Data-Dependent PAC-Bayesian Bounds in the Random-Subset Setting with Applications to Neural Networks

Fredrik Hellström [1] **Giuseppe Durisi** [1]

Abstract

The PAC-Bayesian framework has proven to be a useful tool to obtain nonvacuous generalization bounds for modern learning algorithms, such as overparameterized neural networks. A known heuristic to tighten such bounds is to use data-dependent priors. In this paper, we show how the information-theoretically motivated random-subset setting introduced by Steinke & Zakynthinou (2020) enables the derivation of PAC-Bayesian bounds that naturally involve a data-dependent prior. We evaluate these bounds for neural networks trained on MNIST and Fashion-MNIST, and study their dependence on the training set size, the achieved training accuracy, and the effect of randomized labels.

1. Introduction

Recently, interest in the use of information-theoretic techniques for bounding the loss of learning algorithms has surged. While the first results of this flavor can be traced to the probably approximately correct (PAC)-Bayesian approach (McAllester, 1998; Catoni, 2007) (see also (Guedj, 2019) for a recent review), the connection between loss bounds and classical information-theoretic measures was made explicit in the works of Russo & Zou (2016) and Xu & Raginsky (2017), where bounds on the average population loss were derived in terms of the mutual information between the training data and the output hypothesis. Since then, these average-loss bounds have been tightened (Bu et al., 2020; Negrea et al., 2019). Furthermore, the information-theoretic framework has also been successfully applied to derive tail-probability bounds on the population loss (Bassily et al., 2018; Esposito et al., 2019).

In the PAC-Bayesian framework, one considers stochastic classifiers where, for each prediction, a hypothesis W is drawn from a so-called *posterior* distribution $P_{W|\boldsymbol{Z}}$ given the training data $\boldsymbol{Z}$. The population loss of the stochastic classifier characterized by the posterior is then bounded by a function of the Kullback-Leibler (KL) divergence between the posterior and a reference measure, usually called a *prior*. To obtain tight bounds, selecting a good prior is key. While the prior is traditionally assumed to be data-independent, this is not strictly necessary for PAC-Bayes bounds to hold (Ambroladze et al., 2006; Rivasplata et al., 2020; Dziugaite et al., 2020). A known heuristic for improving the quality of the bound is to use all of $\boldsymbol{Z}$ to select $P_{W|\boldsymbol{Z}}$, but only a part of it, $\boldsymbol{Z}_B$, to evaluate the bound on the population loss. Then, the remaining part, $\boldsymbol{Z}_P$, can be used freely to select a data-dependent prior $Q_{W|\boldsymbol{Z}_P}$ (Ambroladze et al., 2006). Recently, Dziugaite et al. (2020) showed that for some learning settings, such a data-dependent prior may be necessary to enable nonvacuous bounds, and empirically demonstrated its usefulness for neural networks (NN).

The purpose of this paper is to derive and evaluate PAC-Bayesian bounds on the test loss for the *random-subset setting* introduced by Steinke & Zakynthinou (2020), and show how this naturally leads to a procedure for selecting data-dependent priors. In the random-subset setting, $2n$ training samples $\widetilde{\boldsymbol{Z}} = (\tilde{Z}_1, \dots, \tilde{Z}_{2n})$ are available, with all entries of $\widetilde{\boldsymbol{Z}}$ being drawn independently from some distribution P_Z on an instance space $\mathcal{Z}$. However, only a randomly selected subset of cardinality n is actually used for training. It is selected as follows: let $\boldsymbol{S} = (S_1, \dots, S_n)$ be an n-dimensional random vector, the elements of which are drawn independently from a $\mathrm{Bern}(1/2)$ distribution and are independent of $\widetilde{\boldsymbol{Z}}$. Then, for $i = 1, \dots, n$, the ith training sample in $\boldsymbol{Z}(\boldsymbol{S})$ is $Z_i(S_i) = \tilde{Z}_{i+S_i n}$. Based on this training set, a hypothesis $W \in \mathcal{W}$ is chosen through a randomized learning algorithm $P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}} = P_{W|\boldsymbol{Z}(\boldsymbol{S})}$, which is a conditional distribution on $\mathcal{W}$ given $(\widetilde{\boldsymbol{Z}}, \boldsymbol{S})$ that gives rise to the Markov property $(\widetilde{\boldsymbol{Z}}, \boldsymbol{S}) - \boldsymbol{Z}(\boldsymbol{S}) - W$. Let $L_{\boldsymbol{Z}(\boldsymbol{S})}(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i(S_i))$ denote the training loss, where $\ell(\cdot,\cdot)$ is a loss function, which throughout this paper is assumed to have range $[0, 1]$. Furthermore, let $\bar{\boldsymbol{S}}$ denote the modulo-

[1]Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden. Correspondence to: Fredrik Hellström <frehells@chalmers.se>. Part of this material will be presented at the International Symposium on Information Theory (ISIT) 2021 (Hellström & Durisi, 2021).

ICML 2021 Workshop on Information-Theoretic Methods for Rigorous, Responsible, and Reliable Machine Learning. Copyright 2021 by the authors.

2 complement of $\boldsymbol{S}$. Then $L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)$ can be interpreted as a test loss, since W is independent of $\boldsymbol{Z}(\bar{\boldsymbol{S}})$. Note that the average over $(\widetilde{\boldsymbol{Z}}, \boldsymbol{S})$ of the test loss is the population loss $L_{P_Z}(W) = \mathbb{E}_{P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)] = \mathbb{E}_{P_Z}[\ell(W, Z)]$. For the random-subset setting, bounds on the average population loss are derived in (Steinke & Zakynthinou, 2020) in terms of the conditional mutual information (CMI) $I(W; \boldsymbol{S}|\widetilde{\boldsymbol{Z}})$ between the chosen hypothesis W and the random vector $\boldsymbol{S}$ given $\widetilde{\boldsymbol{Z}}$. These bounds are always finite, since $I(W; \boldsymbol{S}|\widetilde{\boldsymbol{Z}})$ is never larger than n bits. In contrast, the bounds obtained in (Xu & Raginsky, 2017) depend on the mutual information $I(W; \boldsymbol{Z})$, a quantity that can be unbounded if W reveals too much about the training set $\boldsymbol{Z}$. The difference between bounds depending on $I(W; \boldsymbol{Z})$ and those given in terms of $I(W; \boldsymbol{S}|\widetilde{\boldsymbol{Z}})$ can also be explained from a PAC-Bayesian perspective in terms of the prior distribution. While $I(W; \boldsymbol{Z})$ compares the posterior $P_{W|\boldsymbol{Z}}$ to the oracle prior P_W, which is the marginalization of $P_{W|\boldsymbol{Z}}P_{\boldsymbol{Z}}$ over $P_{\boldsymbol{Z}}$, the CMI $I(W; \boldsymbol{S}|\widetilde{\boldsymbol{Z}})$ compares $P_{W|\boldsymbol{Z}(\boldsymbol{S})}$ to $P_{W|\widetilde{\boldsymbol{Z}}}$, which is the marginalization of $P_{W|\boldsymbol{Z}(\boldsymbol{S})}P_{\boldsymbol{S}} = P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}}$ over $P_{\boldsymbol{S}}$. Thus, the distribution that plays the role of the prior in the CMI bound is adapted to the data by default.

The following bounds on the average population loss are derived in (Steinke & Zakynthinou, 2020, Thm. 2):

$$\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \sqrt{\frac{2I(W; \boldsymbol{S}|\widetilde{\boldsymbol{Z}})}{n}} \tag{1}$$

$$\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \leq 2\,\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \frac{3I(W; \boldsymbol{S}|\widetilde{\boldsymbol{Z}})}{n}. \tag{2}$$

We will refer to bounds like (1), where the n-dependence is of the form $\sqrt{\mathrm{IM}(n)/n}$ for some information measure $\mathrm{IM}(n)$, as *slow-rate* bounds. We refer to bounds like (2), which have an $\mathrm{IM}(n)/n$-dependence, as *fast-rate* bounds. We note that the results in (Steinke & Zakynthinou, 2020, Thm. 2) pertain to the average population loss, and no PAC-Bayesian bounds are provided.

Contributions In this paper, we derive PAC-Bayesian versions of (1) and (2). We then use these bounds to characterize the test loss of NNs used to classify images from the MNIST and Fashion-MNIST data sets. To obtain nonvacuous bounds for NNs, it is crucial to choose a data-dependent prior (Dziugaite et al., 2020). The random-subset setting provides a natural way to do this by choosing the prior as an approximation of $P_{W|\widetilde{\boldsymbol{Z}}}$. Specifically, we set the posterior to be an isotropic Gaussian distribution centered on the output the output of stochastic gradient descent (SGD) on the training set $\boldsymbol{Z}(\boldsymbol{S})$. For the prior, we form a number of subsets of $\widetilde{\boldsymbol{Z}}$, and set the prior as an isotropic Gaussian centered around the average of the output of SGD on the different subsets. Our numerical results reveal that the bounds are nonvacuous, and in line with previously reported results for similar setups (Dziugaite et al., 2020). Furthermore, we study the impact of the training set size and the target training loss on our bounds and the KL divergence they contain. We find that, for higher training losses, the KL divergence decreases with n, while it increases for lower losses. For fixed n, the bounds initially improve as the training loss decreases, but then grow rapidly once the training loss reaches a certain point. We note that these observations hold for our specific choice of prior, posterior, and training algorithm, and may not be true in general.

2. PAC-Bayesian Random-Subset Bounds

We now present the PAC-Bayesian versions of the average bounds in (1) and (2). Their derivations, which are detailed in Appendix A, are based on the use of exponential inequalities—a framework through which the average bounds (1) and (2) can readily be recovered. Furthermore, single-draw bounds on the test loss, also known as *pointwise* or *de-randomized* PAC-Bayesian bounds (Viallard et al., 2021), can also be derived. These results are deferred to Appendix A.

Theorem 1 *Consider the random-subset setting introduced in Section 1. Let $W \in \mathcal{W}$ be distributed according to $P_{W|\boldsymbol{Z}(\boldsymbol{S})}$. Let $\lambda, \gamma > 0$ be constants such that $\lambda(1-\gamma) + (e^{\lambda} - 1 - \lambda)(1+\gamma^2) \leq 0$. Furthermore, let $Q_{W|\widetilde{\boldsymbol{Z}}}$ be a data-dependent conditional prior such that the distributions $Q_{W|\widetilde{\boldsymbol{Z}}}P_{\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$ and $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}} = P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\widetilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$ are absolutely continuous with respect to each other. Then, with probability at least $1 - 2\delta$ over $P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$, the PAC-Bayesian test loss is bounded as*

$$\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)\right] \leq \min\{B_{\text{slow}}, B_{\text{fast}}\} \tag{3}$$

where B_{slow} and B_{fast} are given by

$$B_{\text{slow}} = \mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \sqrt{\frac{2}{n-1}\left(D(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, Q_{W|\widetilde{\boldsymbol{Z}}}) + \log\frac{\sqrt{n}}{\delta}\right)} \tag{4}$$

$$B_{\text{fast}} = \gamma\,\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \frac{\left(D(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, Q_{W|\widetilde{\boldsymbol{Z}}}) + \log\frac{1}{\delta}\right)}{\lambda n}. \tag{5}$$

Note that the bound in (3) is on the test loss instead of the population loss. One can obtain population-loss bounds by adding a penalty term to (3), as shown in (Hellström &

Table 1. The estimated test loss as well as the slow-rate (4) and fast-rate (5) bounds on the test loss, obtained after training with SGD for 100 epochs, for the different architectures and data sets . The confidence intervals correspond to two standard deviations.

	FCNN, MNIST	FCNN, Fashion-MNIST	LeNet-5, MNIST	LeNet-5, Fashion-MNIST
Test loss	0.098 ± 0.006	0.265 ± 0.057	0.045 ± 0.006	0.301 ± 0.051
Slow-rate bound	0.213 ± 0.010	0.356 ± 0.041	0.291 ± 0.018	0.437 ± 0.030
Fast-rate bound	0.198 ± 0.012	0.471 ± 0.098	0.171 ± 0.016	0.557 ± 0.076

Durisi, 2020, Thm. 2). However, when comparing bounds to the empirical performance of an algorithm, the population loss is unknown. Thus, in practice, one has to resort to evaluating a test loss.

The bounds in (4) and (5) are data-dependent, i.e., they depend on the specific instances of $\widetilde{\boldsymbol{Z}}$ and $\boldsymbol{S}$. This makes them computable for a given data set. They can be turned into data-independent bounds that are functions of the average of the KL divergences appearing in (4) and (5), at the cost of a worse dependence on the confidence parameter δ. Alternatively, one can obtain bounds that have a more benign dependence on δ if one allows the bounds to depend on sufficiently high moments of the KL divergences, or if one replaces these measures by quantities such as conditional maximal leakage or conditional α-divergence. See (Hellström & Durisi, 2020) for further discussion.

Under the assumption that the algorithm always achieves zero training loss, the constants in Theorem 1 can be sharpened. Furthermore, the average bound in (2) can be tightened by considering a sample-wise decomposition of the error. We present these extensions in Appendix A.

3. Experiments

We numerically evaluate the bounds in (4) and (5) for some NNs. Specifically, we consider the convolutional network LeNet-5 and a fully connected NN (FCNN) with two hidden layers of width 600, trained on either MNIST or Fashion-MNIST using SGD. We set the loss function to be the classification error. To evaluate the bounds, we set $\lambda = 1/2.98$ and $\gamma = 1.795$. We select the posterior $P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ to be $\mathcal{N}(W \mid \boldsymbol{\mu}_1, \sigma_1^2 \mathbf{I}_d)$, where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix and $\boldsymbol{\mu}_1$ contains the d NN weights found by SGD on the training set $\boldsymbol{Z}(\boldsymbol{S})$, a randomly chosen subset of n samples from the $2n$ available in $\widetilde{\boldsymbol{Z}}$. The parameter σ_1^2 is chosen to be as large as possible, to some finite precision, so that the training accuracy of the stochastic NN with weights drawn from $\mathcal{N}(W \mid \boldsymbol{\mu}_1, \sigma_1^2 \mathbf{I}_d)$ differs by no more than a specified threshold from the training accuracy of the deterministic NN with weights $\boldsymbol{\mu}_1$.

The marginal $P_{W|\widetilde{\boldsymbol{Z}}}$ can in principle be computed by averaging over all 2^n possible values of $\boldsymbol{S}$. While such an exact computation is prohibitively expensive, this indicates a principled way to choose a prior by approximately performing this procedure. To choose the prior, we therefore form a number of subsets of $\widetilde{\boldsymbol{Z}}$ and train an NN with SGD on each, denoting the average of the output weights as $\boldsymbol{\mu}_2$. We then find $\tilde{\sigma}_2^2$ so that the accuracy on $\widetilde{\boldsymbol{Z}}$ for the stochastic NN with weights drawn from $\mathcal{N}(W \mid \boldsymbol{\mu}_2, \tilde{\sigma}_2^2 \mathbf{I}_d)$ and the deterministic NN with weights $\boldsymbol{\mu}_2$ differs by no more than a specified threshold. Based on $\tilde{\sigma}_2^2$, we create a set of candidate values for σ_2^2. We then set $Q_{W|\widetilde{\boldsymbol{Z}}} = \mathcal{N}(W \mid \boldsymbol{\mu}_2, \sigma_2^2 \mathbf{I}_d)$, where σ_2^2 is chosen so as to minimize the bound, typically leading to $\sigma_1^2 = \sigma_2^2$. For this final bound to be valid, we take a union bound over the set of candidate values. A detailed description of the experimental setup and additional results are given in Appendix B. While we focus on NNs, we note that the procedure given here can be used to obtain bounds for perturbed versions of any deterministic parametric hypothesis. For such bounds to be meaningful, the hypothesis needs to be relatively insensitive to the added Gaussian noise, so that the training loss of the perturbed hypothesis is not too different from the unperturbed training loss.

For each setting, we perform simulations over 10 instances of $\boldsymbol{S}$. Our results are obtained by setting $\delta \approx 0.001$ as the confidence parameter. However, due to the union bound

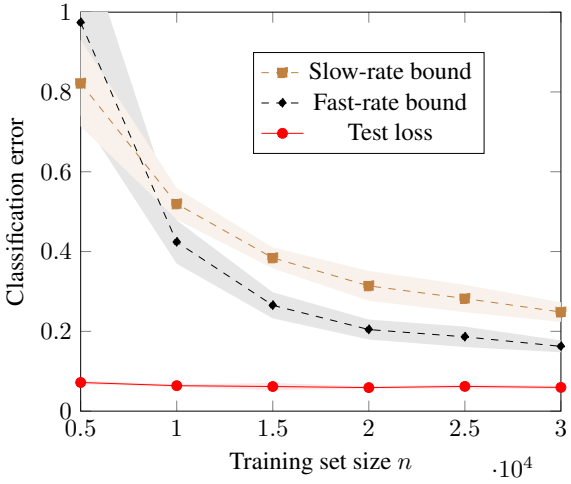

Figure 1. The estimated test loss, as well as the slow-rate (4) and fast-rate (5) bounds on the test loss, for LeNet-5 trained on MNIST with SGD until a training loss of 0.05 is reached. The shaded regions correspond to two standard deviations.

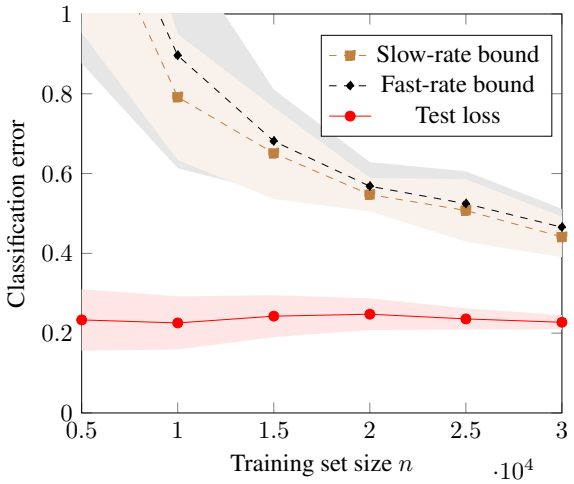

Figure 2. Same as Figure 1, but for Fashion-MNIST.

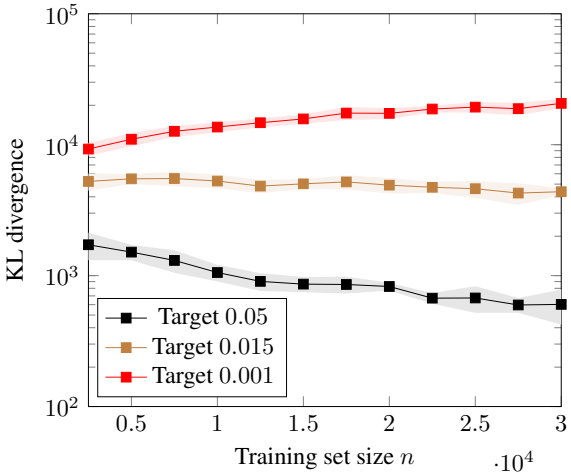

Figure 3. The n-dependence of the KL divergence in (4) and (5) for various target training losses.

over the set of candidates for σ_2^2, the presented bounds hold with probability at least 95%. The test losses and training losses are computed empirically by averaging over the performance of 5 NNs whose weights are sampled from $\mathcal{N}(W \mid \boldsymbol{\mu}_1, \sigma_1^2 \mathbf{I}_d)$. In Table 1, we present our bounds, along with the actual test losses, for several architectures and data sets. The training losses are not given, since they are virtually indistinguishable from the test losses for all of the settings we consider. The quantitative values of the bounds in Table 1 are in line with previously reported results for a similar setup (Dziugaite et al., 2020, Fig. 4), where a similar approach was used. The key differences is that the posterior therein is allowed to depend on the entire data set $\widetilde{\boldsymbol{Z}}$, whereas the training loss and prior depend on randomly selected disjoint subsets of $\widetilde{\boldsymbol{Z}}$. In contrast, in the random-subset setting considered in this paper, the prior is allowed to depend on the entire data set $\widetilde{\boldsymbol{Z}}$, whereas the training loss and posterior depend only on a randomly selected portion $\boldsymbol{Z}(\boldsymbol{S})$ of $\widetilde{\boldsymbol{Z}}$. For the MNIST data set, we see that the fast-rate bound is tighter than its slow-rate counterpart. For the more challenging Fashion-MNIST data set, the slow-rate bound is tighter. This is due to the fact that high training losses and large information measures penalize the fast-rate bound due to its larger constant factors.

In Figure 1, we vary the training set size n and train LeNet-5 on MNIST using SGD with momentum until a training loss of 0.05 is reached. For this setting, the fast-rate bound outperforms its slow-rate counterpart. For small training set sizes, our bounds are conservative. In Figure 2, we instead train LeNet-5 on the Fashion-MNIST data set until a training loss of 0.15 is reached. For this more challenging data set, the bounds are weaker, and the slow-rate bound tends to slightly outperform the fast-rate bound. This can be attributed to higher values of the training loss and relative entropy being obtained since the data set is more challenging. We note that the difference between the slow-rate and fast-rate bounds is less pronounced than in Table 1, where SGD without momentum is used. The results also display a lot more variance, especially for smaller sample sizes.

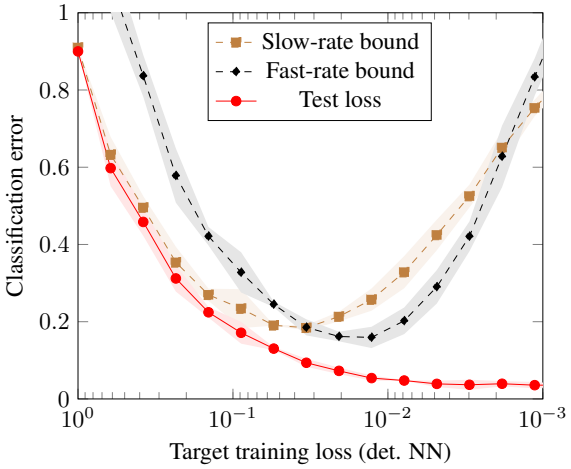

Figure 4. The estimated test loss, as well as the slow-rate (4) and fast-rate (5) bounds, with $n = 3 \cdot 10^4$ as a function of the targeted training loss for the underlying deterministic NN.

For the remaining experiments, we focus on LeNet-5 and MNIST. In Figure 3, we examine the n-dependence of the KL divergence in (4) and (5). For simplicity, we set $\sigma_1 = \sigma_2 = 0.01$. One factor that has a significant impact on the dependence is the target training loss for the underlying deterministic NN. For a target of 0.05, the KL divergence decreases with n, while it increases for a target of 0.001. For an intermediate target of 0.015, it is roughly constant over the studied range.

To more directly probe the effect on our bounds of the target training loss of the underlying deterministic NN, in Figure 4, we fix the training set size $n = 3 \cdot 10^4$ and vary the target training loss.While the bounds track the actual test loss reasonably well for targeted training losses down to around 0.02, the bounds increase rapidly after this point, whereas the actual test loss keeps on decreasing. This observation, which is in line with what is reported in (Dziugaite et al., 2020), illustrates that our PAC-Bayes bounds become vacuous when SGD is run until a very small training error is achieved. We empirically show in Appendix B.3 that this undesired behavior can be mitigated by selecting higher values for σ_1 and σ_2, at the cost of a higher test loss.

References

Ambroladze, A., Parrado-Hernandez, E., and Shawe-Taylor, J. Tighter PAC-Bayes bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, pp. 9–16, Vancouver, Canada, Dec. 2006.

Bassily, R., Moran, S., Nachum, I., Shafer, J., and Yehudayoff, A. Learners that use little information. *J. of Mach. Learn. Res.*, 83:25–55, Apr. 2018.

Bu, Y., Zou, S., and Veeravalli, V. V. Tightening mutual information-based bounds on generalization error. *IEEE J. Sel. Areas Inf. Theory*, 1(1):121–130, May 2020.

Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. IMS Lecture Notes Monogr. Ser., 2007.

Dziugaite, G., Hsu, K., Gharbieh, W., and Roy, D. On the role of data in PAC-Bayes bounds, June 2020. URL https://arxiv.org/abs/2006.10929.

Esposito, A., Gastpar, M., and Issa, I. Generalization error bounds via Rènyi f-divergences and maximal leakage. *arXiv*, Dec. 2019. URL http://arxiv.org/abs/1912.01439.

Guedj, B. A primer on PAC-Bayesian learning. *arXiv*, Jan. 2019. URL http://arxiv.org/abs/1901.05353.

Haghifam, M., Negrea, J., Khisti, A., Roy, D., and Dziugaite, G. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *arXiv*, Apr. 2020. URL http://arxiv.org/abs/2004.12983.

Hellström, F. and Durisi, G. Fast-rate loss bounds via conditional information measures with applications to neural networks. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, July 2021.

Hellström, F. and Durisi, G. Generalization bounds via information density and conditional information density. *IEEE J. Sel. Areas Inf. Theory*, 1(3):824–839, Dec. 2020.

McAllester, D. Some PAC-Bayesian theorems. In *Proc. Conf. Learn. Theory (COLT)*, Madison, WI, July 1998.

Negrea, J., Haghifam, M., Dziugaite, G., Khisti, A., and Roy, D. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

Polyanskiy, Y. and Wu, Y. *Lecture Notes On Information Theory*. 2019. URL http://www.stat.yale.edu/%7Eyw562/teaching/itlectures.pdf.

Rivasplata, O., Kuzborskij, I., Szepesvari, C., and Shawe-Taylor, J. PAC-Bayes analysis beyond the usual bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, pp. 16833–16845, Vancouver, Canada, Dec. 2020.

Rodríguez-Gálvez, B., Bassi, G., Thobaben, R., and Skoglund, M. On random subset generalization error bounds and the stochastic gradient Langevin dynamics algorithm. In *Inf. Theory Workshop (ITW)*, Riva del Garda, Italy, 4 2020.

Russo, D. and Zou, J. Controlling bias in adaptive data analysis using information theory. In *Proc. Artif. Intell. Statist. (AISTATS)*, Cadiz, Spain, May 2016.

Steinke, T. and Zakynthinou, L. Reasoning about generalization via conditional mutual information. In *Proc. Conf. Learn. Theory (COLT)*, Graz, Austria, July 2020.

Viallard, P., Germain, P., Habrard, A., and Morvant, E. A general framework for the derandomization of PAC-Bayesian bounds, 2021. URL https://arxiv.org/abs/2102.08649.

Wainwright, M. J. *High-Dimensional Statistics: a Non-Asymptotic Viewpoint*. Cambridge Univ. Press, Cambridge, U.K., 2019.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, Dec. 2017.

Zhou, R., Tian, C., and Liu, T. Individually conditional individual mutual information bound on generalization error, 2020. URL https://arxiv.org/abs/2012.09922.

Zhou, W., Veitch, V., Austern, M., Adams, R., and Orbanz, P. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *Proc. Int. Conf. Learn. Representations (ICLR)*, New Orleans, LA, USA, May 2019.

A. Proofs and Additional Results

In this section, we prove Theorem 1 and present some additional results. First, we explicitly state and prove the exponential inequalities used in the proof of Theorem 1.

Theorem 2 *Consider the random-subset setting introduced in Section 1. Let $W \in \mathcal{W}$ be distributed according to $P_{W|\boldsymbol{Z}(\boldsymbol{S})}$. Let $\lambda, \gamma > 0$ be constants such that $\lambda(1-\gamma) + (e^{\lambda} - 1 - \lambda)(1+\gamma^2) \leq 0$. Furthermore, let $Q_{W|\widetilde{\boldsymbol{Z}}}$ be a data-dependent conditional prior such that the distributions $Q_{W|\widetilde{\boldsymbol{Z}}} P_{\widetilde{\boldsymbol{Z}}} P_{\boldsymbol{S}}$ and $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}} = P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}} P_{\widetilde{\boldsymbol{Z}}} P_{\boldsymbol{S}}$ are absolutely continuous with respect to each other. Then, the following inequalities hold:*

$$\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg[\exp\Bigg(\lambda n\left(L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) - \gamma L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right) - \log\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}Q_{W|\widetilde{\boldsymbol{Z}}} P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg)\Bigg] \leq 1. \quad (6)$$

$$\mathbb{E}_{P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg[\exp\Bigg(\frac{n-1}{2}(L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) - L_{\boldsymbol{Z}(\boldsymbol{S})}(W))^2 - \log\sqrt{n} - \log\frac{\mathrm{d}P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}Q_{W|\widetilde{\boldsymbol{Z}}} P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg)\Bigg] \leq 1. \quad (7)$$

Proof We begin by proving an exponential inequality for a binary random variable X satisfying $P(X=a) = P(X=b) = 1/2$ where $a, b \in [0,1]$. Let $\bar{X} = b$ if $X = a$ and $\bar{X} = a$ if $X = b$. Finally, let $c = e^{\lambda} - 1 - \lambda$. Then,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X-\gamma\bar{X})}\right] &\leq \mathbb{E}\left[1 + \lambda\left(X - \gamma\bar{X}\right) + c\left(X - \gamma\bar{X}\right)^2\right] \\ &= 1 + \frac{\lambda(1-\gamma)}{2}(a+b) + \frac{c}{2}(a-\gamma b)^2 + \frac{c}{2}(b-\gamma a)^2. \end{aligned} \quad (8)$$

Here, the first inequality follows because $e^y \leq 1 + y + cy^2/\lambda^2$ for all $y \leq \lambda$. Expanding the squares and removing negative terms, we find that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X-\gamma\bar{X})}\right] &\leq 1 + \lambda(1-\gamma) + (e^{\lambda} - 1 - \lambda)(1+\gamma^2) \\ &\leq 1, \end{aligned} \quad (9)$$

where (9) follows from our assumption on λ, γ. Let $Q_{W\widetilde{\boldsymbol{Z}}} = Q_{W|\widetilde{\boldsymbol{Z}}} P_{\widetilde{\boldsymbol{Z}}}$, and apply (9) with $X = \ell(w, Z_i(\bar{S}_i))$ and $\bar{X} = \ell(w, Z_i(S_i))$ for some fixed w and $\widetilde{\boldsymbol{z}}$. It follows that

$$\begin{aligned}&\mathbb{E}_{Q_{W\widetilde{\boldsymbol{Z}}} P_{\boldsymbol{S}}}\left[e^{\lambda n\left(L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) - \gamma L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right)}\right] \\ &= \mathbb{E}_{Q_{W\widetilde{\boldsymbol{z}}}}\left[\prod_{i=1}^{n} \mathbb{E}_{P_{S_i}}\left[e^{\lambda\left(\ell(W, Z_i(\bar{S}_i)) - \gamma\ell(W, Z_i(S_i))\right)}\right]\right] \\ &\leq 1. \end{aligned} \quad (10)$$

The inequality (6) now follows after a change of measure to $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ (Polyanskiy & Wu, 2019, Prop. 17.1).

We obtain (7) as follows. Let $(w, \widetilde{\boldsymbol{z}})$ be fixed, and consider the random variable $\Delta(\boldsymbol{S}) = L_{\boldsymbol{z}(\bar{\boldsymbol{S}})}(w) - L_{\boldsymbol{z}(\boldsymbol{S})}(w)$. Due to the boundedness of $\ell(\cdot,\cdot)$ and the symmetry property $\Delta(\boldsymbol{S}) = -\Delta(\bar{\boldsymbol{S}})$, it follows that $\Delta(\boldsymbol{S})$ is $1/\sqrt{n}$-sub-Gaussian with mean 0 under $P_{\boldsymbol{S}}$. Applying (Wainwright, 2019, Thm. 2.6.(IV)) with $\lambda = 1 - 1/n$, we conclude that

$$\mathbb{E}_{P_{\boldsymbol{S}}}\left[\exp\left(\frac{n-1}{2}(L_{\boldsymbol{z}(\bar{\boldsymbol{S}})}(w) - L_{\boldsymbol{z}(\boldsymbol{S})}(w))^2\right)\right] \leq \sqrt{n}. \quad (11)$$

Taking the expectation with respect to $Q_{W|\widetilde{\boldsymbol{Z}}} P_{\widetilde{\boldsymbol{Z}}}$, changing measure to $P_{W\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$, and rearranging terms, we obtain (7). ■

Equipped with Theorem 2, we now prove Theorem 1.

Proof of Theorem 1 To derive (5), we first apply Jensen's inequality in (6) with respect to $P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$ to get

$$\mathbb{E}_{P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg[\exp\Bigg(\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\lambda n\left(L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) - \gamma L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right)\right] - D(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, Q_{W|\widetilde{\boldsymbol{Z}}})\Bigg)\Bigg] \leq 1. \quad (12)$$

We now use Markov's inequality in the following form. Let $U \sim P_U$ be a nonnegative random variable satisfying $\mathbb{E}[U] \leq 1$. Then,

$$P_U[U \leq 1/\delta] \geq 1 - \mathbb{E}[U]\,\delta \geq 1 - \delta. \quad (13)$$

Applying (13) to (12) we find that, with probability at least $1-\delta$ under $P_{\widetilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\exp\Bigg(\mathbb{E}_{P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\lambda n\left(L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) - \gamma L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right)\right] - D(P_{W|\widetilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, Q_{W|\widetilde{\boldsymbol{Z}}})\Bigg) \leq \frac{1}{\delta}. \quad (14)$$

Taking the logarithm and reorganizing terms, we obtain (5). By the same procedure, starting from (7) instead of (6), we obtain (4) after an additional use of Jensen's inequality. ■

Theorem 2 does not only allow us to derive PAC-Bayesian bounds, but also bounds on the average and single-draw test loss. We present these bounds in the following corollary. We note that, while (7) can be used to obtain an average bound, this includes a suboptimal dependence on n that can be avoided by using an alternative exponential inequality (Hellström & Durisi, 2020, Cor. 5). Hence, we do not present any slow-rate average bound here.

Corollary 3 *Consider the setting of Theorem 2. Then, the*

average population loss is bounded by

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \le \gamma\, \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \frac{\mathbb{E}_{P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, Q_{W|\tilde{\boldsymbol{Z}}})\right]}{\lambda n}. \quad (15)$$

Furthermore, with probability at least $1-2\delta$ *over* $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$*, the single-draw test loss is bounded by*

$$L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) \le \min\{B^{\mathrm{SD}}_{\mathrm{slow}}, B^{\mathrm{SD}}_{\mathrm{fast}}\} \quad (16)$$

where $B^{\mathrm{SD}}_{\mathrm{slow}}$ *and* $B^{\mathrm{SD}}_{\mathrm{fast}}$ *are defined as*

$$B^{\mathrm{SD}}_{\mathrm{fast}} = L_{\boldsymbol{Z}(\boldsymbol{S})}(W) + \sqrt{\frac{2}{n-1}\left(\log\frac{\mathrm{d}P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}Q_{W|\tilde{\boldsymbol{Z}}}P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}} + \log\frac{\sqrt{n}}{\delta}\right)} \quad (17)$$

$$B^{\mathrm{SD}}_{\mathrm{fast}} = \gamma L_{\boldsymbol{Z}(\boldsymbol{S})}(W) + \frac{\left(\log\frac{\mathrm{d}P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}Q_{W|\tilde{\boldsymbol{Z}}}P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}} + \log\frac{1}{\delta}\right)}{\lambda n}. \quad (18)$$

Proof First, we apply Jensen's inequality to (6) to move the expectation inside the exponential. We obtain (15) by taking the logarithm and reorganizing terms. To derive (18), we apply (13) to (6) to conclude that, with probability at least $1-\delta$ under $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\exp\left(\lambda n\big(L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) - \gamma L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\big) - \log\frac{\mathrm{d}P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}Q_{W|\tilde{\boldsymbol{Z}}}P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\right) \le \frac{1}{\delta}. \quad (19)$$

We obtain (18) by reorganizing terms. Similarly, (17) follows by applying (13) to (7) and reorganizing terms. ■

Setting $Q_{W|\tilde{\boldsymbol{Z}}} = P_{W|\tilde{\boldsymbol{Z}}}$, $\gamma = 2$ and $\lambda = 1/3$ in (15), we recover the CMI bound in (Steinke & Zakynthinou, 2020).

As illustrated in Corollary 4 below, for the special case where $Q_{W|\tilde{\boldsymbol{Z}}} = P_{W|\tilde{\boldsymbol{Z}}}$, the bound on the average population loss in (15) can be tightened by replacing the CMI $\mathbb{E}_{P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}})\right] = I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})$ with a sum of samplewise CMIs $I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n})$.

Corollary 4 *Consider the setting of Theorem 2, with the additional assumption that* $Q_{W|\tilde{\boldsymbol{Z}}} = P_{W|\tilde{\boldsymbol{Z}}}$*. Then, the average population loss is bounded by*

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \le \gamma\, \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \sum_{i=1}^{n}\frac{I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n})}{\lambda n}. \quad (20)$$

Proof Consider a fixed $w \in \mathcal{W}$ and $\widetilde{\boldsymbol{z}} \in \mathcal{Z}^{2n}$. By (9),

$$\mathbb{E}_{P_{S_i}}\left[e^{\lambda\left(\ell(w,Z_i(\bar{S}_i)) - \gamma\ell(w,Z_i(S_i))\right)}\right] \le 1. \quad (21)$$

Let $\widetilde{Z}_i^{\pm}$ denote the pair $(\tilde{Z}_i, \tilde{Z}_{i+n})$ and let $P_{S_i|w\tilde{z}_i^{\pm}}$ denote $P_{S_i|W=w,\tilde{Z}_i^{\pm}=\tilde{z}_i^{\pm}}$ for some fixed $w, \tilde{z}_i^{\pm}$. Note that $Z_i(S_i)$ depends on $\widetilde{\boldsymbol{Z}}$ only through $\widetilde{Z}_i^{\pm}$. By changing measure to $P_{S_i|w\tilde{z}_i^{\pm}}$ we obtain

$$\begin{aligned}&\mathbb{E}_{P_{S_i}}\left[e^{\lambda\left(\ell(w,Z_i(\bar{S}_i)) - \gamma\ell(w,Z_i(S_i))\right)}\right] \\ &= \mathbb{E}_{P_{S_i|w\tilde{z}_i^{\pm}}}\left[e^{\lambda\left(\ell(w,Z_i(\bar{S}_i)) - \gamma\ell(w,Z_i(S_i))\right) - \log\frac{\mathrm{d}P_{S_i|w\tilde{z}_i^{\pm}}}{\mathrm{d}P_{S_i}}}\right] \\ &\le 1. \end{aligned} \quad (22)$$

By Jensen's inequality and taking the logarithm, we obtain

$$\begin{aligned}&\mathbb{E}_{P_{S_i|w\tilde{z}_i^{\pm}}}\left[\ell(w,Z_i(\bar{S}_i))\right] \\ &\le \gamma\, \mathbb{E}_{P_{S_i|w\tilde{z}_i^{\pm}}}[\ell(w,Z_i(S_i))] + \frac{1}{\lambda}\,\mathbb{E}_{P_{S_i|w\widetilde{\boldsymbol{z}}}}\left[\log\frac{\mathrm{d}P_{S_i|w\tilde{z}_i^{\pm}}}{\mathrm{d}P_{S_i}}\right] \\ &= \gamma\, \mathbb{E}_{P_{S_i|w\tilde{z}_i^{\pm}}}[\ell(w,Z_i(S_i))] + \frac{D(P_{S_i|w\tilde{z}_i^{\pm}} \,||\, P_{S_i})}{\lambda}. \end{aligned} \quad (23)$$

The desired result follows by noting that, by marginalizing,

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] = \sum_{i=1}^{n}\mathbb{E}_{P_{W\tilde{Z}_i^{\pm}}}\left[\mathbb{E}_{P_{S_i|W\tilde{Z}_i^{\pm}}}\left[\frac{\ell(W,Z_i(\bar{S}_i))}{n}\right]\right] \quad (24)$$

and applying (23) to each term in the sum in (24). ■

The bound in (Rodríguez-Gálvez et al., 2020, Prop. 3) (also found in (Zhou et al., 2020, Cor. 1)), which is a tightening of (Haghifam et al., 2020, Thm. 3.1), depends on a sum of square roots $\sqrt{I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n})}$, whereas our bound does not have a square root. Since $I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n}) \le H(S_i) = \log(2)$, it follows that $I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n}) < \sqrt{I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n})}$. This implies that our bound improves the dependence on $I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n})$ at the cost of greater constants. By (Rodríguez-Gálvez et al., 2020, Lem. 3) (or (Zhou et al., 2020, Lem. 2)), we also have $I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n}) \le I(W;S_i|\tilde{\boldsymbol{Z}})$. Thus, the samplewise bounds presented in Corollary 4 and Corollary 7 are tighter than those reported in (Hellström & Durisi, 2021, Cor. 3, Cor. 6).

For the so-called interpolating setting, where $L_{\boldsymbol{Z}(\boldsymbol{S})}(W) = 0$, one can obtain a different exponential inequality than the one in (6), under the additional assumption that $Q_{W|\tilde{\boldsymbol{Z}}} = P_{W|\tilde{\boldsymbol{Z}}}$. This leads to tighter bounds than the ones in (5), (15), and (18). Specifically, in these alternative bounds,

the factor λ can be set to $\log 2 \approx 0.69$. In contrast, any λ in Theorem 2, regardless of the value of γ, must satisfy $\lambda^2 - 4(e^\lambda - 1)(e^\lambda - 1 - \lambda) \geq 0$, which implies $\lambda < 0.37$. We first prove the following inequality, the derivation of which is similar to part of the proof of (Steinke & Zakynthinou, 2020, Thm. 5.7).

Theorem 5 *Consider the random-subset setting introduced in Section 1. Let $W \in \mathcal{W}$ be distributed according to $P_{W|\boldsymbol{Z}(\boldsymbol{S})}$, and assume that $L_{\boldsymbol{Z}(\boldsymbol{S})}(W) = 0$ a.s. for $W \sim P_{W|\boldsymbol{Z}(\boldsymbol{S})}$. Then,*

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\exp\Big(n\log 2 \cdot L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) - \imath(W, \boldsymbol{S}|\tilde{\boldsymbol{Z}})\Big)\right] \leq 1. \tag{25}$$

Proof Let $\lambda, \gamma > 0$. Then,

$$\begin{aligned}
&\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\Bigg[\prod_{i=1}^{n}\Big(\frac{1}{2}e^{\lambda\ell(W,Z_i(\bar{S}_i))-\gamma\ell(W,Z_i(S_i))} \\
&\qquad + \frac{1}{2}e^{\lambda\ell(W,Z_i(S_i))-\gamma\ell(W,Z_i(\bar{S}_i))}\Big)\Bigg] \\
&\quad = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\left[\prod_{i=1}^{n} e^{\lambda\ell(W,Z_i(\bar{S}_i))-\gamma\ell(W,Z_i(S_i))}\right].
\end{aligned} \tag{26}$$

It follows from (26) that

$$\begin{aligned}
&\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\left[e^{n\left(\lambda L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)-\gamma L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right)}\right] \\
&\quad = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\prod_{i=1}^{n}\left(\frac{1}{2}e^{\lambda\ell(W,Z_i(\bar{S}_i))-\gamma\ell(W,Z_i(S_i))}\right.\right. \\
&\qquad \left.\left. + \frac{1}{2}e^{\lambda\ell(W,Z_i(S_i))-\gamma\ell(W,Z_i(\bar{S}_i))}\right)\right].
\end{aligned} \tag{27}$$

We now change measure to $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ to conclude that

$$\begin{aligned}
&\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[e^{n\left(\lambda L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)-\gamma L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right)-\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})}\right] \\
&\quad = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\prod_{i=1}^{n}\left(\frac{1}{2}e^{\lambda\ell(W,Z_i(\bar{S}_i))-\gamma\ell(W,Z_i(S_i))}\right.\right. \\
&\qquad \left.\left. + \frac{1}{2}e^{\lambda\ell(W,Z_i(S_i))-\gamma\ell(W,Z_i(\bar{S}_i))}\right)\right].
\end{aligned} \tag{28}$$

We now use the interpolating assumption and set $\lambda = \log 2$. If $\ell(W, Z_i(\bar{S}_i)) = 0$, (25) holds trivially for every γ. If $\ell(W, Z_i(\bar{S}_i)) > 0$, we let $\gamma \to \infty$. This, together with the assumption that $\ell(W, Z_i(\bar{S}_i)) \in [0, 1]$, implies (25). ■

Using Theorem 5, we can derive bounds that are analogous to those in (5), (15), and (18). We present these bounds below without proof, since they can be established following steps similar to the ones used to prove (5), (15), and (18).

Corollary 6 *Consider the setting of Theorem 5. Then, the average population loss is bounded by*[1]

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \leq \frac{I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n\log 2}. \tag{29}$$

Furthermore, with probability at least $1-\delta$ over $P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}$, the PAC-Bayesian population loss is bounded by

$$\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)\right] \leq \frac{D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W|\tilde{\boldsymbol{Z}}}) + \log\frac{1}{\delta}}{n\log 2}. \tag{30}$$

Finally, with probability at least $1-\delta$ over $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$, the single-draw population loss is bounded by

$$L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) \leq \frac{\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}}) + \log\frac{1}{\delta}}{n\log 2}. \tag{31}$$

Finally, we present a samplewise average bound for the interpolating setting, which is tighter than (29) by (Haghifam et al., 2020, Rem. 3.5). This result tightens (Rodríguez-Gálvez et al., 2020, Prop. 3) and (Zhou et al., 2020, Cor. 1) for the interpolating setting.

Corollary 7 *Consider the setting of Theorem 5. Then, the average population loss is bounded by*

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \leq \sum_{i=1}^{n}\frac{I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n})}{n\log 2}. \tag{32}$$

Proof Let $\lambda, \gamma > 0$. For all i, by arguing as in (26)—(28),

$$\begin{aligned}
&\mathbb{E}_{P_{W\tilde{Z}_i^{\pm S_i}}}\left[e^{\lambda\ell(W,Z_i(\bar{S}_i))-\gamma\ell(W,Z_i(S_i))-\imath(W,S_i|\tilde{Z}_i,\tilde{Z}_{i+n})}\right] \\
&\quad = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\left(\frac{1}{2}e^{\lambda\ell(W,Z_i(\bar{S}_i))-\gamma\ell(W,Z_i(S_i))}\right.\right. \\
&\qquad \left.\left. + \frac{1}{2}e^{\lambda\ell(W,Z_i(S_i))-\gamma\ell(W,Z_i(\bar{S}_i))}\right)\right].
\end{aligned} \tag{33}$$

Here, $\imath(W, S_i|\tilde{Z}_i, \tilde{Z}_{i+n}) = \log\frac{\mathrm{d}P_{W\tilde{Z}_i^{\pm S_i}}}{\mathrm{d}P_{W\tilde{Z}_i^{\pm}}P_{S_i}}$. Next, we use the interpolating assumption and set $\lambda = \log 2$. If $\ell(W, Z_i(\bar{S}_i)) > 0$, let $\gamma \to \infty$. Since $\ell(W, Z_i(\bar{S}_i)) \in [0, 1]$, this implies that the right-hand side of (33) is at most 1. If $\ell(W, Z_i(\bar{S}_i)) = 0$, this holds trivially for every γ. Thus,

$$\mathbb{E}_{P_{W\tilde{Z}_i^{\pm S_i}}}\left[e^{\log 2\cdot\ell(W,Z_i(\bar{S}_i))-\imath(W,S_i|\tilde{Z}_i,\tilde{Z}_{i+n})}\right] \leq 1. \tag{34}$$

Using Jensen's inequality and reorganizing terms, we obtain

$$\mathbb{E}_{P_{W\tilde{Z}_i^{\pm S_i}}}\left[\ell(W, Z_i(\bar{S}_i))\right] \leq \frac{I(W;S_i|\tilde{Z}_i,\tilde{Z}_{i+n})}{\log 2}. \tag{35}$$

[1] Since $I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq \log 2^n$ for all distributions, the constant $\log 2$ cannot be improved.

Table 2. The LeNet-5 architecture used in Section 3.

Convolutional layer, 20 units, 5×5 size, linear activation, 1×1 stride, valid padding
Max pooling layer, 2×2 size, 2×2 stride
Convolutional layer, 50 units, 5×5 size, linear activation, 1×1 stride, valid padding
Max pooling layer, 2×2 size, 2×2 stride
Flattening layer
Fully connected layer, 500 units, ReLU activation
Fully connected layer, 10 units, $\mathrm{softmax}$ activation

The result now follows because

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{W\tilde{Z}_i^{\pm}\boldsymbol{S}}}\left[\ell(W, Z_i(\bar{S}_i))\right] . \blacksquare \tag{36}$$

B. Experiment Details and Additional Results

B.1. Architectures

The LeNet-5 architecture used in Section 3 is described in Table 2. While it differs from most implementations of LeNet-5, it coincides with the architecture used by (Dziugaite et al., 2020) and (Zhou et al., 2019). It has $431\,080$ parameters. For the binarized MNIST data set considered in Table 3, the number of output units is instead 2, resulting in a network with $427\,072$ parameters. The fully connected neural network denoted by 600^2 consists of an input layer with 784 units, 2 fully connected layers with 600 units and ReLU activations, followed by an output layer with 10 units and $\mathrm{softmax}$ activations. It has $837\,610$ parameters.

B.2. Training procedures

We now provide additional details on the training procedures described in Section 3. The initial weights of all the networks used for each instance of $\boldsymbol{Z}(\boldsymbol{S})$ were set to the same randomly selected values drawn from a zero-mean normal distribution with standard deviation 0.01. All networks were trained using the SGD with a batch size of 512 on the cross-entropy loss, either with momentum and a fixed learning rate or without momentum and a decaying learning rate. First, we describe the details of SGD with momentum. Unless otherwise stated, we used a learning rate of 0.001 for MNIST, and for Fashion-MNIST, we used 0.003. For Figure 4, we used a learning rate of $3 \cdot 10^{-4}$, leading to a better resolution with respect to the obtained training losses. In all experiments, the momentum parameter is set to 0.9. For SGD without momentum we used a decaying learning rate where the learning rate α for epoch E is given by

$$\alpha(E) = \frac{\alpha_0}{1 + \beta \cdot \lfloor E/E_0 \rfloor}. \tag{37}$$

Here, α_0 is the initial learning rate, β is the decay rate, and E_0 is the number of epochs between each decay. In all experiments, we used $\alpha_0 = 0.01$, $\beta = 2$, and $E_0 = 20$.

To choose σ_1, we pick the largest value with one significant digit (i.e., of the form $a \cdot 10^{-b}$ with $a \in [1:9]$ and $b \in \mathbb{Z}$) such that the absolute difference between the training loss on $\boldsymbol{Z}(\boldsymbol{S})$ of the deterministic network with weights $\boldsymbol{\mu}_1$ and empirical average of the training loss of 5 NNs with weights drawn independently from $\mathcal{N}(W \mid \boldsymbol{\mu}_1, \sigma_1^2 \mathbf{I}_d)$ was no larger than some specified threshold. Unless otherwise stated, for MNIST, we use a threshold of 0.05 for selecting σ_1. For Fashion-MNIST, we use a threshold of 0.10.

Next, $\boldsymbol{\mu}_2$ is determined as follows. We form 10 subsets of $\widetilde{\boldsymbol{Z}}$, each of size n. The first subset contains the first n samples in $\widetilde{\boldsymbol{Z}}$, the last contains the last n samples in $\widetilde{\boldsymbol{Z}}$, and the remaining subsets contain the linearly spaced sequences in between. We then train one NN on each subset and denote the average of the final weights of these networks by $\boldsymbol{\mu}_2$. To find σ_2, we use as starting point the same procedure as for determining σ_1, but with $\boldsymbol{\mu}_2$ in place of $\boldsymbol{\mu}_1$ and the training loss evaluated on all of $\widetilde{\boldsymbol{Z}}$. Let us call the value found by this procedure $\sigma_2' = a' \cdot 10^{-b'}$. Then, among the values of the form $a \cdot 10^{-b}$ with $a \in [1:9]$ and $b \in \{b'-1, b', b'+1\}$, we choose σ_2 to be the one that minimizes the bound on the test loss. In all our experiments, this procedure resulted in $\sigma_2 = \sigma_1$. To guarantee that the final bound holds with a given confidence level, all 27 bounds resulting from all possible choices of a and b need to hold with the same confidence level. Since we consider both slow-rate and fast-rate bounds, a total of 54 bounds need to hold simultaneously. We ensure that this is the case via the union bound. Thus, if each individual bound holds with probability at least $1 - \delta$, the optimized bounds hold with probability at least $1 - 54\delta$. We compute the bounds with $\delta = 0.05/54$, so the optimized bounds hold with 95% confidence.

B.3. Additional numerical results

In Table 3, we replace a portion of the data labels with a randomly chosen label, and study how the proportion of corrupt data affects our bounds. To make training with randomized labels more efficient, we use a binarized version of MNIST where the digits $0, \ldots, 4$ are combined into one class and the digits $5, \ldots, 9$ into another. The results show that our bounds become vacuous when randomized labels

Table 3. The estimated training losses, test losses, and corresponding slow-rate (4) and fast-rate (5) bounds on the test loss for LeNet-5 trained on binarized MNIST with partially corrupted labels.

Randomized labels	25%	50%	75%	100%
Training loss	0.106	0.088	0.090	0.081
Test loss	0.216	0.364	0.461	0.494
Slow-rate bound	5.561	9.811	10.45	11.67
Fast-rate bound	44.52	141.0	160.1	200.4

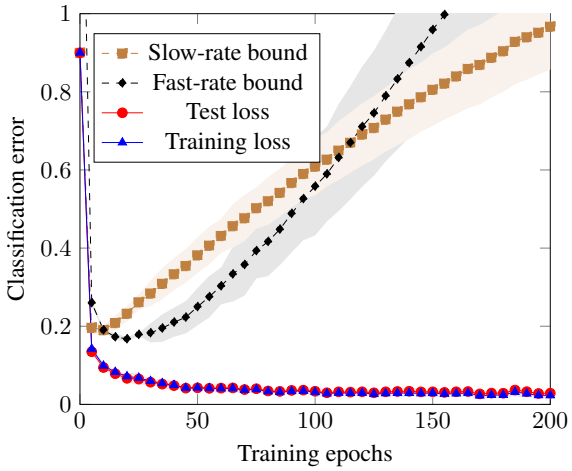

Figure 5. The dependence of the bounds on the training epoch when the threshold 0.05 is used to select σ_1 and σ_2.

are used. The fast-rate bound is significantly worse than its slow-rate counterpart, which is to be expected: when the prior and posterior are selected using randomized labels, a larger discrepancy between them arises. This increases the value of the KL divergence in (4) and (5), which penalizes the fast-rate bound. We note that the qualitative behavior of the bounds is in agreement with the empirically evaluated test error: an increased proportion of randomized labels, and thus an increased test error, increases our bounds.

In Figure 5 and Figure 6, we use SGD with momentum and investigate the role of the variances σ_1 and σ_2 by varying the threshold used to select them. Specifically, in Figure 5, we set the threshold used to determine σ_1 and σ_2 to 0.05, which leads to small values for σ_1 and σ_2. In Figure 6, we use a threshold of 0.15 instead, which allows for larger variances. The results confirm the intuition that larger variances yield better test-loss bounds at the cost of a higher test error.

Finally, in Figure 7 and Figure 8, we investigate the n-dependence of our bounds for a 600^2 FCNN. In Figure 7, we train a 600^2 FCNN on MNIST, and in Figure 8, we consider Fashion-MNIST. The resulting bounds are slightly weaker than in Figure 1 and 2, where LeNet-5 is used, but the overall behavior is very similar.

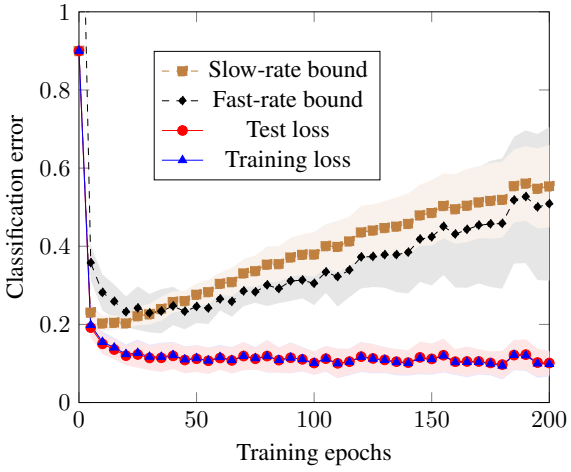

Figure 6. The dependence of the bounds on the training epoch when the threshold 0.15 is used to select σ_1 and σ_2.

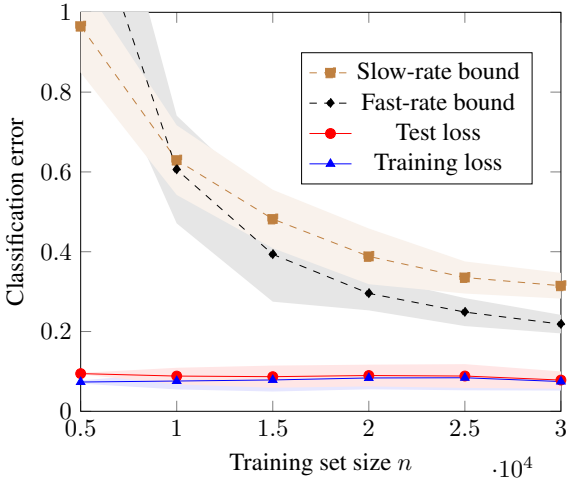

Figure 7. Same as Figure 1, but for a 600^2 FCNN.

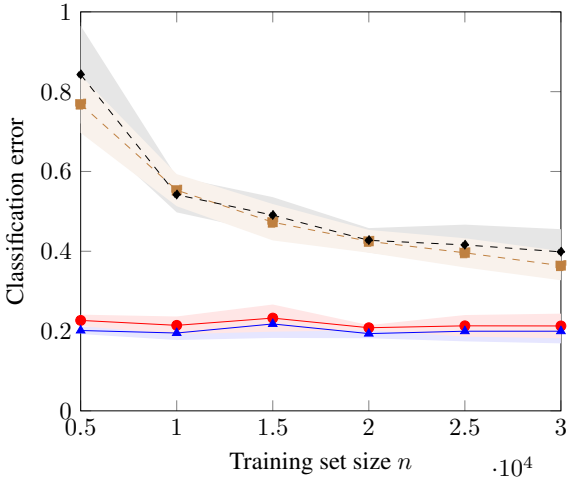

Figure 8. Same as Figure 2, but for a 600^2 FCNN.