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

Konstantinos Pitas

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 $\pi$ 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

$$E \mathcal{L}(\hat{\rho}) \leq E \hat{\mathcal{L}}(\hat{\rho}) + \beta \text{KL}(\hat{\rho}||\pi), \quad (1)$$

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 $\beta \text{KL}(\hat{\rho}||\pi)$ term between the prior and posterior acts as a measure of complexity for the classifier.

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
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 known 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 $\beta$ 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. On the $x$-axis we plot the Empirical Risk $\hat{L}(\hat{\rho})$, while on the $y$-axis we plot the estimated Complexity $\beta \text{KL}(\hat{\rho}||\pi)$ 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 iv) 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 $\mathbb{E}\hat{L}(\hat{\rho})$ 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
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; Germain et al., 2016; Achille & Soatto, 2018; Achille et al., 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).

In Huang et al. (2019) the authors propose “Kronecker flow” approximation.

\[ s_i = W_i a_{i-1}, \]
\[ a_i = \phi_i(s_i), \]

where \( \phi_i \) is an element-wise non-linear function and \( W_i \) is a weight matrix.

We will define \( \theta = [\text{vec}(W_0), \text{vec}(W_q) \cdots \text{vec}(W_l)] \), 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 \( (X, Y) = \{(x_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n \), that contains \( n \) input-output pairs. Samples \( (X, Y) \) are assumed to be sampled randomly from a distribution \( D \). Thus, we denote \( (X, Y) \sim D^n \) the i.i.d observation of \( n \) elements. We consider loss functions \( \ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \), where \( \mathcal{F} \) is a set of predictors \( f : \mathcal{X} \rightarrow \mathcal{Y} \). We also denote the empirical risk \( \mathcal{L}_{\mathcal{D}}(f) = \frac{1}{n} \sum_i \ell(f(x_i), y_i) \) and the risk \( \mathcal{L}_{\mathcal{D}}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell(f, x; y) \).

We will use two loss functions, the non-differentiable zero-one loss \( \ell_0(f, x, y) = \mathbb{I}(\text{arg max}(\ell(f(x))) = y) \), and categorical cross-entropy, which is a commonly used differentiable surrogate \( \ell_{\text{cat}}(f, x, y) = -\sum_y \mathbb{I}[i = y] \log(f(x)_i) \), 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 \( D \) over \( \mathcal{X} \times \mathcal{Y} \), a hypothesis set \( \mathcal{F} \), a loss function \( \ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1] \), a prior distribution \( \pi \) 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 \( (X, Y) \sim D^n \), we have

\[
\mathbb{E}_{f \sim \hat{\rho}} \mathcal{L}_\mathcal{D}(f) \leq \Phi_{\beta}^{-1} \left( \mathbb{E}_{f \sim \hat{\rho}} \ell_{\text{cat}}(f, x; y) + \frac{1}{\beta n} (\text{KL}(|\hat{\rho}| \pi) + \ln \frac{1}{\delta}) \right),
\]

(2)

where \( \Phi_{\beta}^{-1}(x) = \frac{1 - e^{-\beta x}}{1 - e^{-\beta}}. \)

The above PAC-Bayes theorem works with bounded loss functions and as such is typically evaluated with the zero-one loss \( \ell_0 \). However, one might want to optimize the above bound as proposed in Dziugaite & Roy (2017). One approach, is to then parametrize \( f_\theta \) using diagonal Gaussians as \( \hat{\rho}(\theta) = \mathcal{N}(\mu_\theta, \sigma^2_\theta) \) and the prior as \( \pi(\theta) = \mathcal{N}(\mu_\pi, \lambda I) \). Then, one can use the reparametrization trick \( \theta = \mu_\theta + \sqrt{\sigma^2_\theta} \circ \mathcal{N}(0, I) \) and the categorical cross-entropy to optimize the surrogate

\[
\mathbb{E}_{\theta \sim \hat{\rho}(\theta)} \mathcal{L}_{\mathcal{D}}(f_\theta) + \frac{1}{\beta n} (\text{KL}(|\hat{\rho}(\theta)| \pi(\theta)) + \ln \frac{1}{\delta}),
\]

(3)

2. Preliminaries

A neural network transforms its inputs \( a_0 = x \) to an output \( f_\theta(x) = a_l \) through a series of \( l \) layers, each of which consists of a bank of units/neurons. The computation performed by each layer \( i \in \{1, \ldots, l\} \) is given as follows
for $\mu, \sigma$. In practice, one optimizes (3), but wants to evaluate (2). It’s also often beneficial to fine tune $\lambda$ and we want to approximate $E_{f \sim \hat{\rho}} L_{X,Y}^\ell(f)$ with an empirical estimate. We take a union bound over values of $\lambda$, and apply a Chernoff bound for the tail of the empirical estimate of $E_{f \sim \hat{\rho}} L_{X,Y}^\ell(f)$. Putting everything together, as proposed in Dziugaite & Roy (2017), one can obtain valid PAC-Bayes bounds subject to a posterior distribution $\hat{\rho}^*(\theta)$ that hold with probability at least $1 - \delta - \delta'$ and are of the form

$$E_{\theta \sim \hat{\rho}^*(\theta)} L_{\mathcal{D}}^\ell(f_\theta) \leq \Phi_{\beta}^{-1}(\hat{\rho}^*(\theta) || \pi) + \frac{1}{\beta n} \ln(\frac{\pi^2 b^2 \ln(e/\lambda)^2}{6\delta}) + \sqrt{\frac{\ln \frac{2}{\delta'}}{m}},$$

where $\Phi_{\beta}^{-1}(x) = \frac{1}{e} - e^{-\beta x}$. Also $c, b$ are constants, $m$ is the number of samples from $\hat{\rho}$ for approximating $E_{f \sim \hat{\rho}} L_{X,Y}^\ell(f)$ and $L_{X,Y}^\ell(f_\theta)$ 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

input $\rightarrow$ 300FC $\rightarrow$ 300FC $\rightarrow$ #classesFC $\rightarrow$ output

on MNIST, and

input $\rightarrow$ 200FC $\rightarrow$ 200FC $\rightarrow$ #classesFC $\rightarrow$ output

on CIFAR, where xFC denotes a fully connected layer with x neurons.
We also tested four combinations of prior and posterior with a learning rate of $\beta$

$$\eta = \theta - \mathbf{E}[\theta]$$

so that $\eta \sim \hat{\rho}(\theta)$, we get

$$C_\beta(X, Y; \hat{\rho}, \pi) = \mathbf{E}_{\theta \sim \hat{\rho}(\theta)} \hat{L}^{\text{cat}}_{X,Y}(f_\theta) + \beta \mathbf{KL}(\hat{\rho}(\theta) || \pi(\theta))$$

$$\approx \mathbf{E}_{\eta \sim \hat{\rho}(\theta)} [\eta^T \nabla \hat{L}^{\text{cat}}_{X,Y}(f_\theta) + \frac{1}{2} \eta^T \nabla^2 \hat{L}^{\text{cat}}_{X,Y}(f_\theta) \eta] + \beta \mathbf{KL}(\hat{\rho}(\theta) || \pi(\theta))$$

(5)

Figure 3. Empirical evaluation of the categorical cross-entropy loss: We take normalized random directions $v_i$, $i \in \{1, 2, 3, 4\}$ and plot the deterministic categorical cross-entropy loss $\hat{L}^{\text{cat}}_{X,Y}(f_\theta)$ for MNIST2 and CIFAR2 and values on the line $\theta = \theta_* + tv_i$, $t \in [-200, 200]$. We see that the loss closely reassembles a quadratic around the minimum $\theta_*$. 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.
\[ \approx E_{\eta \sim \hat{\rho}\theta}|\frac{1}{2} \eta^T H \eta + \beta KL(\hat{\rho}\theta||\pi(\theta)) \]

where \( H \equiv \nabla^2 \hat{L}_{X,Y}(f_\theta) \) 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 \( \geq 3 \). 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 posterior to have a concentrated behavior it has to be concentrated at locations where the true loss is well approximated by a quadratic. We have thus far dealt with Gaussian posteriors \( \hat{\rho}(\theta) = N(\mu_\theta, \sigma_\theta) \), where \( \forall i \), \( \sigma_{\theta i} \approx \lambda, 0.01 \leq \lambda \leq 1 \). It is well known 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 \( \forall i \), \( \sigma_{\theta i} = \lambda \), we can calculate \( E_{\eta \sim \hat{\rho}(\theta)}||\eta||_2^2 = E_{\eta \sim N(0, \Sigma)}[\sum_{i=0}^d \eta_i^2] = \sum_{i=0}^d \sigma_{\theta i}^2 = \lambda d \). Finally we expect that the radius of the "bubble" is \( E_{\eta \sim \hat{\rho}(\theta)}||\eta||_2 \approx \sqrt{\lambda d} \). 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 \( \hat{\rho}(\theta) = \hat{N}(\mu_\theta, \Sigma_\theta) \) and \( \pi(\theta) = \hat{N}(\mu_\pi, \lambda \Sigma_\pi) \). 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

\[ \min_{\Sigma_\theta} E_{\eta \sim \hat{\rho}(\theta)}|\frac{1}{2} \eta^T H \eta + \beta KL(\hat{\rho}(\theta)||\pi(\theta)) \] where \( \hat{\rho}(\theta) = \hat{N}(\mu_\theta, \Sigma_\theta) \) and \( \pi(\theta) = \hat{N}(\mu_\pi, \lambda \Sigma_\pi) \)

is minimized at

\[ \Sigma_\theta^* = \beta (H + \frac{\lambda}{\Sigma_\pi})^{-1} \]

where \( H \equiv \nabla^2 \hat{L}_{X,Y}(f_\theta) \) captures the curvature at the minimum, while \( \Sigma_\pi \) 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 \( \min_{\Sigma_\theta, \Sigma_\pi} E_{\eta \sim \hat{\rho}(\theta)}|\frac{1}{2} \eta^T H \eta + \beta KL(\hat{\rho}(\theta)||\pi(\theta)) \) with \( \hat{\rho}(\theta) = \hat{N}(\mu_\theta, \Sigma_\theta) \) and \( \pi(\theta) = \hat{N}(\mu_\pi, \lambda \Sigma_\pi) \)

\[ (\Sigma_\theta^*)^{-1} = \frac{1}{2 \beta} [h + \sqrt{h^2 + \frac{4 \beta h}{\lambda (\mu_{\theta i} - \mu_{\pi i})^2}}], \]

\[ (\Sigma_\pi^*)^{-1} = \frac{\lambda}{2 \beta} \left[ h^2 + \frac{4 \beta h}{\lambda (\mu_{\theta i} - \mu_{\pi i})^2} - h \right], \]

where \( H \equiv \nabla^2 \hat{L}_{X,Y}(f_\theta) \) 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 \( p(f_y|x) = \frac{e^{\theta f_y}}{e^{\theta f_\theta}} \) the Fisher Information matrix coincides with the generalized Gauss-Newton approximation of the Hessian (Kunstner et al., 2019). We sample one output \( \hat{y}_i \) from the model distribution \( p(y_i|f_\theta(x_i)) \) for each input \( x_i \), and approximate

\[ \nabla \log p(\hat{y}_i|f_\theta(x_i)) = \nabla \log p(\hat{y}_i|f_\theta(x_i))^T, \]

containing 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 \( \beta \) and \( \lambda \). 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...
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 inherent limitations of the mean-field approximation, as we typically don’t even have access to the optimal prior covariance.

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 Khatiri-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 $\lambda$ and $\beta$, corresponding to $O(10^2)$ 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

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.
Sampling efficiently from the posterior. At the same time we will need to sample efficiently between $O(10^4)$ and $O(10^5)$ 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 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 $\ell_{\text{mae}}(f, x, y) = (1/c) \sum_{i=0}^{c} (f(x)_i - y_i)^2$. Assuming $r$ neurons per layer, $\theta$ has a form $\theta = [\text{vec}(W_0^0) \text{vec}(W_1^1) \cdots \text{vec}(W_r^r)]$. We also denote for layer $i$ and neuron $j$, $\theta_{ij}$, $\mu_{\beta ij}$, $\Sigma_{\beta ij}$, $\mu_{\pi ij}$ the corresponding split variables. We can then motivate optimizing the following surrogate upper bound

$$E_{\theta \sim \hat{p}(\theta)} \tilde{\mathcal{L}}_{X,Y}(f_\theta) + \frac{1}{\beta n} (\text{KL}(\hat{p}(\theta) || \mathcal{N}(\mu_x, \lambda I)) + \ln \frac{1}{\beta})$$  

(9)

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.

5.2. Empirical Results

We now present results on the MNIST datasets. We run a grid search over $\beta$ and $\lambda$, 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


Dissecting Non-Vacuous Generalization Bounds


