Dissecting Non-Vacuous Generalization Bounds based on the Mean-Field Approximation

Konstantinos Pitas [1]

Abstract

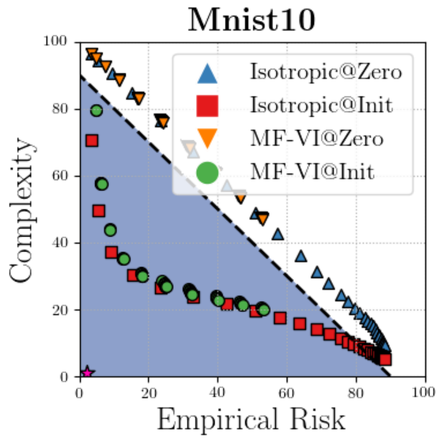
Explaining how overparametrized neural networks simultaneously achieve low risk and zero empirical risk on benchmark datasets is an open problem. PAC-Bayes bounds optimized using variational inference (VI) have been recently proposed as a promising direction in obtaining non-vacuous bounds. We show empirically that this approach gives negligible gains when modeling the posterior as a Gaussian with diagonal covariance—known as the mean-field approximation. We investigate common explanations, such as the failure of VI due to problems in optimization or choosing a suboptimal prior. Our results suggest that investigating richer posteriors is the most promising direction forward.

1. Introduction

Two recent works (Dziugaite & Roy, 2017; Zhou et al., 2018) based on the PAC-Bayes framework (McAllester, 1999) have made remarkable progress towards explaining why overparametrized neural networks simultaneously achieve low risk and zero empirical risk on benchmark datasets. PAC-Bayes bounds deal with randomized classifiers with posterior and prior distributions $\hat{\rho}$ and π respectively. Given that typically one wants to bound the risk of a deterministic classifier f the posterior $\hat{\rho}$ is chosen to be in some sense close to f(i.e. it is usually centered at f). Then, PAC-Bayes theorems make statements that are roughly of the form

$$\mathbf{E}\mathcal{L}(\hat{\rho}) \leq \mathbf{E}\hat{\mathcal{L}}(\hat{\rho}) + \beta \mathrm{KL}(\hat{\rho}||\pi), \tag{1}$$

where $\mathcal{L}(\hat{\rho})$ is the risk, $\hat{\mathcal{L}}(\hat{\rho})$ is the empirical risk and the expectation is over the posterior. The $\beta \mathrm{KL}(\hat{\rho}||\pi)$ term between the prior and posterior acts as a measure of complexity for the classifier.



(a)

Figure 1. **Risk-Complexity plot for MNIST 10**: The area below the dashed line corresponds to non-vacuous pairs of (complexity, empirical risk). The purple star corresponds to the optimal bound implied by the testing set. We parametrize the PAC-Bayes bound with different combinations of diagonal Gaussian priors and posteriors. "Isotropic@" corresponds to isotropic priors and posteriors with the prior centered at 0 and at the deep neural network random initialization. Similarly for "Mean-Field VI@" the posterior is diagonal but non-isotropic and we optimize it with variational inference. Choosing the prior mean to be the random initialization improves the bounds greatly in both cases. On the contrary when optimizing with mean-field variational inference there is negligible improvement over the isotropic case.

The RHS of (1) corresponds to a variational encoding scheme of the deep neural network weights, where the variance of the noise in the posterior measures the level of precision used in the encoding (Blier & Ollivier, 2018). In Dziugaite & Roy (2017), the authors minimize this variational code directly using a differentiable surrogate, by parameterizing the prior and posterior as Gaussians, and optimizing using stochastic variational inference (Hoffman et al., 2013; Kingma & Welling, 2013). They obtain non-vacuous generalization bounds on a simplified MNIST(LeCun & Cortes, 2010) dataset, but are unable to scale their result to larger

[1] École Polytechnique Fédérale de Lausanne, Switzerland. Correspondence to: Konstantinos Pitas <konstantinos.pitas@epfl.ch>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

problems.

Stochastic variational inference is known to result in poor weight encodings, but the reasons behind this are unclear (Blier & Ollivier, 2018). Variational inference, in the context of Bayesian neural networks, is thought to suffer from high gradient variance (Kingma et al., 2015; Wu et al., 2018; Wen et al., 2018). In addition, correlations between parameters are often omitted, as storing and manipulating the full covariance matrix is computationally infeasible. This can be seen as adding independent noise to each weight, an approximation know as *mean-field*. This might be too restrictive in deriving useful posteriors (Ritter et al., 2018; Mishkin et al., 2018), and therefore tight codes.

Consequently, in Zhou et al. (2018) the authors first compress deep neural networks by sparsifying them and deriving a variational code on the remaining parameters. Off the shelf compression algorithms compress remarkably well and thus Zhou et al. (2018) obtain non-vacuous but loose bounds for the much more complex Imagenet (Deng et al., 2009). A significant drawback of this approach is that the bound is derived for a network whose parameters are not similar *even in expectation* to the original ones (Suzuki, 2019).

We thus focus on analyzing the case of applying variational inference directly on the original weights. Importantly, we lack meaningful comparison tools. The techniques in Dziugaite & Roy (2017); Zhou et al. (2018) actually provide *multiple* bounds corresponding to different levels of encoding precision of the weights, which is usually controlled by the the parameter β in (1). However, results are presented in single (empirical risk, complexity) pairs, making drawing conclusions difficult.

Our first contribution is thus to introduce "Risk-Complexity" plots 1. On the x-axis we plot the Empirical Risk $\hat{\mathcal{L}}(\hat{\rho})$, while on the y-axis we plot the estimated Complexity $\beta\mathrm{KL}(\hat{\rho}||\pi)$ or the equivalent complexity metric. The plots have a number of advantages. We can easily plot the region of non-vacuity and the location of the best possible bound implied by the *testing* set. For an optimization based bound method we can then derive multiple (complexity, empirical risk) estimates and plot a Pareto front of all combinations. This results in an intuitive way for comparing bounding methods where one can simply inspect the Pareto fronts in relation to the best possible pair implied by the testing set.

Armed with our new visualization tools we are ready to scrutinize the results of Dziugaite & Roy (2017). The authors combine four elements in deriving non-vacuous bounds: i) changing the prior to be centered at the random initialization instead of at zero ii) optimizing the posterior covariance iii) optimizing the posterior mean iii) simplifying the classification problem by merging the 10 MNIST classes into 2 aggregate ones. In this way it is unclear what is the contribution of each to obtaining non-vacuous bounds.

In particular, separating the effects of i,ii and iii is important. Flatness at the minimum has been frequently cited as a desirable property for good generalization (Keskar et al., 2016). However, current results show mainly empirical correlations with generalization error (Keskar et al., 2016) and the exact effect of flatness is still debated (Dinh et al., 2017). Point ii is related to flatness at the minimum, as increased posterior variance while $\mathbf{E}\hat{\mathcal{L}}(\hat{\rho})$ remains small implies a flat minimum. Importantly, when relating PAC-Bayes to flatness one needs to keep the mean of the posterior fixed. Optimizing the mean and then the covariance corresponds to measuring the flatness of a *different* minimum.

Our contributions. Through detailed experiments we find that for diagonal Gaussian priors and posteriors the dominant element which turns a vacuous bound to non-vacuous is centering the prior at the random initialization instead of at 0. Optimizing the covariance using stochastic variational inference results in negligible or no gains. In fact, a simple isotropic Gaussian baseline in the prior and posterior results in nearly identical bound values.

We are then motivated to investigate two common explanations for this ineffectiveness. First it could be that stochastic variational inference has not properly converged. Secondly, PAC-Bayes theory allows improved bounds by choosing priors that reflect prior knowledge about the problem, as long as these priors don't depend on the training set. Choosing the random initialization to be the prior mean is already a good prior mean choice. It might be that through a better choice of prior *covariance* the mean-field approximation could yield meaningful improvements to the posterior covariance and hence the bound.

Through a simple theoretical analysis, we explore both of these explanations. Specifically, we leverage the fact that the loss landscape around the minimum is empirically quadratic, to derive closed form bound solutions with respect to both posterior and prior covariance. The second result is invalid under the PAC-Bayes framework but is useful as a sanity check. Our results imply both problems with optimization of VI as well as that significantly better priors can in theory be found. At the same time, the closed form results are far from optimal and point to intrinsic limitations of the mean-field approximation.

We then motivate modeling the curvature at the minimum through a simplified version of K-FAC (Martens & Grosse, 2015). This allows us to efficiently sample (complexity, empirical risk) pairs with improved curvature estimates. Using our Risk-Complexity plots, we find that for randomized classifiers with medium to low empirical risk this results in significant improvements in the generalization bound quality, compared to the implied limits of the mean-field approximation.

1.1. Related work

Criticism of uniform convergence. In Nagarajan & Kolter (2019), the authors posit that two sided uniform convergence bounds cannot produce non-vacuous estimates for deep neural networks, even with aggressive pruning of the hypothesis space. To the best of the authors understanding the criticism holds only for derandomized PAC-Bayes bounds. In the following we will be dealing only with bounding the generalization error of stochastic classifiers. Even for the deterministic case the issue is far from resolved (Negrea et al., 2019).

Bounds leveraging the Hessian. A number of bounds incorporating the Hessian have been proposed. Some works provide complexity measures that by design simply correlate with generalization error (Keskar et al., 2016; Li et al., 2019; Rangamani et al., 2019; Liang et al., 2017; Jia & Su, 2019). Others approximate the loss around the minimum using a second order Taylor expansion (Tsuzuku et al., 2019; Wang et al., 2018) and and then optimize the bound with respect to this approximation. In Tsuzuku et al. (2019) the authors first set the prior variance equal to the posterior variance, and then optimize the bound. This results in a suboptimal choice of prior. In Wang et al. (2018) the authors restrict the Hessian to be diagonal and optimize with respect only to the posterior covariance. Both Tsuzuku et al. (2019) and Wang et al. (2018) result in vacuous bounds.

Other bounds and relationship to Bayesian inference. There has been a huge number of works on generalization bounds for deep neural networks(Bartlett et al., 2017; Golowich et al., 2017; Wei & Ma, 2019; Ledent et al., 2019; Pitas et al., 2019). These are typically vacuous by several orders of magnitude. A number of works have pointed out the relationship between PAC-Bayes and Bayesian inference (Germain et al., 2016; Achille & Soatto, 2018; Achille et al., 2019; Dziugaite & Roy, 2017).

In Huang et al. (2019) the authors propose "Kronecker flow" to obtain better PAC-Bayes bounds. While we also test a more flexible posterior, our emphasis is on a detailed criticism of the mean-field approximation. Furthermore, as we discuss in section 5, flow based methods face a number of challenges in our testing setup.

2. Preliminaries

A neural network transforms its inputs $\mathbf{a}_0 = \boldsymbol{x}$ to an output $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathbf{a}_l$ through a series of l layers, each of which consists of a bank of units/neurons. The computation performed by each layer $i \in \{1, ..., l\}$ is given as follows

$$\begin{aligned} \boldsymbol{s}_i &= \mathbf{W}_i \mathbf{a}_{i-1}, \\ \mathbf{a}_i &= \phi_i(\boldsymbol{s}_i), \end{aligned}$$

where ϕ_i is an element-wise non-linear function and $\mathbf{W}_i$ is a weight matrix.

We will define $\boldsymbol{\theta} = [\text{vec}(\mathbf{W}_0)\text{vec}(\mathbf{W}_0)\cdots\text{vec}(\mathbf{W}_l)]$, which is the vector consisting of all the network's parameters concatenated together, where vec is the operator which vectorizes matrices by concatenating their rows horizontally.

We denote the learning sample $(X, Y) = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, that contains n input-output pairs. Samples (X, Y) are assumed to be sampled randomly from a distribution $\mathcal{D}$. Thus, we denote $(X, Y) \sim \mathcal{D}^n$ the i.i.d observation of n elements. We consider loss functions $\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{F}$ is a set of predictors $f : \mathcal{X} \to \mathcal{Y}$. We also denote the empirical risk $\hat{\mathcal{L}}^{\ell}_{X,Y}(f) = (1/n)\sum_i \ell(f, \boldsymbol{x}_i, y_i)$ and the risk $\mathcal{L}^{\ell}_{\mathcal{D}}(f) = \mathbf{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\ell(f, \boldsymbol{x}, y)$.

We will use two loss functions, the non-differentiable zero-one loss $\ell_{01}(f, x, y) = \mathbb{I}(\arg\max(f(x)) = y)$, and categorical cross-entropy, which is a commonly used differentiable surrogate $\ell_{\text{cat}}(f, x, y) = -\sum_i \mathbb{I}[i = y] \log(f(x)_i)$, where we assume that the outputs of f are normalized to form a probability distribution.

We will also use the following PAC-Bayes formulation, by Catoni (2007)

Theorem 2.1. *(Catoni, 2007) Given a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, a hypothesis set $\mathcal{F}$, a loss function $\ell' : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$, a prior distribution π over $\mathcal{F}$, a real number $\delta \in (0, 1]$, and a real number $\beta > 0$, with probability at least $1 - \delta$ over the choice of $(X, Y) \sim \mathcal{D}^n$, we have*

$$\begin{aligned} \forall \hat{\rho} \text{ on } \mathcal{F} : \mathbf{E}_{f\sim\hat{\rho}}\mathcal{L}^{\ell'}_{\mathcal{D}}(f) \leq& \Phi^{-1}_{\beta}(\mathbf{E}_{f\sim\hat{\rho}}\hat{\mathcal{L}}^{\ell'}_{X,Y}(f) \\ &+ \frac{1}{\beta n}(\text{KL}(\hat{\rho}||\pi) + \ln\frac{1}{\delta})), \end{aligned} \tag{2}$$

where $\Phi^{-1}_{\beta}(x) = \frac{1-e^{-\beta x}}{1-e^{-\beta}}$.

The above PAC-Bayes theorem works with bounded loss functions and as such is typically evaluated with the zero-one loss ℓ_{01}. However, one might want to optimize the above bound as proposed in Dziugaite & Roy (2017). One approach, is to then parametrize $f_{\boldsymbol{\theta}}$ using diagonal Gaussians as $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \boldsymbol{\sigma}_{\hat{\rho}})$ and the prior as $\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\pi}, \lambda\mathbf{I})$. Then, one can use the reparametrization trick $\boldsymbol{\theta} = \boldsymbol{\mu}_{\hat{\rho}} + \sqrt{\boldsymbol{\sigma}_{\hat{\rho}}} \odot \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the categorical cross-entropy to optimize the surrogate

$$\mathbf{E}_{\boldsymbol{\theta}\sim\hat{\rho}(\boldsymbol{\theta})}\hat{\mathcal{L}}^{\ell_{\text{cat}}}_{X,Y}(f_{\boldsymbol{\theta}}) + \frac{1}{\beta n}(\text{KL}(\hat{\rho}(\boldsymbol{\theta})||\mathcal{N}(\boldsymbol{\mu}_{\pi}, \lambda\mathbf{I})) + \ln\frac{1}{\delta}), \tag{3}$$

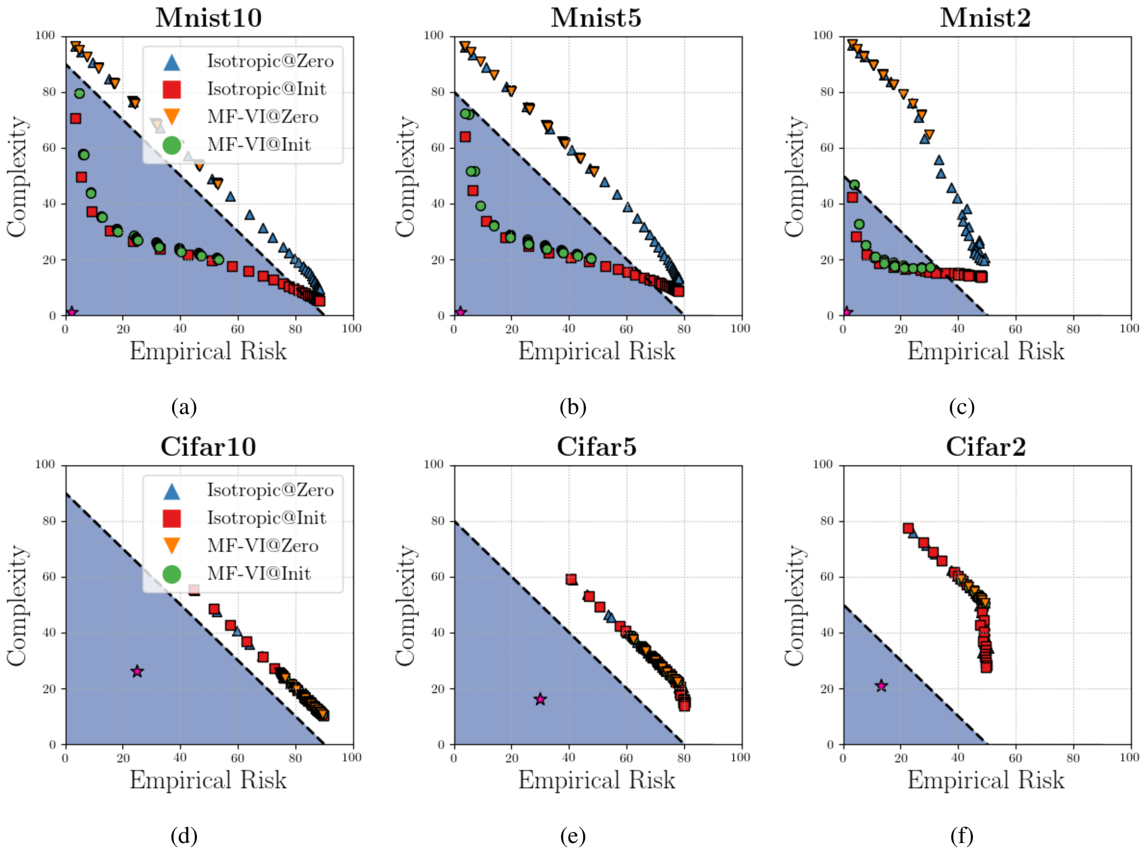


Figure 2. **Detailed comparison of posterior and prior choices**: The area below the dashed line corresponds to non-vacuous pairs of (complexity, empirical risk). The purple star corresponds to the optimal bound implied by the testing set. For the MNIST case there is a significant improvement when changing from a prior centered at 0 to a prior centered at the random initialization. The baseline isotropic bounds are non-vacuous for a prior centered at 0. Optimizing the mean-field approximation using variational inference provides no improvements over the baseline, regardless of prior choice. In the CIFAR case all modelling choices result in vacuous bounds.

for $\boldsymbol{\mu}_{\hat\rho}$, $\boldsymbol{\sigma}_{\hat\rho}$. In practice, one optimizes (3), but wants to evaluate (2). It's also often beneficial to fine tune λ and we want to approximate $\mathbf{E}_{f\sim\hat\rho}\hat{\mathcal{L}}^{\ell_{01}}_{X,Y}(f)$ with an empirical estimate. We take a union bound over values of λ, and apply a Chernoff bound for the tail of the empirical estimate of $\mathbf{E}_{f\sim\hat\rho}\hat{\mathcal{L}}^{\ell_{01}}_{X,Y}(f)$. Putting everything together, as proposed in Dziugaite & Roy (2017), one can obtain valid PAC-Bayes bounds subject to a posterior distribution $\hat\rho^*(\boldsymbol{\theta})$ that hold with probability at least $1-\delta-\delta'$ and are of the form

$$\mathbf{E}_{\boldsymbol{\theta}\sim\hat\rho^*(\boldsymbol{\theta})}\mathcal{L}^{\ell_{01}}_{\mathcal{D}}(f_{\boldsymbol{\theta}}) \le \Phi^{-1}_{\beta}(\tilde{\mathcal{L}}^{\ell_{01}}_{X,Y}(f_{\boldsymbol{\theta}}) + \frac{1}{\beta n}\mathrm{KL}(\hat\rho^*(\boldsymbol{\theta})||\pi) + \frac{1}{\beta n}\ln(\frac{\pi^2 b^2 \ln(c/\lambda)^2}{6\delta}) + \sqrt{\frac{\ln\frac{2}{\delta'}}{m}}), \tag{4}$$

where $\Phi^{-1}_{\beta}(x) = \frac{1-e^{-\beta x}}{1-e^{-\beta}}$. Also c, b are constants, m is the number of samples from $\hat\rho$ for approximating $\mathbf{E}_{f\sim\hat\rho}\hat{\mathcal{L}}^{\ell_{01}}_{X,Y}(f)$ and $\tilde{\mathcal{L}}^{\ell_{01}}_{X,Y}(f_{\boldsymbol{\theta}})$ the empirical estimate.

It is not difficult to see, that for a high enough number of samples n and m, the terms in line 2 of (4) have a negligible effect on the bound. All proofs are deferred to the Appendix.

3. Empirical results

We tested 6 different datasets. These consist of the original MNIST-10 and CIFAR-10 (Krizhevsky & Hinton, 2010) datasets, as well as simplified versions, where we collapsed the 10 classes into 5 and 2 aggregate classes, potentially simplifying the classification problem. All had 50000 training samples. We test the architectures

$$\text{input} \rightarrow 300\text{FC} \rightarrow 300\text{FC} \rightarrow \#classes\text{FC} \rightarrow \text{output}$$

on MNIST, and

$$\text{input} \rightarrow 200\text{FC} \rightarrow 200\text{FC} \rightarrow \#classes\text{FC} \rightarrow \text{output}$$

on CIFAR, where xFC denotes a fully connected layer with x neurons.

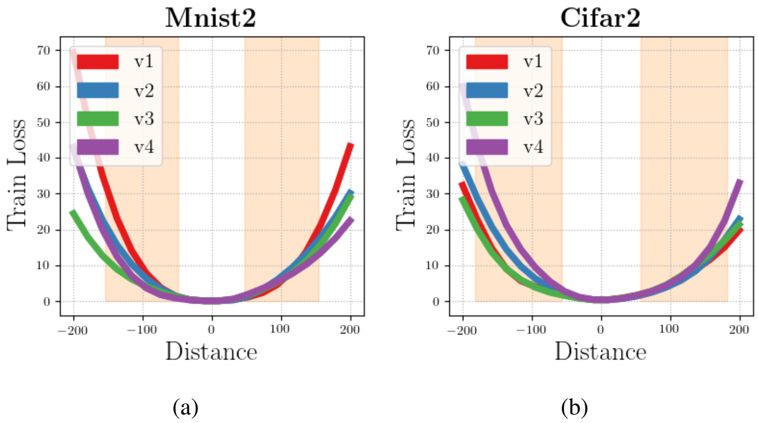


Figure 3. **Empirical evaluation of the categorical cross-entropy loss**: We take normalized random directions $\mathbf{v}_i, \ i \in \{1,2,3,4\}$ and plot the deterministic categorical cross-entropy loss $\hat{\mathcal{L}}^{\ell_{\text{cat}}}_{X,Y}(f_{\boldsymbol{\theta}})$ for MNIST2 and CIFAR2 and values on the line $\boldsymbol{\theta} = \boldsymbol{\theta}_* + t\mathbf{v}_i, \ t \in [-200, 200]$. We see that the loss closely reassembles a quadratic around the minimum $\boldsymbol{\theta}_*$. High dimensional Gaussian vectors concentrate close to a hypersphere centered on the mean. We find the radius of the hyperspheres and shade the corresponding 1 dimensional cross sections in the plots. Posteriors relevant to our experiments concentrate within an area well approximated by the quadratic.

We also tested four combinations of prior and posterior

1. $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \lambda\mathbf{I})$, $\pi(\boldsymbol{\theta}) = \mathcal{N}(0, \lambda\mathbf{I})$
2. $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \lambda\mathbf{I})$, $\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\text{init}}, \lambda\mathbf{I})$
3. $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \boldsymbol{\sigma}_{\hat{\rho}})$, $\pi(\boldsymbol{\theta}) = \mathcal{N}(0, \lambda\mathbf{I})$.
4. $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \boldsymbol{\sigma}_{\hat{\rho}})$, $\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\text{init}}, \lambda\mathbf{I})$.

Isotropic posterior. Isotropic combinations 1 and 2 differ only in the prior mean. The first prior is centered at 0, while the second prior is centered at the random deep neural network initialization. In practice, to derive multiple (complexity, empirical risk) pairs we sample λ,β in the range $\lambda \in [0.031, 0.3]$ and $\beta \in [1, 5]$. For these we compute $\hat{\mathcal{L}}(\hat{\rho})$ and $\text{KL}(\hat{\rho}||\pi)$. The second can be computed analytically, while we approximate the first using Monte Carlo sampling with $m = 1000$ samples from $\hat{\rho}$. We then plug the results into (4). We set the estimated complexity as $\text{Complexity} \equiv [\Phi^{-1}_{\beta^*}(\hat{\mathcal{L}}(\hat{\rho}^*) + \frac{1}{\beta^* n}\text{KL}(\hat{\rho}^*||\pi)) - \hat{\mathcal{L}}(\hat{\rho}^*)]$, where β^* is the optimal β.

Diagonal posterior (VI). Combinations 3 and 4 correspond to a posterior with diagonal covariance and a non-informative prior centered at 0 and at the random initialization. For MNIST we do a grid search over $\beta \in [1, 5]$ and $\lambda \in [0.03, 0.1]$ while for CIFAR we search in $\beta \in [1, 5]$ and $\lambda \in [0.1, 0.3]$. For each (β, λ) pair we optimize $\boldsymbol{\sigma}_{\hat{\rho}}$ using the surrogate (3). Specifically, we use the state of the art Flipout estimator (Wen et al., 2018). We used 5 epochs of training using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of $1e - 1$. Increasing the number of epochs didn't affect the results. We calculate the complexity and empirical risk as in the isotropic case.

We plot the Pareto fronts of all modeling choices in 2. For the case of MNIST, changing from the prior centered at 0 to the prior centered at the random initialization resulted in a significant improvement of the bound. The resulting bounds with a prior at the random initialization are non-vacuous, even for the simple isotropic posterior. Optimizing the covariance with VI yields negligible or no improvements, regardless of the prior choice.

For CIFAR, we do not see significant variation in the bounds. The Catoni bound has a saturating effect above the line $y = 1 - x, \ \text{s.t.} \ x \in [0, 1]$. All (complexity,empirical risk) pairs fall into this saturating region. Specifically, mean-field VI fails to meaningfully improve the bound. Looking at the optimal bound points (star shapes), one explanation for the difference with MNIST, is that CIFAR DNNs have overfit the data significantly.

4. Quadratic Approximation

The stochastic and non-convex objective (3) is difficult to analyze theoretically. As such we first propose to approximate the cross-entropy loss at the mean of the posterior using a second order Taylor expansion which will make the subsequent analysis tractable (this corresponds to a *Laplace* approximation (Bishop, 2006) to the posterior). We introduce the centered random variable $\boldsymbol{\eta} = \boldsymbol{\theta} - \mathbf{E}[\boldsymbol{\theta}]$ so that $\boldsymbol{\eta} \sim \hat{\rho}'(\boldsymbol{\theta})$, we get

$$\begin{aligned} &C_\beta(X, Y; \hat{\rho}, \pi) = \mathbf{E}_{\boldsymbol{\theta}\sim\hat{\rho}(\boldsymbol{\theta})}\hat{\mathcal{L}}^{\ell_{\text{cat}}}_{X,Y}(f_{\boldsymbol{\theta}}) + \beta\text{KL}(\hat{\rho}(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \\ &\approx \mathbf{E}_{\boldsymbol{\eta}\sim\hat{\rho}'(\boldsymbol{\theta})}[\boldsymbol{\eta}^T\nabla\hat{\mathcal{L}}^{\ell_{\text{cat}}}_{X,Y}(f_{\boldsymbol{\theta}}) + \frac{1}{2}\boldsymbol{\eta}^T\nabla^2\hat{\mathcal{L}}^{\ell_{\text{cat}}}_{X,Y}(f_{\boldsymbol{\theta}})\boldsymbol{\eta}] \\ &+ \beta\text{KL}(\hat{\rho}(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) \end{aligned} \quad (5)$$

$$\approx \mathbf{E}_{\boldsymbol{\eta}\sim\hat{\rho}'(\boldsymbol{\theta})}[\frac{1}{2}\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta}] + \beta\mathrm{KL}(\hat{\rho}(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})).$$

where $\mathbf{H} \equiv \nabla^2\hat{\mathcal{L}}^{\ell_{\text{cat}}}_{X,Y}(f_{\boldsymbol{\theta}})$ is the Hessian and captures the curvature at the minimum.

In the above we made a number of assumptions. First, we assumed that the gradient at the point of expansion is zero. For a well trained overparametrized DNN this is a reasonable assumption. Secondly, we omit terms of the Taylor expansion of order ≥ 3. This results in a quadratic approximation. We conduct experiments to see whether this is reasonable. Specifically, we take random directions along the loss landscape and plot along them the value of the loss. We see in Figure 3 that the loss is indeed approximately quadratic around the minimum. At the same time, we note that approximating the loss as quadratic has been used to obtain state of the art results in the DNN compression literature (Dong et al., 2017; Wang et al., 2019; Peng et al., 2019; LeCun et al., 1990; Hassibi & Stork, 1993).

For the expectation of the quadratic loss to be a good approximation of the expectation of the categorical loss, the mass of the posterior has to be concentrated at locations where the true loss is well approximated by a quadratic. We have thus far dealt with Gaussian posteriors $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \boldsymbol{\sigma}_{\hat{\rho}})$, where $\forall i,\ \sigma_{\hat{\rho}i} \approx \lambda$, $0.01 \leq \lambda \leq 1$. It is well know that Gaussians in high dimensions concentrate on a thin "bubble" away from the origin. We can use this intuition and make a rough calculation of the radius of this bubble (Vershynin, 2018). Specifically, assuming that $\forall i,\ \boldsymbol{\sigma}_{\hat{\rho}i} = \lambda$, we can calculate $\mathbf{E}_{\boldsymbol{\eta}\sim\hat{\rho}'(\boldsymbol{\theta})}||\boldsymbol{\eta}||_2^2 = \mathbf{E}_{\boldsymbol{\eta}\sim\mathcal{N}(0,\boldsymbol{\sigma}_{\hat{\rho}})}[\sum_{i=0}^d \eta_i^2] = \sum_{i=0}^d \sigma_{\hat{\rho}i} = \lambda d$. Finally we expect that the radius of the "bubble" is $\mathbf{E}_{\boldsymbol{\eta}\sim\hat{\rho}'(\boldsymbol{\theta})}||\boldsymbol{\eta}||_2 \approx \sqrt{\lambda d}$. We plot these regions in Figure 3. We see that posteriors concentrate within areas where the quadratic approximation is reasonable.

4.1. Optimal Posterior

Compared to the diagonal modeling of the previous section, we now make the slightly more general modeling choices $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \boldsymbol{\Sigma}_{\hat{\rho}})$ and $\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\pi}, \lambda\boldsymbol{\Sigma}_{\pi})$. We can then show that the optimal posterior covariance of the objective (5) for fixed prior and posterior means has a closed form solution.

Lemma 4.1. *The convex optimization problem* $\min_{\boldsymbol{\Sigma}_{\hat{\rho}}} \mathbf{E}_{\boldsymbol{\eta}\sim\hat{\rho}'(\boldsymbol{\theta})}[\frac{1}{2}\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta}] + \beta\mathrm{KL}(\hat{\rho}(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))$ *where* $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \boldsymbol{\Sigma}_{\hat{\rho}})$ *and* $\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\pi}, \lambda\boldsymbol{\Sigma}_{\pi})$ *is minimized at*

$$\boldsymbol{\Sigma}^*_{\hat{\rho}} = \beta(\mathbf{H} + \frac{\beta}{\lambda}\boldsymbol{\Sigma}_{\pi}^{-1})^{-1}, \tag{6}$$

where $\mathbf{H} \equiv \nabla^2\hat{\mathcal{L}}^{\ell_{cat}}_{X,Y}(f_{\boldsymbol{\theta}})$ *captures the curvature at the minimum, while* $\boldsymbol{\Sigma}_{\pi}$ *is the prior covariance.*

4.2. Optimal Prior

We can relax the modeling choices further by noting that PAC-Bayesian theory allows one to choose an informative prior, with the restriction that the prior can only depend on the data generating distribution and *not* the training set. A number of previous works (Parrado-Hernández et al., 2012; Catoni, 2003; Ambroladze et al., 2007) have used this insight mainly on simpler linear settings and usually by training a classifier on a separate training set and using the result as a prior. Recently, Dziugaite & Roy (2018) have proposed to use the original training set to derive valid priors by imposing differential privacy constraints.

We ignore these concerns for the moment, and optimize the prior covariance directly. The objective is non-convex, however for the case of *diagonal* prior and posterior covariances we can find the global minimum.

Lemma 4.2. *The optimal prior and posterior covariances for* $\min_{\boldsymbol{\sigma}_{\hat{\rho}},\boldsymbol{\sigma}_{\pi}} \mathbf{E}_{\boldsymbol{\eta}\sim\hat{\rho}'(\boldsymbol{\theta})}[\frac{1}{2}\boldsymbol{\eta}^T\mathbf{H}\boldsymbol{\eta}] + \beta\mathrm{KL}(\hat{\rho}(\boldsymbol{\theta})||\pi(\boldsymbol{\theta}))$ *with* $\hat{\rho}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}}, \boldsymbol{\sigma}_{\hat{\rho}})$ *and* $\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\pi}, \lambda\boldsymbol{\sigma}_{\pi})$ *have elements*

$$(\sigma^*_{\hat{\rho}i})^{-1} = \frac{1}{2\beta}[h_i + \sqrt{h_i^2 + \frac{4\beta h_i}{(\mu_{i\hat{\rho}} - \mu_{i\pi})^2}}], \tag{7}$$

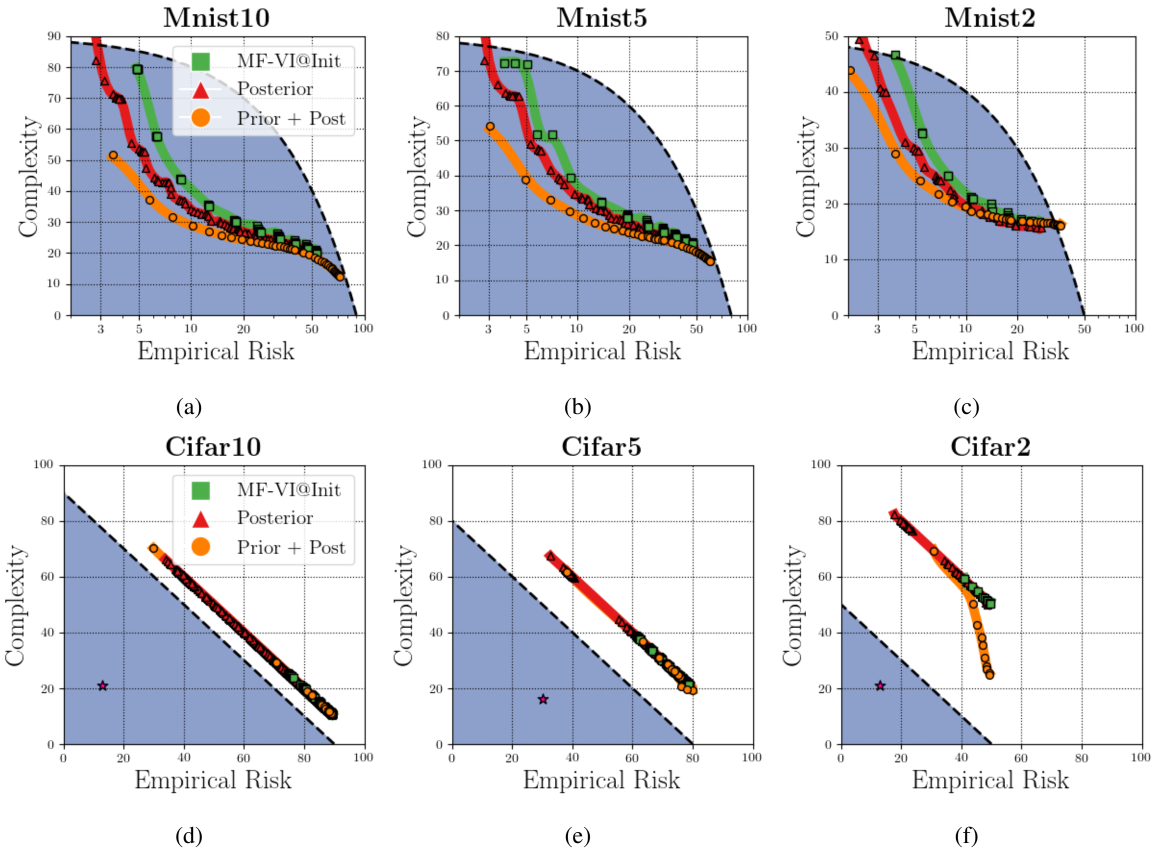
$$(\sigma^*_{\pi i})^{-1} = \frac{\lambda}{2\beta}[\sqrt{h_i^2 + \frac{4\beta h_i}{(\mu_{i\hat{\rho}} - \mu_{i\pi})^2}} - h_i], \tag{8}$$

where $\mathbf{H} \equiv \nabla^2\hat{\mathcal{L}}^{\ell_{cat}}_{X,Y}(f_{\boldsymbol{\theta}})$ *captures the curvature at the minimum.*

We *cannot* prove generalization using this result. Rather we use it as a sanity check for what is achievable through the mean-field approximation and an optimal informative prior covariance.

To approximate the Hessian we note that for the cross entropy loss and the softmax activation function $p(y = c|f_{\boldsymbol{\theta}}) = \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x})_c)/\sum_i \exp(f_{\boldsymbol{\theta}}(\boldsymbol{x})_i)$ the Fisher Information matrix coincides with the generalized Gauss-Newton approximation of the Hessian (Kunstner et al., 2019). We sample one ouput $\tilde{y}_i$ from the model distribution $p(y_i|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$ for each input $\boldsymbol{x}_i$, and approximate $\mathbf{H} \approx \sum_{i=0}^n \nabla_{\boldsymbol{\theta}} \log p(\tilde{y}_i|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))\nabla_{\boldsymbol{\theta}} \log p(\tilde{y}_i|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))^{\mathrm{T}}$, retaining only the diagonal elements.

Keeping the posterior and prior means fixed, we optimize the posterior covariance, as well as the posterior and *prior* covariance jointly in closed form. We plot the results in Figure 4, using the same approach as section 3 with $m = 1000$ for (7),(8), $m = 100$ for (6) and sampling over β and λ. For MNIST, valid bounds where we only optimize the posterior in closed form get significant benefits over VI of between 5-10%. Thus, even though Adam is very robust



(a) (b) (c)

(d) (e) (f)

Figure 4. **Closed form posterior and prior**: We plot the results obtained by mean-field variational inference, as well as the closed form bounds with optimized posterior and jointly optimized posterior and prior covariances. For MNIST, we plot the empirical risk in logarithmic scale for ease of exposition. Valid bounds where we only optimize the posterior in closed form get significant benefits over VI of between 5-10%. Optimizing the prior results in further improvements of 5-10%, implying that in theory better priors can be found. The results are far from tight even when optimizing the prior and for CIFAR all bounds are vacuous. This implies inherent limitations of the mean-field approximation, as we typically don't even have access to the optimal prior covariance.

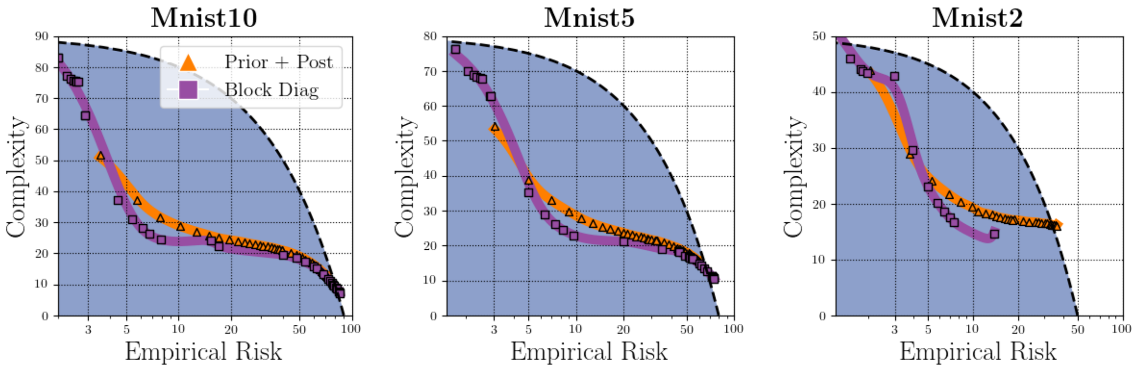
to hyperparameter selection, and the Flipout estimator is state of the art, one might look to hyperparameter tuning for better results. We present arguments in the next section, that hold also for the mean-field case, as to why it should be beneficial to avoid hyperparameter tuning. Invalid bounds where we optimize the prior and posterior jointly result in further improvements of 5-10%, implying that in theory better priors can be found. The bounds are far from tight, even when optimizing the prior, and for CIFAR all bounds are vacuous. This implies that the mean-field approximation is limited in the bound improvements it can provide.

5. Beyond the mean-field approximation

5.1. Computational Issues

Given the implied shortcomings of the mean-field approximation it is interesting to look at richer posterior distributions. A number of approximations exist to model such posteriors. In Mishkin et al. (2018), the authors model the covariance as having a low-rank + diagonal structure. In normalizing flows (Rezende & Mohamed, 2015) a simple initial density is transformed into a more complex one, by applying a sequence of invertible transformations, until a desired level of complexity is attained. In K-FAC (Martens & Grosse, 2015), the Hessian can be approximated as a Khatri-Rao product to construct a Laplace approximation of the posterior (Ritter et al., 2018). Considerable effort has been placed into

Optimizing multiple variational objectives. To obtain Pareto fronts we will perform a grid search over λ and β, corresponding to $\mathcal{O}(10^2)$ classifiers with different empirical risk and complexity. Optimizing variational objectives is known to be unstable, to scale badly and to require extensive hyperparameter tuning (Wu et al., 2018). Optimizing each posterior using SGD as in Mishkin et al. (2018); Rezende & Mohamed (2015), for even a few minutes, can add several hours to obtaining the full grid. Hyperparameter tuning objectives that do not converge can quickly make the task



(a) (b) (c)

Figure 5. **Beyond the mean field approximation**: We compared the simplified K-FAC curvature approximation to the closed form *invalid* mean-field inference. Invalid results correspond to an optimal prior and posterior covariance to which we don't typically have access. For medium to low empirical risk the block diagonal curvature improves the bound for MNIST10-5-2 by $8.2\%, 7.5\%, 4.4\%$ respectively.

infeasible.

Sampling efficiently from the posterior. At the same time we will need to sample efficiently between $\mathcal{O}(10^4)$ and $\mathcal{O}(10^5)$ posterior samples. This is because we will be applying a Chernoff bound on the tail of the empirical risk. For flow based methods, the KL term also has to be approximated with MC sampling.

In the non-flow based methods, one typically seeks to factor $\Sigma = LL^T$. Then $y = Lz$, where z is standard normal, has the appropriate distribution, and can be sampled efficiently. While Mishkin et al. (2018) provide an efficient Cholesky factorization of their low-rank + diagonal approximation, the Khatri-Rao product (Martens & Grosse, 2015) has no obvious Cholesky factorization. Finally inference time in flow based methods will be influenced by the number of mappings used in the flow.

Simplified K-FAC Laplace. We assume a multiclass classification problem with c classes, and that the labels y are one-hot encoded. We then define the mean square error loss $\ell_{\text{mse}}(f, x, y) = (1/c)\sum_{i=0}^{c}(f(x)_i - y_i)^2$. Assuming r neurons per layer, $\boldsymbol{\theta}$ has a form $\boldsymbol{\theta} = [\text{vec}(\mathbf{W}_0^{0,:})\text{vec}(\mathbf{W}_0^{1,:})\cdots\text{vec}(\mathbf{W}_l^{r,:})]$. We also denote for layer i and neuron j, $\boldsymbol{\theta}_{ij}$, $\boldsymbol{\mu}_{\hat{\rho}ij}$, $\boldsymbol{\Sigma}_{\hat{\rho}ij}$, $\boldsymbol{\mu}_{\pi ij}$ the corresponding split variables. We can then motivate optimizing the following surrogate upper bound

Lemma 5.1. *Assuming negligible layerwise derivatives of order other than 2, the differentiable surrogate objective*

$$\mathbf{E}_{\boldsymbol{\theta}\sim\hat{\rho}(\boldsymbol{\theta})}\hat{\mathcal{L}}^{\ell_{mse}}_{X,Y}(f_{\boldsymbol{\theta}}) + \frac{1}{\beta n}(\text{KL}(\hat{\rho}(\boldsymbol{\theta})||\mathcal{N}(\boldsymbol{\mu}_{\pi}, \lambda\mathbf{I})) + \ln\frac{1}{\delta}), \tag{9}$$

has the following upper bound

$$\sum_{i,j}[\mathbf{E}_{\boldsymbol{\eta}_{ij}\sim\hat{\rho}'_{ij}(\boldsymbol{\theta})}[\frac{1}{2}\boldsymbol{\eta}_{ij}^T\mathbf{H}_i\boldsymbol{\eta}_{ij}] + \frac{1}{\beta n}\text{KL}(\hat{\rho}_{ij}(\boldsymbol{\theta})||\pi_{ij}(\boldsymbol{\theta})] + \mathcal{O}(c^l), \tag{10}$$

where $\hat{\rho}_{ij}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\hat{\rho}ij}, \boldsymbol{\Sigma}_{\hat{\rho}ij})$, $\pi_{ij}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{\pi ij}, \lambda\mathbf{I})$, $\mathbf{H}_i = (1/n)\sum_{k=0}^{n}\mathbf{a}_i^k\mathbf{a}_i^{k^T}$*, are neuronwise posteriors, priors and Hessians.*

The above corresponds to a greatly simplified version of K-FAC, where each layer has a posterior with covariance $\boldsymbol{\Sigma}_i = \mathbf{H}_i \otimes \mathbf{I}$, i.e. we assume correlations only for parameters in each neuron. While this approximation covers our needs, one could in principle use the slightly more expressive $\boldsymbol{\Sigma}_i = \mathbf{H}_i \otimes \mathbf{G}_i$, where $\mathbf{G}_i = \mathbf{E}[\mathbf{g}_i\mathbf{g}_i^T]$ and $\mathbf{g}_i$ are the backpropagated layerwise errors for layer i (Ritter et al., 2018), and we note that optimizing more expressive versions of K-FAC efficiently is an active area of research (Zhang et al., 2018; Bae et al., 2018). We've broken the original into many much smaller subproblems. We can now compute the Hessian efficiently and in a stable way once, and then sample the posterior efficiently in closed form at different variance levels λ.

5.2. Empirical Results

We now present results on the MNIST datasets. We run a grid search over β and λ, with 20 samples each, for $\beta \in [0.001, 0.02]$ and $\lambda \in [0.001, 0.1]$. We use $m = 1000$ samples for estimating the empirical risk. For computing the Pareto fronts we optimize (10) with (6) and evaluate (4) following the procedure of Section 3. The running time for each experiment was 33h, 30h and 25h respectively.

We plot the results in Figure 5 where we compare with the case of jointly optimized diagonal prior and posterior. At very low and very high empirical risk levels the complexity estimates saturate. However, for medium empirical risk levels the block diagonal covariance yields significant improvements to the bounds. The effect is more pronounced on the more difficult MNIST10 and MNIST5 experiments, where using the block diagonal posterior results in a decrease in the estimated complexity of $\sim 10\%$.

6. Conclusion and Future Work

We have presented several arguments in favor of richer posterior distributions under the PAC-Bayes framework. We've only scratched the surface, as we've relaxed only slightly from the diagonal case, getting significant gains. Of course, another line of approach would be to optimize further the prior *mean* in a valid way, an area that has been little investigated. As research moves closer to solving the generalization puzzle of deep learning, we hope that our plots provide a more intuitive way to compare new bounds.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Achille, A., Paolini, G., Mbeng, G., and Soatto, S. The information complexity of learning tasks, their structure and their distance. *arXiv preprint arXiv:1904.03292*, 2019.

Ambroladze, A., Parrado-Hernández, E., and Shawe-taylor, J. S. Tighter pac-bayes bounds. In *Advances in neural information processing systems*, pp. 9–16, 2007.

Bae, J., Zhang, G., and Grosse, R. Eigenvalue corrected noisy natural gradient. *arXiv preprint arXiv:1811.12565*, 2018.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Bishop, C. M. *Pattern recognition and machine learning*. springer, 2006.

Blier, L. and Ollivier, Y. The description length of deep learning models. In *Advances in Neural Information Processing Systems*, pp. 2216–2226, 2018.

Catoni, O. A pac-bayesian approach to adaptive classification. *preprint*, 840, 2003.

Catoni, O. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1019–1028. JMLR. org, 2017.

Dong, X., Chen, S., and Pan, S. Learning to prune deep neural networks via layer-wise optimal brain surgeon. In *Advances in Neural Information Processing Systems*, pp. 4857–4867, 2017.

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

Dziugaite, G. K. and Roy, D. M. Data-dependent pac-bayes priors via differential privacy. In *Advances in Neural Information Processing Systems*, pp. 8430–8441, 2018.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. Pac-bayesian theory meets bayesian inference. In *Advances in Neural Information Processing Systems*, pp. 1884–1892, 2016.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

Hassibi, B. and Stork, D. G. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in neural information processing systems*, pp. 164–171, 1993.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Huang, C.-W., Touati, A., Vincent, P., Dziugaite, G. K., Lacoste, A., and Courville, A. Stochastic neural network with kronecker flow. *arXiv preprint arXiv:1906.04282*, 2019.

Jia, Z. and Su, H. Information-theoretic local minima characterization and regularization. *arXiv preprint arXiv:1911.08192*, 2019.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pp. 2575–2583, 2015.

Krizhevsky, A. and Hinton, G. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7): 1–9, 2010.

Kunstner, F., Balles, L., and Hennig, P. Limitations of the empirical fisher approximation. *arXiv preprint arXiv:1905.12558*, 2019.

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

LeCun, Y., Denker, J. S., and Solla, S. A. Optimal brain damage. In *Advances in neural information processing systems*, pp. 598–605, 1990.

Ledent, A., Lei, Y., and Kloft, M. Improved generalisation bounds for deep learning through covering numbers. *arXiv preprint arXiv:1905.12430*, 2019.

Li, X., Gu, Q., Zhou, Y., Chen, T., and Banerjee, A. Hessian based analysis of sgd for deep nets: Dynamics and generalization. *arXiv preprint arXiv:1907.10732*, 2019.

Liang, T., Poggio, T., Rakhlin, A., and Stokes, J. Fisher-rao metric, geometry, and complexity of neural networks. *arXiv preprint arXiv:1711.01530*, 2017.

Martens, J. and Grosse, R. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417, 2015.

McAllester, D. A. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Mishkin, A., Kunstner, F., Nielsen, D., Schmidt, M., and Khan, M. E. Slang: Fast structured covariance approximations for bayesian deep learning with natural gradient. In *Advances in Neural Information Processing Systems*, pp. 6245–6255, 2018.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. *arXiv preprint arXiv:1902.04742*, 2019.

Negrea, J., Dziugaite, G. K., and Roy, D. M. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. *arXiv preprint arXiv:1912.04265*, 2019.

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. Pac-bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13(Dec):3507–3531, 2012.

Peng, H., Wu, J., Chen, S., and Huang, J. Collaborative channel pruning for deep networks. In *International Conference on Machine Learning*, pp. 5113–5122, 2019.

Pitas, K., Loukas, A., Davies, M., and Vandergheynst, P. Some limitations of norm based generalization bounds in deep neural networks. *arXiv preprint arXiv:1905.09677*, 2019.

Rangamani, A., Nguyen, N. H., Kumar, A., Phan, D., Chin, S. H., and Tran, T. D. A scale invariant flatness measure for deep network minima. *arXiv preprint arXiv:1902.02434*, 2019.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Ritter, H., Botev, A., and Barber, D. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning, 2018.

Suzuki, T. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. *arXiv preprint arXiv:1909.11274*, 2019.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. *arXiv preprint arXiv:1901.04653*, 2019.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wang, C., Grosse, R., Fidler, S., and Zhang, G. Eigendamage: Structured pruning in the kronecker-factored eigenbasis. *arXiv preprint arXiv:1905.05934*, 2019.

Wang, H., Keskar, N. S., Xiong, C., and Socher, R. Identifying generalization properties in neural networks. *arXiv preprint arXiv:1809.07402*, 2018.

Wei, C. and Ma, T. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. *arXiv preprint arXiv:1905.03684*, 2019.

Wen, Y., Vicol, P., Ba, J., Tran, D., and Grosse, R. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.

Wu, A., Nowozin, S., Meeds, E., Turner, R. E., Hernández-Lobato, J. M., and Gaunt, A. L. Deterministic variational inference for robust bayesian neural networks. *arXiv preprint arXiv:1810.03958*, 2018.

Zhang, G., Sun, S., Duvenaud, D., and Grosse, R. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, pp. 5852–5861, 2018.

Zhou, W., Veitch, V., Austern, M., Adams, R. P., and Orbanz, P. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. *arXiv preprint arXiv:1804.05862*, 2018.