

The Fine Print on Tempered Posteriors

Konstantinos Pitas
and Julyan Arbel

PITAS.KONSTANTINOS@INRIA.FR

JULYAN.ARBEL@INRIA.FR

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

We conduct a detailed investigation of tempered posteriors and uncover a number of crucial and previously undiscussed points. Contrary to previous results, we first show that for realistic models and datasets and the tightly controlled case of the Laplace approximation to the posterior, stochasticity does not in general improve test accuracy. The coldest temperature is often optimal. One might think that Bayesian models with some stochasticity can at least obtain improvements in terms of calibration. However, we show empirically that when gains are obtained this comes at the cost of degradation in test accuracy. We then discuss how targeting Frequentist metrics using Bayesian models provides a simple explanation of the need for a temperature parameter λ in the optimization objective. Contrary to prior works, we finally show through a PAC-Bayesian analysis that the temperature λ cannot be seen as simply fixing a misspecified prior or likelihood.

Keywords: Tempered Posteriors; Cold Posteriors; PAC-Bayes

1. Introduction

An influential line of work starting from [Wenzel et al. \(2020\)](#) highlighted that Bayesian neural networks typically exhibit better test-time predictive performance if the posterior distribution is “sharpened” through tempering. Tempering with a temperature λ is either applied to both the data likelihood and the prior resulting in cold posteriors or only to the data likelihood resulting in tempered posteriors. The empirical need for this tempering highlights a fundamental mismatch between Bayesian theory and practice. As the number of training samples increases, Bayesian theory states that the posterior distribution should be concentrating more and more on the true model parameters, under mild regularity assumptions. At any time, the posterior is our best guess at the true model parameters, without having to resort to heuristics.

The experimental setups where the need for tempering arises have been hard to pinpoint precisely. Likelihood misspecification has proven to be a prominent explanation. [Nabarro et al. \(2022\)](#); [Bachmann et al. \(2022\)](#); [Kapoor et al. \(2022\)](#); [Aitchison \(2021\)](#) argue that the aleatoric uncertainty in the likelihood function becomes misspecified in many experiments due to data augmentation and/or due to curation of the data. A common thread in the above works is that a correct likelihood would lead to optimal test performance and tempering would be unnecessary. In the original paper [Wenzel et al. \(2020\)](#) as well as subsequent ones ([Nabarro et al., 2022](#); [Bachmann et al., 2022](#); [Kapoor et al., 2022](#); [Aitchison, 2021](#)) another conclusion is that Bayesian neural networks outperform deterministic ones in test metrics for the correct value of λ .

We challenge both theoretically and experimentally the above assumptions and conclusions. In a strict sense, we will be focusing on tempered posteriors, while noting that for Gaussian priors and posteriors the two can be seen as equivalent (Aitchison, 2021). We make the following four contributions.

We first observe that in previous works (Nabarro et al., 2022; Bachmann et al., 2022; Kapoor et al., 2022; Aitchison, 2021) even after “fixing” the likelihood with various approaches, the test accuracy seems to be improving as the temperature is decreasing, though it is implied that some stochasticity is beneficial. In the tightly controlled case of the Laplace approximation, we show experimentally that in terms of test accuracy, stochasticity does not help (Section 4). In contrast to previous works, our conclusion is that the *deterministic* network is nearly always optimal.

Second, one can question if such an evaluation is fair for Bayesian models. Indeed practitioners often motivate the use of Bayesian models so as to improve the calibration of predictions. We empirically show that for the negative log-likelihood (NLL) and the expected calibration error (ECE), multiple values of λ are optimal, i.e. some stochasticity helps (Section 4). Crucially though, improving the test ECE comes in general at the cost of reducing test accuracy.

Third, given the poor performance of purely Bayesian models for Frequentist metrics, we then posit that there exists a simpler explanation for cold and tempered posteriors, specifically that Bayesian inference *does not readily provide high-probability guarantees on out-of-sample performance* (Section 5). Existing theorems simply describe *posterior contraction* to the true posterior (Ghosal et al., 2000; Blackwell and Dubins, 1962). However, practitioners (Nabarro et al., 2022; Bachmann et al., 2022; Kapoor et al., 2022; Aitchison, 2021) evaluate posteriors on Frequentist test metrics such as test accuracy. This discrepancy between using Bayesian models and evaluating on Frequentist metrics is enough to justify the use of a temperature parameter. To prove out-of-sample performance with high probability, one needs a generalization bound such as a PAC-Bayes bound (McAllester, 1999; Catoni, 2007; Alquier et al., 2016; Dziugaite and Roy, 2017). The tightest of these bounds (Catoni, 2007) naturally incorporates a temperature parameter λ .

Finally, we investigate the question of whether tempered posteriors can be seen as fixing a misspecified likelihood (Section 6). With a simplified PAC-Bayesian analysis, we argue that tempered posteriors do not only compensate for a misspecified aleatoric uncertainty, contrary to Nabarro et al. (2022); Bachmann et al. (2022); Kapoor et al. (2022); Aitchison (2021). Our bound implies that the specific weight instantiation of a neural network influences the choice of λ . We confirm this experimentally for the case of the Laplace approximation. We use different MAP estimates to fit the Laplace approximation. For a fixed value of λ , the test accuracy changes even though *both the likelihood function and the inherent aleatoric uncertainty of the data remain the same*.

We also include a detailed FAQ section in the Appendix, together with all the proofs and additional experiments.

2. Related work

Aitchison (2021) argues that tempering results from the fact that common computer vision benchmarks are curated. That is, each label has been independently verified by multiple

labelers, thus the aleatoric uncertainty for all labels is low. He proposes a generative model that implies a tempered likelihood that accounts for data curation. [Kapoor et al. \(2022\)](#) describe the softmax-cross-entropy loss as resulting from a multinomial likelihood function together with a Dirichlet prior distribution. They propose a new likelihood and a corresponding Dirichlet prior which explicitly models aleatoric uncertainty through an additional hyper-parameter. They argue that a softmax-cross-entropy loss does not explicitly allow tuning the aleatoric uncertainty of the data, thus necessitating tempered posteriors. [Bachmann et al. \(2022\)](#) argue that data augmentation reduces the effective number of training samples. Thus tempering reduces the nominal number of training samples so as to match the effective number. [Nabarro et al. \(2022\)](#) explore whether data-augmentation *increases* the effective sample size. They propose a new valid likelihood that takes into account data augmentations and discover that the cold-posterior effect persists. Despite the above explanations, the need for tempering persists even in settings without data curation and augmentation, and when the likelihood is suitably modified.

Certifying generalization to out-of-sample data has been extensively explored for the case of deep neural networks ([Dziugaite et al., 2021](#); [Dziugaite and Roy, 2017](#); [Zhou et al., 2018](#); [Lotfi et al., 2022](#); [Bartlett et al., 2017](#)). The current state-of-the-art bounds ([Dziugaite et al., 2021](#); [Lotfi et al., 2022](#)) utilize the Catoni PAC-Bayes bound [Catoni \(2007\)](#), together with a data-dependent prior, and find non-vacuous and in some cases tight bounds. We base part of our arguments on the Catoni bound.

The posterior contraction properties of Bayesian inference have been extensively studied ([Ghosal et al., 2000](#); [Blackwell and Dubins, 1962](#)). Of note are results that show contraction even in the case of misspecification [Grünwald \(2012\)](#); [Grünwald and Van Ommen \(2017\)](#); [Grünwald and Langford \(2007\)](#); [Bhattacharya et al. \(2019\)](#), using tempered posteriors restricted to $\lambda < 1$ (warm posteriors).

[Grünwald and Langford \(2007\)](#); [Grünwald \(2012\)](#); [Grünwald and Van Ommen \(2017\)](#) explore Safe-Bayes, which finds a temperature parameter λ by taking a sequential view of Bayesian inference. They find a Cèsaro averaged posterior, which is an average of the posteriors at different optimization steps, and which does not coincide with the standard posterior. The Safe-Bayes analysis is also restricted to the case where $\lambda < 1$. By contrast we provide an analytical expression of the bound on true risk, given λ , and also numerically investigate the case of $\lambda > 1$. Our analysis thus provides intuition regarding which parameters (for example the curvature) might result in tempering. [Catoni \(2007\)](#) discusses the optimal value of the temperature λ for PAC-Bayes bounds, for *fixed* priors and posteriors. By contrast we investigate the case where the posterior is optimized for different λ and which is the relevant one for the tempering literature. [Bhattacharya et al. \(2019\)](#) derive a general form of the PAC-Bayes inequality albeit to analyze posterior contraction. In several prediction settings, out-of-sample performance can however be established directly using this result.

Modern contraction analyses and PAC-Bayes bounds are closely linked ([Bhattacharya et al., 2019](#); [Grünwald, 2012](#)). It is thus important to base our results on the tightest analyses available for the deep learning setting. As previously mentioned, for Bayesian neural networks the particular Catoni form of the PAC-Bayesian inequality has been empirically shown to result in state-of-the-art bounds. We, therefore, use the Catoni bound in our analyses where applicable.

3. Limitations

We base our analysis on the Laplace approximation to the posterior. Therefore the question arises as to whether our results generalize to other inference methods. Here we simply note that our empirical results align with existing results in the MCMC case such as Kapoor et al. (2022); Nabarro et al. (2022); Noci et al. (2021). A further limitation is that PAC-Bayes bounds are sometimes (relatively) tight but more often they are loose. This brings into question the practical utility of a PAC-Bayes theoretical analysis.

4. Stochasticity typically hurts performance

We first explore whether any stochasticity improves the test accuracy, compared to a deterministic network. Here, working in the Laplace approximation setting is crucial. Previous studies dealt with MCMC methods and have not adequately explored what happens when the temperature becomes “frozen”. The case of Wenzel et al. (2020) is striking as the “frozen” network and the deterministic network test accuracies do not match (Figure 1 of Wenzel et al., 2020). The Laplace approximation has a simple intuitive behaviour: the posterior collapses on a Dirac delta at the MAP as $\lambda \rightarrow +\infty$. We then investigate how λ affects the other most popular test metrics (NLL and ECE). In particular we explore whether one can improve the ECE without hurting test accuracy, which is the most pertinent question for practitioners.

4.1. Definitions

We denote the learning sample $(X, Y) = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, that contains n input-output pairs. Observations (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 $\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, where \mathcal{F} is a set of predictors $f : \mathcal{X} \rightarrow \mathcal{Y}$. We also denote the risk $\mathcal{L}_{\mathcal{D}}^{\ell}(f) = \mathbf{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \ell(f, \mathbf{x}, y)$ and the empirical risk $\hat{\mathcal{L}}_{X, Y}^{\ell}(f) = (1/n) \sum_i \ell(f, \mathbf{x}_i, y_i)$. We consider two probability measures, the prior $\pi \in \mathcal{M}(\mathcal{F})$ and the approximate posterior $\hat{\rho} \in \mathcal{M}(\mathcal{F})$. Here, $\mathcal{M}(\mathcal{F})$ denotes the set of all probability measures on \mathcal{F} . We encounter cases where we make predictions using the approximate posterior predictive distribution $\mathbf{E}_{f \sim \hat{\rho}}[p(y|\mathbf{x}, f)]$. We will use two loss functions, the non-differentiable zero-one loss $\ell_{01}(f, \mathbf{x}, y) = \mathbb{I}(\arg \max_j f(\mathbf{x})_j \neq y)$, and the negative log-likelihood (NLL), which is a commonly used differentiable surrogate $\ell_{\text{nll}}(f, \mathbf{x}, y) = -\log(p(y|\mathbf{x}, f))$. We assume that outputs of f form a probability distribution $p(y|\mathbf{x}, f)$ either through a Gaussian likelihood (in the case of regression) or using the softmax activation function (in the case of classification). Given the above, the Evidence Lower Bound (ELBO) has the following form

$$-\mathbf{E}_{f \sim \hat{\rho}} \hat{\mathcal{L}}_{X, Y}^{\ell_{\text{nll}}}(f) - \frac{1}{\lambda n} \text{KL}(\hat{\rho} \parallel \pi), \quad (1)$$

where $\lambda = 1$. Note that our temperature parameter λ is the *inverse* of the parameter $T = 1/\lambda$ typically used in tempered posterior works. Here, cold posteriors are the result of $\lambda > 1$. Our setup is discussed in Wenzel et al. (2020), p3 Section 2.3, and used in Bachmann et al. (2022); Aitchison (2021); Kapoor et al. (2022). While Wenzel et al. (2020) use MCMC to conduct their experiments, we opt for the ELBO for analytical tractability.

4.2. Experimental setup: Laplace approximation

The ELBO (1) is minimized at the probability density $\rho^*(f)$ given by [Catoni \(2007\)](#):

$$\rho^*(f) := \pi(f) e^{-\lambda n \hat{\mathcal{L}}_{X,Y}^{\text{nl}}(f)} / \mathbf{E}_{f \sim \pi} \left[e^{-\lambda n \hat{\mathcal{L}}_{X,Y}^{\text{nl}}(f)} \right].$$

We will use the Laplace approximation to the posterior in our experiments. This is equivalent to approximating $\lambda n \hat{\mathcal{L}}_{X,Y}^{\text{nl}}(f)$ using a second order Taylor expansion around a minimum $\mathbf{w}_{\hat{\rho}}$, such that

$$\lambda n \hat{\mathcal{L}}_{X,Y}^{\text{nl}}(f_{\mathbf{w}}) \approx \lambda n \hat{\mathcal{L}}_{X,Y}^{\text{nl}}(f_{\mathbf{w}_{\hat{\rho}}}) + \frac{\lambda}{2} (\mathbf{w} - \mathbf{w}_{\hat{\rho}})^\top \mathbf{H} (\mathbf{w} - \mathbf{w}_{\hat{\rho}}), \quad (2)$$

where \mathbf{H} is the network Hessian $n \nabla \nabla \hat{\mathcal{L}}_{X,Y}^{\text{nl}}(f_{\mathbf{w}})|_{\mathbf{w}=\mathbf{w}_{\hat{\rho}}}$. Assuming a Gaussian prior $\pi = \mathcal{N}(0, \sigma_\pi^2 \mathbf{I})$, the Laplace approximation to the posterior $\hat{\rho}$ is again Gaussian

$$\hat{\rho} = \mathcal{N}(\mathbf{w}_{\hat{\rho}}, (\lambda \mathbf{H} + \mathbf{I}/\sigma_\pi^2)^{-1})$$

The Hessian is generally infeasible to compute in practice for modern deep neural networks, such that many approaches employ the generalized Gauss–Newton (GGN) approximation $\mathbf{H}^{\text{GGN}} := \sum_{i=1}^n \mathcal{J}_{\mathbf{w}}(\mathbf{x}_i)^\top \mathbf{\Lambda}(\mathbf{y}_i; f_i) \mathcal{J}_{\mathbf{w}}(\mathbf{x}_i)$, where $\mathcal{J}_{\mathbf{w}}(\mathbf{x})$ is the network per-sample Jacobian $[\mathcal{J}_{\mathbf{w}}(\mathbf{x})]_c = \nabla_{\mathbf{w}} f_c(\mathbf{x}; \mathbf{w}_{\hat{\rho}})$, and $\mathbf{\Lambda}(\mathbf{y}; f) = -\nabla_{f f}^2 \log p(\mathbf{y}; f)$ is the per-input noise matrix ([Kunstner et al., 2019](#)). We will use two simplified versions of the GGN: (i) an isotropic approximation with variance $\sigma_{\hat{\rho}}^2(\lambda)$ equal to $\left(\frac{\lambda h}{d} + \frac{1}{\sigma_\pi^2}\right)^{-1}$, and h equal to $\text{tr}(\mathbf{H}^{\text{GGN}}) = \sum_{i,j,k} [\mathbf{\Lambda}(\mathbf{y}_i; f)]_{kk} (\nabla_{\mathbf{w}} f_k(\mathbf{x}_i; \mathbf{w}_{\hat{\rho}})_j)^2$, and (ii) the Kronecker-Factorized Approximate Curvature (KFAC) approximation ([Martens and Grosse, 2015](#)) which only retains a block-diagonal part of the GGN.

4.3. Experimental setup: training

When making predictions, we use the posterior predictive distribution $\mathbf{E}_{\mathbf{w} \sim \hat{\rho}}[p(y|\mathbf{x}, f_{\mathbf{w}})]$ of the *full neural network model*, meaning that samples from $\hat{\rho}$ are inputted to the full neural network (as opposed to, for example, its linearization). Since the 0-1 loss is not differentiable, the posterior estimated with the cross-entropy loss will be used for classification problems.

We have tested extensively in realistic classification tasks. We used the CIFAR-10, CIFAR-100 ([Krizhevsky and Hinton, 2009](#)), SVHN ([Netzer et al., 2011](#)) and FashionMnist ([Xiao et al., 2017](#)) datasets. In all experiments, we split the dataset into three sets. These three are the typical prediction tasks sets: training set Z_{train} , testing set Z_{test} , and validation set $Z_{\text{validation}}$. We use Monte Carlo sampling with 100 samples to make predictions. For the CIFAR-10, CIFAR-100, and SVHN datasets, we use a WideResNet22 ([Zagoruyko and Komodakis, 2016](#)), with Fixup initialization ([Zhang et al., 2019](#)). For the FashionMnist dataset, we use a convolutional architecture with three convolutional layers, followed by two fully connected non-linear layers. More details on the experimental setup can be found in the Appendix.

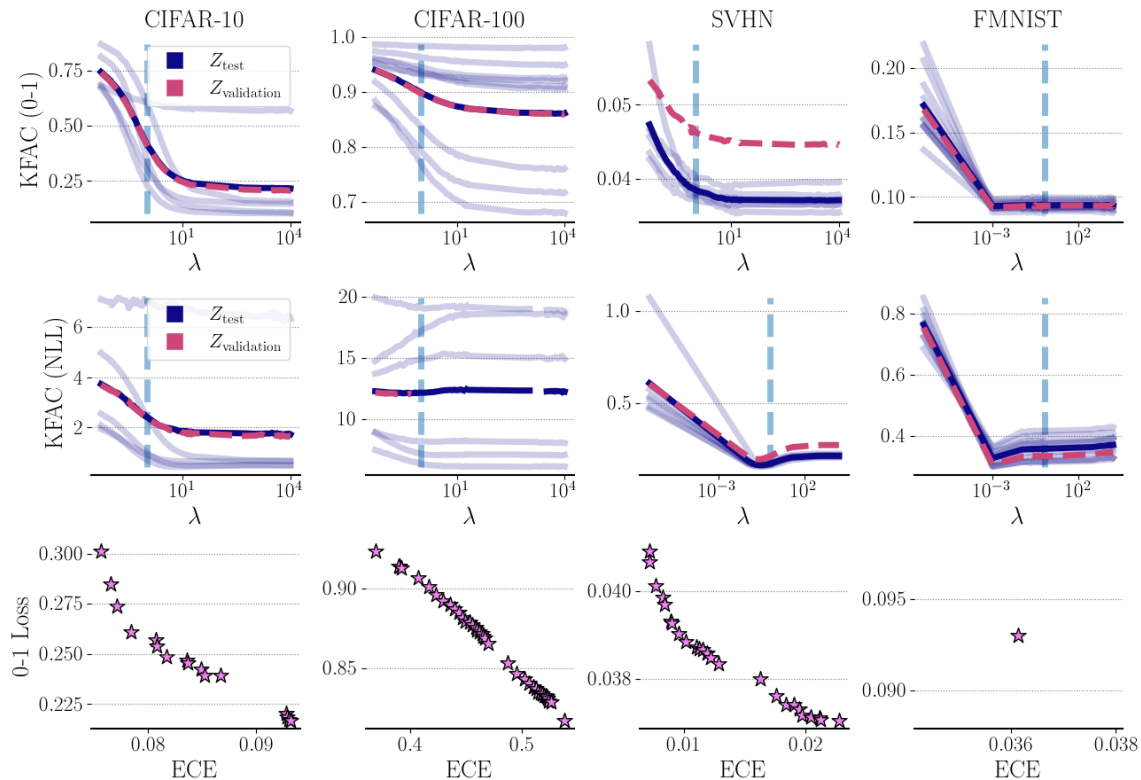


Figure 1: Test 0-1 loss mean — , as well as 10 MAP trials — , along with the mean of the validation set -- -- (we denote $\lambda = 1$ by -- --): KFAC Laplace 0-1 loss (top) and KFAC NLL (middle) and 0-1 loss vs ECE (bottom row). For the standard KFAC posteriors the test and validation 0-1 loss have a rapid improvement as $\lambda \uparrow$ followed by a plateau. Coldest posteriors $\lambda \gg 1$ are almost always best. The NLL has more complex behaviour, with cold and warm posteriors depending on the dataset. There is a clear tradeoff between 0-1 loss and ECE.

We find ten MAP estimates for the neural network weights of the CIFAR-10, CIFAR-100, SVHN and FMNIST datasets by training on Z_{train} using SGD. We then fit a KFAC Laplace approximation to each MAP estimate using $(X, Y) = Z_{\text{train}}$ and also choose the prior through optimizing the marginal likelihood. We estimate the 0-1 loss, negative log-likelihood (NLL), and Expected Calibration Error (ECE) of the posterior predictive on Z_{test} and $Z_{\text{validation}}$.

4.4. Observations

We first test a setting without data augmentation. We plot the results for the 0-1 loss in Figure 1 (top row). *Contrary to Wenzel et al. (2020) in terms of test 0-1 loss, the MAP estimate obtained where $\lambda \gg 1$ and the posterior is “coldest” is almost always optimal.* This result is more coherent than the one in Wenzel et al. (2020) (Figure 1) where the coldest temperature 0-1 test loss and the MAP 0-1 test loss don’t match. It highlights that in a

tightly controlled setting, Bayesian approaches often don't improve at all over deterministic ones in terms of test Frequentist metrics such as the 0-1 loss, *for any temperature* λ .

We then plot in Figure 1 (middle row) the results for the NLL. Even without data augmentation and even when we optimize the prior variance using the marginal likelihood, we find that all three cases of temperatures (cold posterior, warm posterior, as well as posterior with $\lambda = 1$) can be optimal, for varying datasets. *This highlights the importance of the choice of the evaluation metric when discussing the cold posterior effect, as results can vary significantly depending on our choice.*

We then elaborate on this finding and further ask the crucial question: *If the Laplace approximation does not improve the test 0-1 loss in general, can it retain (for some value of λ) the same test 0-1 loss while improving calibration?* In Figure 1, we see that in most of our experiments, we couldn't find such a temperature λ . Apart from FMNIST, *there seems to be a clear tradeoff between test 0-1 loss and calibration error.*

We repeat the experiment for the case of data augmentation, for the CIFAR-10 and CIFAR-100 datasets. We use the standard augmentations of random crops and rotations. We perform augmentation both when estimating the MAP, as well as when fitting the Laplace approximation. We include the results in the Appendix. We see similar behavior to the setting without data augmentation. Our only additional observation is that test accuracy has improved for both CIFAR-10 and CIFAR-100 (without augmentation not only does training suffer but numerical issues arise when performing matrix inversion for the Laplace approximation).

Empirically (for the Laplace case) the coldest temperature is almost always optimal in terms of test accuracy. Even though one can improve calibration for a given temperature, this comes at the cost of accuracy.

5. Tempered Posteriors: A simple explanation

In light of the generally poor performance of purely Bayesian methods in terms of test Frequentist metrics for deep learning, we explore a simple explanation regarding the need for tempering: the mismatch between Bayesian guarantees and practice. We start by noting that in supervised prediction, what we often try to minimize is

$$\text{KL}(p_{\mathcal{D}}(y|\mathbf{x})\|\mathbf{E}_{f\sim\hat{\rho}}[p(y|\mathbf{x}, f)]) = \mathbf{E}_{\mathbf{x}, y\sim\mathcal{D}} \left[\ln \frac{p_{\mathcal{D}}(y|\mathbf{x})}{\mathbf{E}_{f\sim\hat{\rho}}[p(y|\mathbf{x}, f)]} \right], \quad (3)$$

the conditional relative entropy (Cover, 1999) between the true conditional distribution $p_{\mathcal{D}}(y|\mathbf{x})$ and the posterior predictive distribution $\mathbf{E}_{f\sim\hat{\rho}}[p(y|\mathbf{x}, f)]$. For example, this is implicitly the minimized quantity when optimizing classifiers with the cross-entropy loss (Masegosa, 2020; Morningstar et al., 2022). It is also on this and similar predictive metrics that the tempering effect appears. In the following, we will outline the relationship between the ELBO, PAC-Bayes, and (3).

5.1. ELBO

We assume a training sample $(X, Y) \sim \mathcal{D}^n$ as before, denote $p(\mathbf{w}|X, Y)$ the true posterior probability over predictors f parameterized by \mathbf{w} (typically weights for neural networks), and π and $\hat{\rho}$ respectively the prior and variational posterior distributions as before. The ELBO has the following form

$$\text{KL}(\hat{\rho}(\mathbf{w})\|p(\mathbf{w}|X, Y)) = -n \underbrace{\left(-\mathbf{E}_{f \sim \hat{\rho}} \hat{\mathcal{L}}_{X, Y}^{\ell_{\text{all}}}(f) - \frac{1}{n} \text{KL}(\hat{\rho} \|\pi) \right)}_{\text{ELBO}} + \ln p(Y|X).$$

Thus, maximizing the ELBO can be seen as minimizing the KL divergence between the true posterior and the variational posterior over the weights $\text{KL}(\hat{\rho}(\mathbf{w})\|p(\mathbf{w}|X, Y))$. The true posterior distribution $p(\mathbf{w}|X, Y)$ gives more probability mass to predictors which are more likely given the training data, *however these predictors do not provably minimize* $\text{KL}(p_{\mathcal{D}}(y|\mathbf{x})\|\mathbf{E}_{f \sim \hat{\rho}}[p(y|\mathbf{x}, f)])$, the evaluation metric of choice for supervised prediction. Similar issues exist with theorems on posterior contraction. These typically guarantee that the posterior concentrates on the most probable set of parameters, in the well-specified regime (Blackwell and Dubins, 1962). When we are not interested in inferring the most probable parameters but in predicting on new data, it is important to derive a certificate of generalization through a generalization bound, which directly bounds out-of-sample performance. In the following, we focus on analyzing a PAC-Bayes bound in order to obtain insights into when tempering is necessary. As discussed in Section 2, modern contraction analyses are closely related to PAC-Bayes, and one can sometimes move from a contraction result to a generalization bound. However, the specific forms discussed in the following section currently provide state-of-the-art generalization bounds for modern deep neural networks (Dziugaite et al., 2021; Lotfi et al., 2022).

5.2. PAC-Bayes

We first look at the following bound denoted by $\mathcal{B}_{\text{Alquier}}$. It was considered by Alquier et al. (2016) (Theorem 4.1); see also Theorem 1 by Masegosa (2020) for a statement under the same conditions. We will see that it is natural to assume a temperature λ when trying to bound out-of-sample performance, as opposed to guaranteeing posterior contraction.

Theorem 1 ($\mathcal{B}_{\text{Alquier}}$, Alquier et al., 2016) *Given a distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$, a hypothesis set \mathcal{F} , a loss function $\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, a prior distribution π over \mathcal{F} , real numbers $\delta \in (0, 1]$ and $\lambda > 0$, with probability at least $1 - \delta$ over the choice $(X, Y) \sim \mathcal{D}^n$, we have for all $\hat{\rho}$ on \mathcal{F}*

$$\mathbf{E}_{f \sim \hat{\rho}} \mathcal{L}_{\mathcal{D}}^{\ell}(f) \leq \mathbf{E}_{f \sim \hat{\rho}} \hat{\mathcal{L}}_{X, Y}^{\ell}(f) + \frac{1}{\lambda n} \left[\text{KL}(\hat{\rho} \|\pi) + \ln \frac{1}{\delta} + \Psi_{\ell, \pi, \mathcal{D}}(\lambda, n) \right]$$

where $\Psi_{\ell, \pi, \mathcal{D}}(\lambda, n)$ is equal to $\ln \mathbf{E}_{f \sim \pi} \mathbf{E}_{X', Y' \sim \mathcal{D}^n} \exp \left[\lambda n \left(\mathcal{L}_{\mathcal{D}}^{\ell}(f) - \hat{\mathcal{L}}_{X', Y'}^{\ell}(f) \right) \right]$.

There are three different terms in the above bound. The empirical risk term $\mathbf{E}_{f \sim \hat{\rho}} \hat{\mathcal{L}}_{X, Y}^{\ell}(f)$ is the empirical mean of the loss of the classifier over all training samples. The KL term

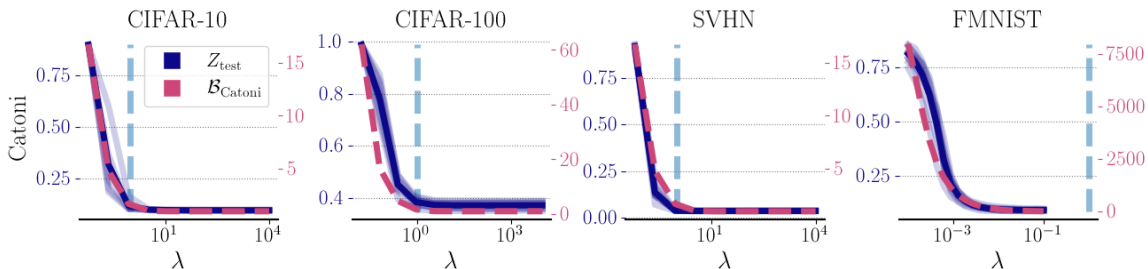


Figure 2: We plot the Catoni bound, as well as the test misclassification error for the CIFAR-10, CIFAR-100, SVHN and FMNIST datasets and the Isotropic Laplace approximation. We see that the Catoni bounds tightly correlate with the test error. As the temperature λ increases, the bound predicts a smaller test error, this is reflected in the empirical data as the need for a tempered posterior.

$1/(\lambda n)\text{KL}(\hat{\rho}||\pi)$ is the complexity of the model, which in this case is measured as the KL-divergence between the posterior and prior distributions. The Moment term $1/(\lambda n)\Psi_{\ell,\pi,\mathcal{D}}(\lambda, n)$ is the log-Laplace transform of the difference between the risk and empirical risk for a reversal of the temperature. We can typically make assumptions on the tails of the loss to ensure that the Moment term is bounded (see eg [Germain et al., 2016](#)). Using Theorem 1 together with Jensen’s inequality, one can bound (3) directly as follows

$$\begin{aligned}
 & \text{KL}(p_{\mathcal{D}}(y|\mathbf{x})||\mathbf{E}_{f\sim\hat{\rho}}[p(y|\mathbf{x}, f)]) \\
 &= \mathbf{E}_{\mathbf{x},y\sim\mathcal{D}}[-\ln \mathbf{E}_{f\sim\hat{\rho}}[p(y|\mathbf{x}, f)]] + \mathbf{E}_{\mathbf{x},y\sim\mathcal{D}}[\ln p_{\mathcal{D}}(y|\mathbf{x})] \\
 &\leq \mathbf{E}_{\mathbf{x},y\sim\mathcal{D}}[\mathbf{E}_{f\sim\hat{\rho}}[-\ln p(y|\mathbf{x}, f)]] + \mathbf{E}_{\mathbf{x},y\sim\mathcal{D}}[\ln p_{\mathcal{D}}(y|\mathbf{x})] \\
 &\leq \underbrace{\mathbf{E}_{f\sim\hat{\rho}}[\hat{\mathcal{L}}_{X,Y}^{\ell_{\text{nl}}}(f)] + \frac{1}{\lambda n} \left[\text{KL}(\hat{\rho}||\pi) + \ln \frac{1}{\delta} + \Psi_{\ell_{\text{nl}},\pi,\mathcal{D}} \right]}_{\text{PAC-Bayes}} + \mathbf{E}_{\mathbf{x},y\sim\mathcal{D}}[\ln p_{\mathcal{D}}(y|\mathbf{x})].
 \end{aligned}$$

The last line holds under the conditions of Theorem 1 and in particular with probability at least $1 - \delta$ over the choice $(X, Y) \sim \mathcal{D}^n$. Notice here the presence of the temperature parameter $\lambda \geq 0$, which need not be $\lambda = 1$.

It is easy to see that maximizing the ELBO is equivalent to minimizing a PAC-Bayes bound for $\lambda = 1$, which might not necessarily be optimal for a finite sample size. More specifically even for exact inference the Bayesian posterior predictive distribution does not necessarily minimize $\text{KL}(p_{\mathcal{D}}(y|\mathbf{x})||\mathbf{E}_{f\sim\hat{\rho}}[p(y|\mathbf{x}, f)])$.

We note that the temperature is inherent in the derivation of PAC-Bayes bounds ([Bégin et al., 2016](#)) (see also Appendix). It more generally occurs when trying to apply Chernoff bounds on the tails of random variables. It is necessary here because we are trying to bound out-of-sample performance, with high probability.

5.3. Classification tasks

For classification tasks, we are typically mainly interested in achieving low expected zero-one risk $\mathbf{E}_{f\sim\hat{\rho}}\mathcal{L}_{\mathcal{D}}^{\ell_{01}}(f)$. The ELBO objective is not directly related to this risk. However in the

PAC-Bayesian literature, there exist bounds specifically adapted to it. In the following, we will use one of the tightest and most commonly used bounds, the ‘‘Catoni’’ bound, denoted $\mathcal{B}_{\text{Catoni}}$ from [Catoni \(2007\)](#) Theorem 1.2.6.

Theorem 2 ($\mathcal{B}_{\text{Catoni}}$, [Catoni, 2007](#)) *Given a distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$, a hypothesis set \mathcal{F} , the 0-1 loss function $\ell_{01} : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]$, a prior distribution π over \mathcal{F} , a real number $\delta \in (0, 1]$, and a real number $\lambda > 0$, with $\Phi_\lambda^{-1}(x) = \frac{1-e^{-\lambda x}}{1+e^{-\lambda x}}$, we have with probability at least $1 - \delta$ over $(X, Y) \sim \mathcal{D}^n$*

$$\forall \hat{\rho} \text{ on } \mathcal{F} : \mathbf{E}_{f \sim \hat{\rho}} \mathcal{L}_{\mathcal{D}}^{\ell_{01}}(f) \leq \Phi_\lambda^{-1} \left(\mathbf{E}_{f \sim \hat{\rho}} \hat{\mathcal{L}}_{X,Y}^{\ell_{01}}(f) + \frac{1}{\lambda n} \left[\text{KL}(\hat{\rho} \parallel \pi) + \ln \frac{1}{\delta} \right] \right).$$

Similarly to the Alquier bound, the empirical risk term is the empirical mean of the loss of the classifier over all training samples. The KL term is the complexity of the model, which in this case is measured as the KL divergence between the posterior and prior distributions. The Moment term has been absorbed in this case in the Φ_λ^{-1} function.

5.4. Experiments

PAC-Bayes bounds require correct control of the prior mean as the ℓ_2 distance between prior and posterior means in the KL term is often the dominant term in the bound. To control this distance, we follow a variation of the approach in [Dziugaite et al. \(2021\)](#) to construct our classifiers. We first use Z_{train} to find a prior mean \mathbf{w}_π . We then set the posterior mean equal to the prior mean $\mathbf{w}_{\hat{\rho}} = \mathbf{w}_\pi$ but evaluate the r.h.s of the bounds on $Z_{\text{validation}}$. Note that in this way $\|\mathbf{w}_{\hat{\rho}} - \mathbf{w}_\pi\|_2^2 = 0$, while the bound is still valid since the prior is independent of the evaluation set $(X, Y) = Z_{\text{validation}}$. We fit an Isotropic Laplace approximation to the posterior (using $(X, Y) = Z_{\text{validation}}$). This is because for the isotropic Laplace approximation and a Gaussian isotropic prior, the KL divergence has a simple analytical expression $\text{KL}(\hat{\rho} \parallel \pi) = \frac{1}{2} \left(d \frac{\sigma_{\hat{\rho}}^2(\lambda)}{\sigma_\pi^2} + \frac{\|\mathbf{w}_{\hat{\rho}} - \mathbf{w}_\pi\|_2^2}{\sigma_\pi^2} - d - d \ln \sigma_{\hat{\rho}}^2(\lambda) + d \ln \sigma_\pi^2 \right)$.

For different values of λ we then estimate the Catoni bound (Theorem 2) using $Z_{\text{validation}}$. We also estimate the *test* 0-1 loss. We use the prior variance $\sigma_\pi^2 = 0.1$, as optimizing the marginal likelihood leads to $\sigma_{\hat{\rho}}^2 \approx 0$ which is not relevant for BNNs. We plot the results in [Figure 2](#). The Catoni bound correlates tightly with test 0-1 loss for all datasets.

The temperature in PAC-Bayes bounds (which is also present in all Chernoff bounds) is a plausible explanation for the need of tempering. It should also be considered first by researchers, before resorting to more involved explanations.

6. Temperature λ , aleatoric uncertainty and curvature

How does PAC-Bayes align with prior explanations of tempering? In particular, it is interesting to explore the most prominent explanation, which suggests that the temperature λ fixes our wrong assumptions about aleatoric uncertainty, encoded in our choice of the likelihood and/or the prior. PAC-Bayes objectives are difficult to analyze theoretically in non-convex cases. Thus in the following we make a number of simplifying assumptions.

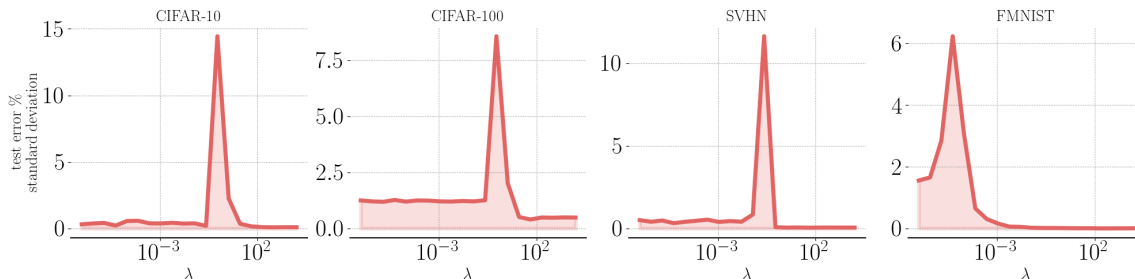


Figure 3: λ does not simply fix a misspecified likelihood or prior. For fixed λ and noting that the prior σ_π^2 and the dataset are also fixed there is large variability in the test 0-1 loss given multiple Laplace approximations. Since the inherent aleatoric uncertainty in this experiment is fixed, we shouldn’t be observing any variability in test error for fixed λ . For $\lambda \gg 1$ all the Laplace approximations degenerate to deterministic networks. Conversely, for $\lambda \ll 1$ all the Laplace approximations degenerate to random predictions. In both cases, the standard deviation of the error is almost 0.

First of all we focus our analysis on a linearized model around a minimum of the loss landscape. There are a number of factors supporting this choice. The Laplace approximation with the Generalized Gauss–Newton approximation to the Hessian corresponds to a linearization of the neural network around the MAP estimate $\mathbf{w}_\rho \in \mathbb{R}^d$ (Immer et al., 2021) $f_{\text{lin}}(\mathbf{x}; \mathbf{w}) = f(\mathbf{x}; \mathbf{w}_\rho) + \nabla_{\mathbf{w}} f(\mathbf{x}; \mathbf{w}_\rho)^\top (\mathbf{w} - \mathbf{w}_\rho)$. Also, prior work has shown that linearization is reasonable when analyzing minima of the loss landscape. This is without having to resort to common assumptions about infinite width (Zancato et al., 2020; Maddox et al., 2021). We adopt the aforementioned linear form together with the Gaussian likelihood $\ell_{\text{nl}}(\mathbf{w}, \mathbf{x}, y) = \frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y - f_{\text{lin}}(\mathbf{x}; \mathbf{w}))^2$. We also make the following modeling choices: 1) Prior over weights: $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_\pi, \sigma_\pi^2 \mathbf{I})$. 2) Gradients as Gaussian mixture: $\nabla_{\mathbf{w}} f(\mathbf{x}; \mathbf{w}_\rho) \sim \sum_{i=1}^k \phi_i \mathcal{N}(\boldsymbol{\mu}_i, \sigma_{\mathbf{x}i}^2 \mathbf{I})$. Note that this assumption should be plausible for trained neural networks, in that previous works have shown that per sample gradients with respect to the weights, at \mathbf{w}_ρ , are clusterable (Zancato et al., 2020). We consider that a Gaussian Mixture model for these clusters is a crude but reasonable modeling choice. 3) Labeling function: $y = f(\mathbf{x}; \mathbf{w}_\rho) + \nabla_{\mathbf{w}} f(\mathbf{x}; \mathbf{w}_\rho)^\top (\mathbf{w}_* - \mathbf{w}_\rho) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. We also assume that we have a deterministic estimate of the posterior weights \mathbf{w}_ρ which we keep fixed, and we model the posterior as $\hat{\rho} = \mathcal{N}(\mathbf{w}_\rho, \sigma_\rho^2(\lambda) \mathbf{I})$, similarly to our experimental section. Therefore estimating the posterior corresponds to estimating the variance $\sigma_\rho^2(\lambda)$.

Proposition 3 ($\mathcal{B}_{\text{approximate}}$) *With the above modeling choices, given a distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$, real numbers $\delta \in (0, 1]$, $\lambda \in (0, \frac{1}{c})$ with $c = 2n\sigma_{\mathbf{x}}^2\sigma_\pi^2$, we have with probability at least*

$1 - \delta$ over $(X, Y) \sim \mathcal{D}^n$

$$\begin{aligned}
 2\mathbf{E}_{\mathbf{w} \sim \hat{\rho}} \mathcal{L}_{\mathcal{D}}^{\text{nl}}(\mathbf{w}) &\leq \frac{\|\mathbf{y} - f(\mathbf{X}; \mathbf{w}_{\hat{\rho}})\|_2^2}{n \sigma^2} + \ln(2\pi \sigma^2) + \sigma_{\epsilon}^2 \\
 &+ \frac{h}{2n \sigma^2} \left(\frac{\lambda h}{d \sigma^2} + \frac{1}{\sigma_{\pi}^2} \right)^{-1} + \frac{2\sigma_{\mathbf{x}}^2(\sigma_{\pi}^2 d + \|\mathbf{w}_{*}\|_2^2)}{1 - 2\lambda n \sigma_{\mathbf{x}}^2 \sigma_{\pi}^2} \\
 &+ \frac{1}{\lambda n} \left[\frac{d}{\sigma_{\pi}^2} \left(\frac{\lambda h}{d \sigma^2} + \frac{1}{2\sigma_{\pi}^2} \right)^{-1} + \frac{\|\mathbf{w}_{\hat{\rho}} - \mathbf{w}_{\pi}\|_2^2}{\sigma_{\pi}^2} - d \right. \\
 &\left. + d \ln \left(\frac{\lambda h}{d \sigma^2} + \frac{1}{\sigma_{\pi}^2} \right) + d \ln \sigma_{\pi}^2 + 2 \ln \frac{1}{\delta} \right]
 \end{aligned}$$

where $h = \sum_i \sum_j (\nabla_{\mathbf{w}} f(\mathbf{x}_i; \mathbf{w}_{\hat{\rho}})_j)^2$ is a curvature parameter, $\sigma_{\mathbf{x}}^2 = \sum_{j=1}^k \phi_j \sigma_{\mathbf{x}_j}^2$ is the posterior gradient variance, and σ^2 is the variance of the likelihood function.

Given the Gaussian likelihood function, previous works (Bachmann et al., 2022; Nabarro et al., 2022; Aitchison, 2021) have identified σ^2 (our estimate of the aleatoric uncertainty of the data) as a key source of misspecification, which the temperature λ attempts to fix. Under the PAC-Bayesian modeling of the risk our conclusions are different. Contrary to this prior work our bound suggests that λ cannot be seen as simply fixing a misspecified likelihood variance σ^2 or prior variance σ_{π}^2 . In particular it does not simply rescale the aforementioned quantities. Tempering is the result of complex interactions between various parameters resulting from 1) the likelihood through σ^2 , 2) the prior π through \mathbf{w}_{π} and σ_{π}^2 , 3) data through \mathbf{w}_{*} and $\sigma_{\mathbf{x}}^2$, 4) as well as various parameters such as the curvature of the minimum h and the MAP estimate $\mathbf{w}_{\hat{\rho}}$ that depend on the deep neural network architecture, the optimization procedure and the data. Furthermore, our bound implies that *even for fixed prior, likelihood, and data*, the same λ can imply different test risks based on each MAP estimates properties such as the curvature h and the distance from initialization $\|\mathbf{w}_{\hat{\rho}} - \mathbf{w}_{\pi}\|_2^2$.

We can observe this in our empirical data. We first make 10 MAP estimates for the CIFAR-10, CIFAR-100, SVHN and FMNIST datasets and take care such that the training and testing errors of the MAP are similar. We fit for each MAP a Laplace approximation using the Isotropic approximation to the posterior. For a fixed value of λ we compute the test 0-1 loss of all the Laplace approximations at the different MAP estimates. We then compute the standard deviation of the test 0-1 loss for this fixed value of λ . We keep the prior variance fixed $\sigma_{\pi}^2 = 0.1$. We combine the resulting differences for all fixed values of $\lambda \in [10^{-7}, 10^4]$ in Figure 3. We see that the standard deviation is large for some fixed values of λ . *As both the prior, the temperature λ and the dataset are kept fixed*, the variability can only be explained in terms of the different posteriors induced by the different MAP estimates. In particular we identify the curvature at the minimum h as important, given that we've taken care such that the training and testing accuracy of the different MAP estimates is similar. Our results imply that the temperature parameter λ cannot simply be seen as fixing our wrong initial assumption about the aleatoric uncertainty.

Assume that different MAP estimates $\mathbf{w}_{\hat{\rho}}$ are generated from some randomized algorithm $\mathbf{w}_{\hat{\rho}} \sim \mathcal{P}(X, Y)$ where X, Y are the training data. Our results furthermore imply that it might be difficult to find a single value of λ that gives a *target test loss*, simply because of the stochasticity of $\mathbf{w}_{\hat{\rho}} \sim \mathcal{P}(X, Y)$. This is especially interesting for neural networks because inference is always approximate, and the loss is multimodal.

λ does not simply fix a misspecified estimate of aleatoric uncertainty. In our case, λ also varies due to the specific MAP estimate used for Laplace approximation. More generally, λ is influenced by the fact that inference is approximate and stochastic.

7. Discussion

We argued that using Bayesian models to target Frequentist metrics provides a simple explanation for needing a temperature in the optimization objective. We hope that this will motivate Bayesian practitioners to use heuristics “guilt-free” when targeting Frequentist performance metrics, or target posterior contraction instead. We showcased the importance of choosing a metric when discussing tempering, and our empirical results on the tradeoff between metrics point towards the need for the use of Pareto curves when evaluating Bayesian approaches. Finally our results on the effect of the posterior on tempering might imply that some fine-tuning of λ might be inevitable given the approximate and stochastic nature of inference in neural networks.

Acknowledgments

This work was granted access to the HPC resources of IDRIS under the allocation AD011013876 made by GENCI.

References

- Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. In *International Conference on Learning Representations*, 2021.
- Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1): 8374–8414, 2016.
- Gregor Bachmann, Lorenzo Noci, and Thomas Hofmann. How Tempering Fixes Data Augmentation in Bayesian Neural Networks. *International Conference on Machine Learning*, 2022.
- Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.
- Luc Bégin, Pascal Germain, François Laviolette, and Jean-François Roy. Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2016.

- Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.
- David Blackwell and Lester Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33(3):882–886, 1962.
- Olivier Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56 of *Monograph Series*. Institute of Mathematical Statistics Lecture Notes, 2007.
- Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.
- Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *Uncertainty in Artificial Intelligence*, 2017.
- Gintare Karolina Dziugaite, Kyle Hsu, Waseem Gharbieh, Gabriel Arpino, and Daniel Roy. On the role of data in PAC-Bayes. In *International Conference on Artificial Intelligence and Statistics*, pages 604–612. PMLR, 2021.
- Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing Systems*, 29, 2016.
- Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.
- Peter Grünwald. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.
- Peter Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2):119–149, 2007.
- Peter Grünwald and Thijs Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.
- Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of Bayesian neural nets via local linearization. In *International Conference on Artificial Intelligence and Statistics*, pages 703–711. PMLR, 2021.
- Sanyam Kapoor, Wesley J Maddox, Pavel Izmailov, and Andrew Gordon Wilson. On uncertainty, tempering, and data augmentation in bayesian classification. *arXiv preprint arXiv:2203.16481*, 2022.
- Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Citeseer*, 2009.
- Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32, 2019.

- Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. Pac-bayes compression bounds so tight that they can explain generalization. *arXiv preprint arXiv:2211.13609*, 2022.
- Wesley Maddox, Shuai Tang, Pablo Moreno, Andrew Gordon Wilson, and Andreas Damianou. Fast adaptation with linearized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 2737–2745. PMLR, 2021.
- James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.
- Andres Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 33:5479–5491, 2020.
- David A McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.
- Warren R Morningstar, Alex Alemi, and Joshua V Dillon. PAC^m-Bayes: Narrowing the empirical risk gap in the misspecified Bayesian regime. In *International Conference on Artificial Intelligence and Statistics*, pages 8270–8298. PMLR, 2022.
- Seth Nabarro, Stoil Ganey, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. Data augmentation in Bayesian neural networks and the cold posterior effect. In *Uncertainty in Artificial Intelligence*, pages 1434–1444. PMLR, 2022.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *arxiv*, 2011.
- Lorenzo Noci, Kevin Roth, Gregor Bachmann, Sebastian Nowozin, and Thomas Hofmann. Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect. *Advances in Neural Information Processing Systems*, 34, 2021.
- Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? *International Conference on Machine Learning*, 2020.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arxiv*, 2017.
- Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.
- Luca Zancato, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Predicting training time without training. *Advances in Neural Information Processing Systems*, 33:6136–6146, 2020.

Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *International Conference on Learning Representations*, 2019.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. *arXiv preprint arXiv:1804.05862*, 2018.