $\mathsf{I}(\phi(\Theta, X)W; y_i)$, where $\phi(\Theta, X) \triangleq (\Theta X^\top X \Theta^\top)^{-1} \Theta (X^\top X)$ and $W \triangleq \arg\min_{w \in \mathbb{R}^D} \widehat{\mathcal{R}}_n(w)$, 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 the rest of this subsection.

Problem statement. Consider n i.i.d. samples $(x_1, \ldots, x_n)$ and a response variable $y = (y_1, \ldots, y_n)$, where $x_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. Consder a learning algorithm $\mathcal{A}^{(d)}$ whose objective is $\arg\min_{w \in \mathrm{W}_{\Theta,d}} \widehat{\mathcal{R}}_n(w)$, with

$$\forall w \in \mathbb{R}^D, \quad \widehat{\mathcal{R}}_n(w) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top w)^2 = \frac{1}{n} \|y - Xw\|^2 \,. \tag{67}$$

where $X \in \mathbb{R}^{n \times D}$ is the design matrix. This objective is equivalent to finding $W' = \arg\min_{w' \in \mathbb{R}^d} \widehat{\mathcal{R}}_n^\Theta(w')$, where

$$\forall w' \in \mathbb{R}^d, \;\; \widehat{\mathcal{R}}_n^\Theta(w') = \frac{1}{n} \|y - X\Theta w'\|^2 \,. \tag{68}$$

We assume the problem is over-determined, i.e. $D \leq n$. Solving $\nabla_{w'} \widehat{\mathcal{R}}_n^\Theta(w') = 0$ yields

$$W' = (\Theta X^\top X \Theta^\top)^{-1} \Theta X^\top y \,. \tag{69}$$

On the other hand, we know the solution of $\arg\min_{w \in \mathbb{R}^D} \widehat{\mathcal{R}}_n(w)$ is the ordinary least squares (OLS) estimator, given by

$$W = (X^\top X)^{-1} X^\top y \,. \tag{70}$$

Hence, by (70) with (69), we deduce that

$$W' = (\Theta X^\top X \Theta^\top)^{-1} \Theta (X^\top X) W \tag{71}$$

Generalization error. In the remainder of this section, we assume that X is deterministic and there exists $W^\star \in \mathbb{R}^D$ such that $y_i = x_i^\top W^\star + \varepsilon_i$ where $(\varepsilon_i)_{i=1}^n$ are i.i.d. from $\mathcal{N}(0, \sigma^2)$. By using similar techniques as in Appendix A.4, one can show that $\mathrm{gen}(\mu, \mathcal{A}^{(d)}) = 2\sigma^2 d/n$.

Generalization error bound. Since $y_i \sim \mathcal{N}(x_i^\top W^\star, \sigma^2)$, and by (69),

$$x_i^\top \Theta^\top W' \sim \mathcal{N}(x_i^\top \Theta_X W^\star, \sigma^2 x_i^\top \Theta^\top [\Theta X^\top X \Theta^\top]^{-1} \Theta x_i) \tag{72}$$

where $\Theta_X = \Theta^\top (\Theta X^\top X \Theta^\top)^{-1} \Theta (X^\top X) \in \mathbb{R}^{D \times D}$. Therefore,

$$(\tilde{y}_i - x_i^\top \Theta^\top \tilde{W}') \sim \mathcal{N}(x_i^\top (\boldsymbol{I}_D - \Theta_X) W^\star, \sigma^2 (1 + x_i^\top \Theta^\top [\Theta X^\top X \Theta^\top]^{-1} \Theta x_i)) \,, \tag{73}$$

and

$$\ell^\Theta(\tilde{W}', \tilde{y}_i) \sim \sigma_i^2 \, \chi^2(1, \lambda_i) \,, \tag{74}$$

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

$$\forall t < \frac{1}{2\sigma_i^2}, \quad \mathbb{E}\big[e^{t\, \ell^\Theta(\tilde{W}', \tilde{y}_i)}\big] = \frac{e^{(\lambda_i \sigma_i^2 t)/(1 - 2\sigma_i^2 t)}}{\sqrt{1 - 2\sigma_i^2 t}} \tag{75}$$

and its expectation is $\mathbb{E}[\ell^\Theta(\tilde{W}', \tilde{y}_i)] = \sigma_i^2(1 + \lambda_i)$. Therefore, for $t < 1/(2\sigma_i^2)$ and $u_i = 2\sigma_i^2 t$,

$$K_{\ell^\Theta(\tilde{W}', \tilde{y}_i)}(t) = \frac{\lambda_i u_i}{2(1 - u_i)} - \frac{1}{2}\log(1 - u_i) - \frac{1}{2}(1 + \lambda_i) u_i \tag{76}$$

$$= \frac{1}{2}\{-\log(1 - u_i) - u_i\} + \frac{\lambda_i u_i^2}{2(1 - u_i)} \,. \tag{77}$$

Since $-\log(1 - x) - x \le x^2/2$ for $x < 0$, we deduce that for $t < 0$,

$$K_{\ell^\Theta(\tilde{W}', \tilde{y}_i)}(t) \le \frac{u_i^2}{4} + \frac{\lambda_i u_i^2}{2(1 - u_i)} \tag{78}$$

$$= \sigma_i^4 t^2 + \frac{2\lambda_i \sigma_i^4 t^2}{1 - 2\sigma_i^2 t} \,. \tag{79}$$

By applying Theorem A.2, we conclude that

$$\mathrm{gen}(\mu, \mathcal{A}^{(d)}) \le \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\Theta \left[\inf_{t > 0} \frac{\mathsf{I}(W'; y_i) + \sigma_i^4 t^2 \big(1 + 2\lambda_i (1 + 2\sigma_i^2 t)^{-1}\big)}{t} \right] . \tag{80}$$

By (71), W' is the projection of W along $\phi(\Theta, X) \triangleq (\Theta X^\top X \Theta^\top)^{-1} \Theta (X^\top X)$. The right-hand side term in (80) can thus be interpreted as a generalized SMI with a non-isotropic slicing distribution that depends on the fixed X.

As d converges to D, $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$ converges to $\mathbf{0}_n$. Indeed, consider the compact singular value decomposition (SVD) of $X\Theta^\top$, i.e. $X\Theta^\top = USV^\top$ where $S \in \mathbb{R}^{d \times d}$ is diagonal, $U \in \mathbb{R}^{n \times d}$, $V \in \mathbb{R}^{d \times m}$ s.t. $U^\top U = V^\top V = \boldsymbol{I}_d$. By using the pseudo-inverse expression of SVD,

$$X\Theta_X = X\Theta^\top (\Theta X^\top X \Theta^\top)^{-1} \Theta (X^\top X) \tag{81}$$

$$= USV^\top V S^{-1} U^\top X \tag{82}$$

$$= UU^\top X \tag{83}$$

Therefore, $\sqrt{\lambda} = X(\boldsymbol{I}_D - UU^\top) W^\star$. Since $U^\top U = \boldsymbol{I}_d$ with $U \in \mathbb{R}^{n \times d}$, then $\boldsymbol{I}_D - UU^\top$ has $(D - d)$ eigenvalues equal to 1 and d eigenvalues equal to 0. Hence, λ converges to $\mathbf{0}_n$ as $d \to D$.

B. Postponed Proofs for Section 5

B.1. Proof of Theorems 5.1 and 5.2

Proof of Theorem 5.1. By the triangle inequality, for any pair of models $(\mathcal{A}, \mathcal{A}')$,

$$|\mathrm{gen}(\mu, \mathcal{A})| \leq |\mathrm{gen}(\mu, \mathcal{A}) - \mathrm{gen}(\mu, \mathcal{A}')| + |\mathrm{gen}(\mu, \mathcal{A}')| \,. \tag{84}$$

Consider $\mathcal{A} : \mathrm{Z}^n \to \mathrm{W}$ and $\mathcal{A}' : \mathrm{Z}^n \to \mathrm{W}_{\Theta,d}$ such that $\mathcal{A}(S_n) = W$ may depend on $\Theta \sim P_\Theta$, and $\mathcal{A}'(S_n) = \Theta(\Theta^\top W)$. On the one hand, by applying Lemma A.1 with $X = \Theta^\top W$, $Y = Z_i$, $U = \Theta$ and $f^U(X, Y) = \ell^\Theta(\Theta^\top W, Z_i)$, we obtain

$$|\mathrm{gen}(\mu, \mathcal{A}')| \leq \frac{C}{n} \sum_{i=1}^{n} \mathbb{E}_{P_\Theta} \left[\sqrt{\frac{\mathsf{I}^\Theta(\Theta^\top W; Z_i)}{2}} \right] . \tag{85}$$

On the other hand, by the definition of the generalization error, one can show that

$$|\mathrm{gen}(\mu, \mathcal{A}) - \mathrm{gen}(\mu, \mathcal{A}')| = |\mathbb{E}[\mathcal{R}(W) - \widehat{\mathcal{R}}_n(W)] - \mathbb{E}[\mathcal{R}^\Theta(\Theta^\top W) - \widehat{\mathcal{R}}_n^\Theta(\Theta^\top W)]| \tag{86}$$

$$\leq |\mathbb{E}[\mathcal{R}(W) - \mathcal{R}^\Theta(\Theta^\top W)]| + |\mathbb{E}[\widehat{\mathcal{R}}_n(W) - \widehat{\mathcal{R}}_n^\Theta(\Theta^\top W)]| \tag{87}$$

where the expectations are computed over $P_{W|\Theta,S_n} \otimes P_\Theta \otimes \mu^{\otimes n}$. Additionally,

$$|\mathbb{E}[\mathcal{R}(W) - \mathcal{R}^\Theta(\Theta^\top W)]| = |\mathbb{E}_{P_{W|\Theta} \otimes P_\Theta \otimes \mu}[\ell(W, Z) - \ell(\Theta\Theta^\top W, Z)]| \tag{88}$$

$$\leq \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta \otimes \mu} |\ell(W, Z) - \ell(\Theta\Theta^\top W, Z)| \tag{89}$$

$$\leq L \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\Theta^\top W\| \,, \tag{90}$$

where (88) follows from the definition of the population risks $\mathcal{R}(w)$ and $\mathcal{R}^\Theta(\Theta^\top w)$, and (90) results from the assumption that $\ell(\cdot, z) : \mathrm{W} \to \mathbb{R}_+$ is L-Lipschitz for all $z \in \mathrm{Z}$.

Using similar arguments, one can show that

$$|\mathbb{E}[\widehat{\mathcal{R}}_n(W) - \widehat{\mathcal{R}}_n^\Theta(\Theta^\top W)]| \leq L \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\Theta^\top W\| \,, \tag{91}$$

and we conclude that

$$|\mathrm{gen}(\mu, \mathcal{A}) - \mathrm{gen}(\mu, \mathcal{A}')| \leq 2L \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\Theta^\top W\| \,. \tag{92}$$

The final result follows from bounding (84) using (85) and (92).

□

Proof of Theorem 5.2. Consider $\mathcal{A} : \mathrm{Z}^n \to \mathrm{W}$ and $\mathcal{A}' : \mathrm{Z}^n \to \mathrm{W}_{\Theta,d}$ such that $\mathcal{A}(S_n) = W$ may depend on $\Theta \sim P_\Theta$, and $\mathcal{A}'(S_n) = \Theta\mathcal{Q}(\Theta^\top W)$. Using the same techniques as in the proof of Theorem 5.1, we obtain

$$|\mathrm{gen}(\mu, \mathcal{A})| \leq 2L \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\mathcal{Q}(\Theta^\top W)\| + |\mathrm{gen}(\mu, \mathcal{A}')| \tag{93}$$

$$\leq 2L \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\mathcal{Q}(\Theta^\top W)\| + C \, \mathbb{E}_{P_\Theta} \left[\sqrt{\frac{\mathsf{I}^\Theta(\mathcal{Q}(\Theta^\top W); S_n)}{2n}} \right] \tag{94}$$

where eq. (94) follows from applying Theorem 4.1.

Then, by using the triangle inequality, the fact that $\|\Theta\| = \|\Theta^\top \Theta\| = 1$, and the properties of $\mathcal{Q}$,

$$\mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\mathcal{Q}(\Theta^\top W)\| \tag{95}$$

$$\leq \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\Theta^\top W\| + \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|\Theta\Theta^\top W - \Theta\mathcal{Q}(\Theta^\top W)\| \tag{96}$$

$$\leq \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\Theta^\top W\| + \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \left[\|\Theta\| \|\Theta^\top W - \mathcal{Q}(\Theta^\top W)\| \right] \tag{97}$$

$$\leq \mathbb{E}_{P_{W|\Theta} \otimes P_\Theta} \|W - \Theta\Theta^\top W\| + \delta \,. \tag{98}$$

Finally, since $\mathcal{Q}(\Theta^\top W)$ is a discrete random variable and $\|\Theta^\top W\| \leq M$, we use the same arguments as in Section 4.1 to bound $\mathsf{I}^\Theta(\mathcal{Q}(\Theta^\top W); S_n)$ by $d \log(2M\sqrt{d}/\delta)$.

□

B.2. Rate-distortion bounds applied to feedforward neural networks

In the following, we determine the conditions under which feedforward networks meet the assumptions outlined in Theorems 5.1 and 5.2. Suppose $\mathrm{Z} = \{(x, y) \in \mathrm{X} \times \{1, \dots, K\}\}$ is the set of feature-label pairs.

For $q \in \mathbb{N}^*$, a q-layer feedforward network is characterized by a mapping $f : \mathrm{W} \times \mathrm{X} \to \mathbb{R}^K$ such that the output of its i-th layer, $X^{(i)}$, satisfies $X^{(1)} \triangleq W^{(1)}X$ and for $i \in \{2, \dots, q\}$, $X^{(i)} \triangleq W^{(i)}(\psi_i(X^{(i-1)}))$, where ψ_i is an activation function applied element-wise, $W^{(i)} \in \mathbb{R}^{d_{out}^{(i)} \times d_{in}^{(i)}}$, and X is the input feature vector.

For any such feedforward network f, we denote by $\bar{f} : \mathrm{W}_{\Theta,d} \times \mathrm{X} \to \mathbb{R}^K$ a feedforward network with the same architecture as f (i.e., same number of layers and neurons, and same type of activation functions), but with its parameter space restricted to $\mathrm{W}_{\Theta,d}$. Denote by $\bar{X}^{(i)}$ and $\bar{W}^{(i)}$ the output and weight matrix of the i-th layer of $\bar{f}$.

Theorem B.1. *Let f be a q-layer feedforward neural network. Assume that for $i \in \{2, \dots, q-1\}$, ψ_i is α_i-Lipschitz continuous and $\psi_i(\mathbf{0}) = \mathbf{0}$. Assume for $i \in \{1, \dots, q\}$, $\|W^{(i)}\|_2 \leq M$ and $\|\bar{W}^{(i)}\|_2 \leq M$. Then, for $i \in \{1, \dots, q\}$,*

$$\|X^{(i)} - \bar{X}^{(i)}\|_2 \leq M^{i-1}\|X\|_2 \left(\prod_{j=1}^{i} \alpha_j\right) \sum_{j=1}^{i} \|W^{(j)} - \bar{W}^{(j)}\|_2 \,, \tag{99}$$

where $\alpha_1 \triangleq 1$.

Proof. We prove this result by induction. By definition, $\|X^{(1)} - \bar{X}^{(1)}\|_2 = \|(W^{(1)} - \bar{W}^{(1)})X\|_2$. Since the spectral norm is consistent with the Euclidean norm, $\|X^{(1)} - \bar{X}^{(1)}\|_2 \leq \|W^{(1)} - \bar{W}^{(1)}\|_2\|X\|_2$, so (99) is true for $i = 1$.

Now, let $i > 1$ and assume that (99) holds for $j \in \{1, \dots, i-1\}$. Then,

$$\begin{aligned}
\|X^{(i)} - \bar{X}^{(i)}\|_2 &= \|W^{(i)}\psi_i(X^{(i-1)}) - \bar{W}^{(i)}\psi_i(\bar{X}^{(i-1)})\|_2 && (100)\\
&= \|(W^{(i)} - \bar{W}^{(i)})\psi_i(X^{(i-1)}) + \bar{W}^{(i)}(\psi_i(X^{(i-1)}) - \psi_i(\bar{X}^{(i-1)}))\|_2 && (101)\\
&\leq \|W^{(i)} - \bar{W}^{(i)}\|_2\|\psi_i(X^{(i-1)})\|_2 + \|\bar{W}^{(i)}\|_2\|\psi_i(X^{(i-1)}) - \psi_i(\bar{X}^{(i-1)})\|_2 \,, && (102)
\end{aligned}$$

where (102) results from applying the triangle inequality and $\|Mx\|_2 \leq \|M\|_2\|x\|_2$. Since ψ_i is α_i-Lipschitz continuous and $\psi_i(\mathbf{0}) = \mathbf{0}$, we obtain

$$\|X^{(i)} - \bar{X}^{(i)}\|_2 \leq \alpha_i \left(\|W^{(i)} - \bar{W}^{(i)}\|_2\|X^{(i-1)}\|_2 + \|\bar{W}^{(i)}\|_2\|X^{(i-1)} - \bar{X}^{(i-1)}\|_2\right) . \tag{103}$$

By recursively using the definition of $X^{(i)}$ and $\|W^{(i)}\psi_i(X^{(i-1)})\|_2 \leq \alpha_i\|W^{(i)}\|_2\|X^{(i-1)}\|_2$, one can show that

$$\|X^{(i-1)}\|_2 \leq \|X\|_2 \prod_{j=1}^{i-1} \alpha_j\|W^{(j)}\|_2 \leq M^{i-1}\|X\|_2 \prod_{j=1}^{i-1} \alpha_j \,. \tag{104}$$

Additionally, since we assume (99) holds for $j \in \{1, \dots, i-1\}$,

$$\|X^{(i-1)} - \bar{X}^{(i-1)}\|_2 \leq M^{i-2}\|X\|_2 \left(\prod_{j=1}^{i-1} \alpha_j\right) \sum_{j=1}^{i-1} \|W^{(j)} - \bar{W}^{(j)}\|_2 \,. \tag{105}$$

By plugging (104) and (105) in (103), we obtain

$$\begin{aligned}
\|X^{(i)} - \bar{X}^{(i)}\|_2 &\leq \|X\|_2 \left(\prod_{j=1}^{i} \alpha_j\right) \left(M^{i-1}\|W^{(i)} - \bar{W}^{(i)}\|_2 + \|\bar{W}^{(i)}\|_2 M^{i-2} \sum_{j=1}^{i-1} \|W^{(j)} - \bar{W}^{(j)}\|_2\right) && (106)\\
&\leq M^{i-1}\|X\|_2 \left(\prod_{j=1}^{i} \alpha_j\right) \sum_{j=1}^{i} \|W^{(j)} - \bar{W}^{(j)}\|_2 \,, && (107)
\end{aligned}$$

which concludes the proof.

□

Theorem B.2. *Let f and $\bar{f}$ be two q-layer feedforward neural networks satisfying the assumptions in Theorem B.1. Denote by $\mathcal{A}$ (respectively, $\bar{\mathcal{A}}$) the learning algorithm consisting in training f (resp., $\bar{f}$) using the loss function $\ell : \mathrm{W} \times \mathrm{Z} \to \mathbb{R}_+$, where $\mathrm{Z} = \mathrm{X} \times \{1, \dots, K\}$. Let $\tilde{\ell} : \mathbb{R}^K \times \{1, \dots, K\} \to \mathbb{R}_+$ be the mapping such that for any $w \in \mathrm{W}$ and $z = (x, y) \in \mathrm{Z}$, $\ell(w, z) = \tilde{\ell}(f(w, x), y)$ (resp., $\ell(w, z) = \tilde{\ell}(\bar{f}(w, x), y)$). Assume that $\tilde{\ell}$ is β-Lipschitz w.r.t. the first variable. Suppose additionally that $\forall X \in \mathrm{X}$, $\|X\|_2 \leq R$. Then,*

$$|\mathrm{gen}(\mu, \mathcal{A}) - \mathrm{gen}(\mu, \bar{\mathcal{A}})| \leq 2\beta M^{q-1} R \left(\prod_{i=1}^{q} \alpha_i \right) \mathbb{E}\Big[\sum_{i=1}^{q} \|W^{(i)} - \bar{W}^{(i)}\|_2 \Big] . \tag{108}$$

Proof. By definition of the generalization error,

$$\begin{aligned}
|\mathrm{gen}(\mu, \mathcal{A}) - \mathrm{gen}(\mu, \bar{\mathcal{A}})| &= |\mathbb{E}[\mathcal{R}(W) - \widehat{\mathcal{R}}_n(W)] - \mathbb{E}[\mathcal{R}(\bar{W}) - \widehat{\mathcal{R}}_n(\bar{W})]| && (109)\\
&= |\mathbb{E}[\mathcal{R}(W) - \mathcal{R}(\bar{W}) - (\widehat{\mathcal{R}}_n(W) - \widehat{\mathcal{R}}_n(\bar{W}))]| && (110)\\
&\leq \mathbb{E}[|\mathcal{R}(W) - \mathcal{R}(\bar{W})| + |\widehat{\mathcal{R}}_n(W) - \widehat{\mathcal{R}}_n(\bar{W})|] && (111)
\end{aligned}$$

For any $(W, \bar{W}) \sim P_{W|S_n} \otimes P_{\bar{W}|S_n}$,

$$\begin{aligned}
|\mathcal{R}(W) - \mathcal{R}(\bar{W})| &= |\mathbb{E}_{Z\sim\mu}[\ell(W, Z)] - \mathbb{E}_{Z\sim\mu}[\ell(\bar{W}, Z)]| && (112)\\
&\leq |\mathbb{E}_{(X,Y)\sim\mu}[\tilde{\ell}(f(W, X), Y) - \tilde{\ell}(\bar{f}(\bar{W}, X), Y)]| && (113)\\
&\leq \mathbb{E}_{(X,Y)\sim\mu}\big[|\tilde{\ell}(f(W, X), Y) - \tilde{\ell}(\bar{f}(\bar{W}, X), Y)|\big] && (114)\\
&\leq \beta\, \mathbb{E}\big[\|f(W, X) - \bar{f}(\bar{W}, X)\|_2\big] . && (115)
\end{aligned}$$

We bound (115) using Theorem B.1 and we obtain,

$$\begin{aligned}
|\mathcal{R}(W) - \mathcal{R}(\bar{W})| &\leq \beta M^{q-1} \mathbb{E}\big[\|X\|_2\big] \left(\prod_{i=1}^{q} \alpha_i \right) \sum_{i=1}^{q} \|W^{(i)} - \bar{W}^{(i)}\|_2 && (116)\\
&\leq \beta M^{q-1} R \left(\prod_{i=1}^{q} \alpha_i \right) \sum_{i=1}^{q} \|W^{(i)} - \bar{W}^{(i)}\|_2 . && (117)
\end{aligned}$$

Similarly,

$$\begin{aligned}
|\widehat{\mathcal{R}}_n(W) - \widehat{\mathcal{R}}_n(\bar{W})| &= \left| \frac{1}{n} \sum_{i=1}^{n} \ell(W, Z_i) - \frac{1}{n} \sum_{i=1}^{n} \ell(\bar{W}, Z_i) \right| && (118)\\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(f(W, X_i), Y_i) - \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(\bar{f}(\bar{W}, X_i), Y_i) \right| && (119)\\
&\leq \frac{1}{n} \sum_{i=1}^{n} |\tilde{\ell}(f(W, X_i), Y_i) - \tilde{\ell}(\bar{f}(\bar{W}, X_i), Y_i)| && (120)\\
&\leq \frac{\beta}{n} \sum_{i=1}^{n} \|f(W, X_i) - \bar{f}(\bar{W}, X_i)\|_2 && (121)\\
&\leq \frac{\beta}{n} M^{q-1} \left(\sum_{i=1}^{n} \|X_i\|_2 \right) \left(\prod_{i=1}^{q} \alpha_i \right) \sum_{i=1}^{q} \|W^{(i)} - \bar{W}^{(i)}\|_2 && (122)\\
&\leq \beta M^{q-1} R \left(\prod_{i=1}^{q} \alpha_i \right) \sum_{i=1}^{q} \|W^{(i)} - \bar{W}^{(i)}\|_2 . && (123)
\end{aligned}$$

We obtain the final result by plugging (117) and (123) in (111).

□

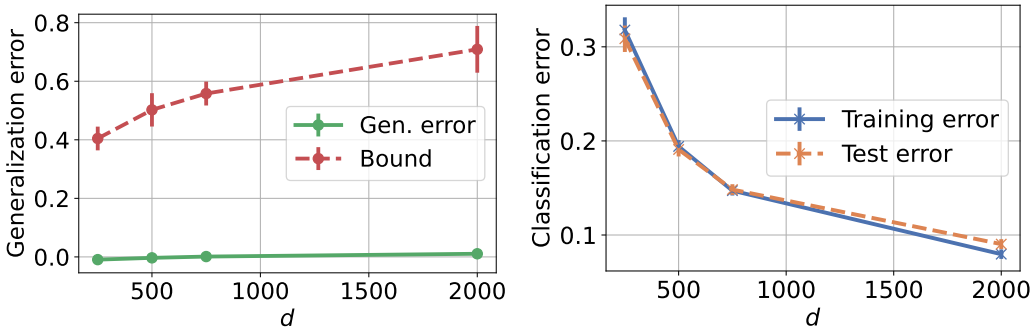


Figure 5: Generalization bounds on MNIST classification with neural networks trained on $\mathrm{W}_{\Theta,d}$

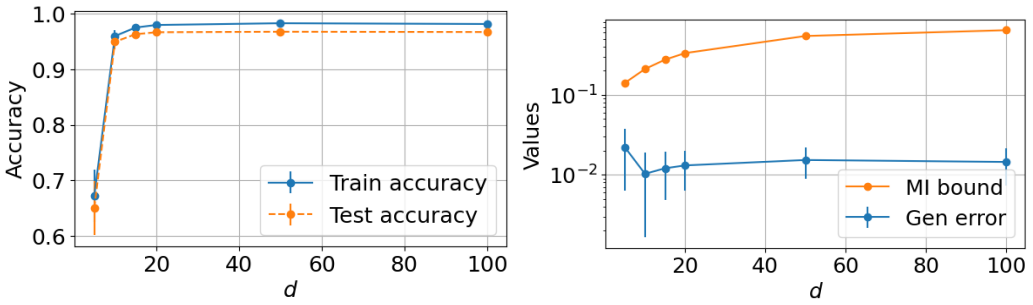


Figure 6: Generalization bounds on Iris dataset classification with neural networks trained on $\mathrm{W}_{\Theta,d}$

C. Additional Experimental Details for Section 6

C.1. Methodological details

Architecture for MINE. In all our experiments, the MI terms are estimated with MINE (Belghazi et al., 2018) based on a fully-connected neural network with one single hidden layer of dimension 100. The network is trained for 200 epochs and a batch size of 64, using the Adam optimizer (Kingma & Ba, 2017) with default parameters (on PyTorch).

Quantization method. We use the quantization scheme of Lotfi et al. (2022) with minor modifications. We learn $c = [c_1, ..., c_L] \in \mathbb{R}^L$ quantization levels in 16-precision during training using the straight through estimator, and quantize the weights $W' = [W_1, \cdots, W_d] \in \mathbb{R}^d$ into $\widehat{W}_i = c_{q(i)}$, where $q(i) = \arg\min_{k\in\{1,...,L\}} |W_i - c_k|$. Post quantization, arithmetic coding is employed for further compression, to take into account the fact that quantization levels are not uniformly distributed in the quantized weights. Denote by p_k the empirical probability of c_k. Arithmetic coding uses at most $\lceil d \times H(p) \rceil + 2$ bits, where $H(p) = -\sum_{k=1}^{L} p_k \log_2 p_k$. The total bit requirement for the quantized weights, the codebook c, and the probabilities $(p_1, \ldots, p_L)$ is bounded by $\lceil d \times H(p) \rceil + L \times (16 + \lceil \log_2 d \rceil) + 2$.

C.2. Additional details and empirical results

Binary classification with logistic regression (Section 6.1). We consider the binary classification problem solved with logistic regression as described in (Bu et al., 2019, §VI), with features dimension $s = 20$, hence $D = s + 1$ (weights and intercept). We train our model on $\mathrm{W}_{\Theta,d}$ for different values of $d < D$, using n training samples. We compute the test error on $\lfloor 20n/80 \rfloor$ observations. For each value of n and d, we approximate the generalization error for 30 samples of Θ independently drawn from the SVD-based projector (see Section 6). We estimate the MI term in the bounds via MINE (with the aforementioned architecture) using 30 samples of $(W', Z_i) \sim P_{W'|S_n,\Theta} \otimes \mu$ for each Θ.

Classification with NNs (Section 6.1). We consider a feedforward neural network with 2 fully-connected layers of width 200 to classify MNIST (LeCun & Cortes, 2010) and CIFAR-10 (Krizhevsky et al., 2009). The random projections are

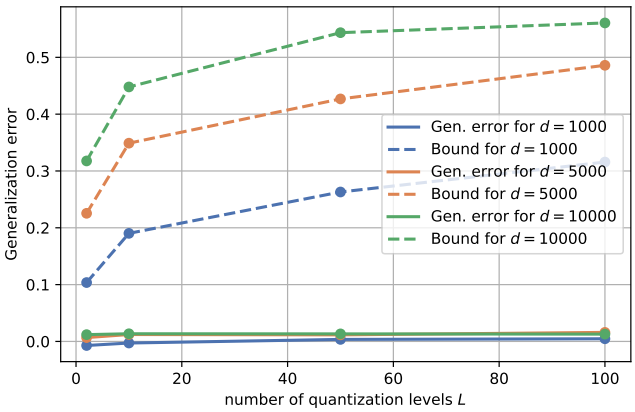


Figure 7: Influence of the number of quantization levels L on the generalization error and our bounds, for MNIST classification with NNs.

sampled using the Kronecker product projector, in order to scale better with the high-dimensionality of our models (see Appendix C.2). We train our NNs on $\mathrm{W}_{\Theta,d}$ for different values of d, including the intrinsic dimensions reported in (Li et al., 2018). We approximate the generalization error for 30 samples of Θ and estimate our MI-based bounds given by Theorem A.2. The MI terms are estimated using MINE over 100 samples of $(W', Z_i) \sim P_{W'|S_n,\Theta} \otimes \mu$ for each Θ. As MINE requires multiple runs, which can be very expensive, we only estimate MI for datasets and models of reasonable sizes: see Figure 5 for results on MNIST. For MNIST and CIFAR-10, we quantize W' and evaluate our quantization-based generalization bounds. To train our NNs, we run Adam (Kingma & Ba, 2017) with default parameters for 30 epochs and batch size of 64 or 128.

We also classify the Iris dataset (Fisher, 1936). We train a two-hidden-layer NN with width 100 (resulting in $D = 10\,903$ parameters) on $\mathrm{W}_{\Theta,d}$. We use Adam with a learning rate of 0.1 as optimizer, for 200 epochs and batch size of 64. We approximate the generalization error for 20 samples of Θ independently drawn from the SVD-based projector. We evaluate our generalization bounds (Theorem 4.2) using MINE over 500 samples of $(W', Z_i) \sim P_{W'|\Theta,S_n} \otimes \mu$ for each Θ. We report results for $d \in \{5, 10, 15, 20, 50, 100\}$ in Figure 6. 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 of our bounds are non-vacuous.

Influence of the number of quantization levels. We analyze the influence of the quantization levels L on the generalization error and our bounds in practice. We consider the MNIST classification task with NNs in Section 6.1 and train for different values of L. We report the results in Figure 7 for several values of d. We observe that for all tested dimensions, the generalization error increases with increasing L. Our bound exhibits the same behavior, as anticipated given the dependence on L (see paragraph "Quantization" in Section 6). This experiment illustrates that *(i)* the more aggressive the compression, the better the generalization, *(ii)* our bounds accurately reflect the behavior of the generalization error, and is tighter for lower values of d and L.

Classification with NNs (Section 6.2). We consider a feedforward neural network f with 3 fully-connected layers and ReLU activations, as formally described in Appendix B.2. We parameterize f with $w = \Theta(\Theta^\top w_1) + \bar{\Theta}(\bar{\Theta}^\top w_2) \in \mathbb{R}^D$, where $\Theta \in \mathrm{St}(d, D)$ is randomly generated at initialization and $\bar{\Theta} \in \mathbb{R}^{D\times(D-d)}$ is such that $[\Theta, \bar{\Theta}] \in \mathbb{R}^{D\times D}$ forms an orthogonal basis of $\mathbb{R}^D$. The projection matrix Θ is generated with the sparse projector by Li et al. (2018). Each run consists in randomly selecting a subset of MNIST of $n = 1000$ samples and training f on that dataset for 5 different samples of Θ. For each Θ, we train for 20 epochs using the Adam optimizer with a batch size of 256, learning rate $\eta = 0.01$ for w_1 and $\eta/10$ for w_2, and other parameters set to their default values (Kingma & Ba, 2017). During training, we clamp the norm of each layer's weight matrix at the end of each iteration to satisfy the condition in Theorem B.2. All hyperparameters, including C and M, were chosen so that the neural network trained on MNIST with $d = D$ achieves a training accuracy of at least 99% in almost all runs.