How Tempering Fixes Data Augmentation in Bayesian Neural Networks

Gregor Bachmann [*1] **Lorenzo Noci** [*1] **Thomas Hofmann** [1]

Abstract

While Bayesian neural networks (BNNs) provide a sound and principled alternative to standard neural networks, an artificial sharpening of the posterior usually needs to be applied to reach comparable performance. This is in stark contrast to theory, dictating that given an adequate prior and a well-specified model, the untempered Bayesian posterior should achieve optimal performance. Despite the community's extensive efforts, the observed gains in performance still remain disputed with several plausible causes pointing at its origin. While data augmentation has been empirically recognized as one of the main drivers of this effect, a theoretical account of its role, on the other hand, is largely missing. In this work we identify two interlaced factors concurrently influencing the strength of the cold posterior effect, namely the correlated nature of augmentations and the degree of invariance of the employed model to such transformations. By theoretically analyzing simplified settings, we prove that tempering implicitly reduces the misspecification arising from modeling augmentations as i.i.d. data. The temperature mimics the role of the effective sample size, reflecting the gain in information provided by the augmentations. We corroborate our theoretical findings with extensive empirical evaluations, scaling to realistic BNNs. By relying on the framework of group convolutions, we experiment with models of varying inherent degree of invariance, confirming its hypothesized relationship with the optimal temperature.

[*]Equal contribution [1]Department of Computer Science, ETH Zürich, Zürich, Switzerland. Correspondence to: Gregor Bachmann <gregor.bachmann@inf.ethz.ch>, Lorenzo Noci <lorenzo.noci@inf.ethz.ch>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1. Introduction

Deep learning has lead to tremendous advances in a variety of tasks such as computer vision (He et al., 2016), natural language processing (Devlin et al., 2019) and reinforcement learning (Silver et al., 2016) to name but a few. While such deep models exhibit astonishing predictive power, their black-box nature renders uncertainty estimation very difficult and often leads to over-confident decisions (Nguyen et al., 2015; Szegedy et al., 2013). To overcome this difficulty, Bayesian neural networks (BNNs) have been introduced, combining the functional form of deep models with the framework of Bayesian inference (Graves, 2011b; Blundell et al., 2015; Hernández-Lobato & Adams, 2015). By forming a posterior distribution over the model parameters instead of a point estimate, uncertainty estimations become significantly better calibrated, leading to more informed decisions in safety-critical applications. Moreover, due to the multi-modal nature of neural loss landscapes (Garipov et al., 2018; Draxler et al., 2018), Bayesian models are particularly well-suited as they naturally form an ensemble of models (Wilson, 2020).

Recently however, Wenzel et al. (2020) observed that BNNs evaluated on standard benchmarks yield suboptimal performance, even being outperformed significantly by a simple SGD baseline. They noticed on the other hand that an artificial sharpening of the posterior distribution (so-called *cold* posteriors) leads to very strong performance. The sub-optimality of Bayesian models is worrisome as the Bayesian framework is equipped with theoretical guarantees regarding its optimality (Kolmogorov, 1960; Savage, 1954; Jaynes, 2003). As a consequence, a multitude of works have been put forth, exploring different potential causes of this so-called cold posterior effect (CPE), including the curated nature of standard datasets (Aitchison, 2021), the non-Bayesian nature of data augmentation (Izmailov et al., 2021) as well as possibly poorly chosen priors (Noci et al., 2021; Fortuin et al., 2021). Data augmentation has been observed to play a particularly pronounced role, being instrumental in causing the cold posterior effect in a variety of settings (Izmailov et al., 2021; Noci et al., 2021). Nabarro et al. (2021) develop a formalism to incorporate data augmentation into the model but unfortunately the CPE still persists. In this work we aim to bridge this gap by mathematically analyzing the effect of data augmentation on the

resulting posterior. Inspired by the seminal works on "longitudinal" data (Liang & Zeger, 1986; Laird & Ware, 1982; Ware, 1985), we approach the process of data augmentation in a similar spirit, taking into account the strong statistical dependence between augmented examples which breaks the i.i.d. assumption implicit in most Bayesian inference pipelines. By incorporating the correlation structure into the model, we prove that tempering approximates the correct Bayesian posterior and even matches it in simplified settings. Intuitively, the temperature plays the role of the effective sample size, adjusting for the fact that data augmentation leads to a sample size that lies between the original and the augmented one.

We perform exhaustive experiments to validate our theoretical insights. In particular, by relying on the framework of group convolutions (Cohen & Welling, 2016), we design architectures for which the invariance with respect to certain augmentations is approximately built into the model. In turn, we observe a clear correlation between the degree of model invariance and optimal temperature. Under the view of tempering as adjusting the effective sample size of the augmented dataset, using approximately invariant models decreases the need for cold posteriors, as the effective sample size is close to that of the original dataset.

In summary, our contributions are the following:

- We identify the source of misspecification introduced by data augmentation in a Bayesian context. We show how augmenting samples, in conjunction with the model's invariance, induces highly correlated errors, rendering the implicit i.i.d. assumption underlying standard BNN pipelines incorrect.
- We show in simplified settings that tempering the (wrong) posterior can alleviate the misspecification, resulting in a potential explanation of the CPE.
- We test our theory using group convolutions, confirming our hypothesis and theoretical results on the role of invariance. Furthermore, we show that BNNs with group convolutions outperform the standard convolution counterpart in the considered settings.

2. Background

In this section we introduce the relevant mathematical notation and provide background on Bayesian neural networks as well as the cold posterior effect.

Notation. We study the standard supervised learning setting, where we have a dataset consisting of $n \in \mathbb{N}$ i.i.d. input-target pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ distributed according to some (unknown) data distribution, $(\boldsymbol{x}_i, y_i) \sim \mathcal{D}$. We assume that the inputs come from some possibly high-dimensional space $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ and that the targets are univariate, $y \in \mathbb{R}$. In a classification context, y will for instance denote a binary encoding. All our results however extend to the multivariate target case. Occasionally we will find it useful to summarize inputs and targets into matrices, we will then denote $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$. We consider a family of functions $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$ parameterized by $\boldsymbol{\theta} \in \mathbb{R}^p$, serving as models for the data distribution $\mathcal{D}$.

BNNs. Finding optimal values for $\boldsymbol{\theta}$ is a very challenging task due to the high-dimensional nature of the parameter space and its inherent degeneracy. A Bayesian approach to this problem specifies a *prior* distribution $p(\boldsymbol{\theta})$ over the parameters (possibly incorporating domain knowledge) and leverages the data by means of Bayes' rule, resulting in the *posterior* distribution

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}) p(\boldsymbol{\theta}) \tag{1}$$

Prediction on a test sample $\boldsymbol{x}$ is then performed by marginalizing out the parameters $\boldsymbol{\theta}$ over the posterior distribution:

$$p(y|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y}) = \int p(y|\boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) d\boldsymbol{\theta}. \tag{2}$$

Eq. 2, also called *Bayesian model average* (BMA), naturally exploits the multi-modal landscape by averaging over all models that are compatible with the data, leading to richer explanations. Moreover, having access to the predictive distribution $p(y|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y})$ directly allows for uncertainty estimation.

Inference. In practice however, the integral in Eq. 2 is intractable and as a consequence, approximate inference methods have been designed to estimate it. In particular, in this work we will make use of gradient-based Monte Carlo methods (MCMC), where a finite number of K samples from the posterior are obtained and the BMA is approximated as $p(y|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y}) \approx \frac{1}{K} \sum_{k=1}^K p(y|\boldsymbol{x}, \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$. To that end, we introduce the posterior energy function

$$U(\boldsymbol{\theta}) := -\log\left(p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})\right),$$

along with the discretized Langevin dynamics which govern the parameters' evolution,

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \frac{\alpha_t}{2} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) + \sqrt{\alpha_t} \mathcal{N}(0, \mathbb{1}). \tag{3}$$

In deep learning, we often deal with very big data corpora, rendering the gradient operation $\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})$ intractable. Inspired by the mini-batching operation in SGD, Welling & Teh (2011) introduced the same idea, crucially relying on

the fact that for i.i.d. data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, we can write

$$U(\boldsymbol{\theta}) \overset{i.i.d.}{=} -\sum_{i=1}^{n} \log\left(p(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i)\right) - \log(p(\boldsymbol{\theta})). \tag{4}$$

A noisy estimate of the full gradient is then formed on a mini-batch $S_t \subset \{1, \dots, n\}$ of the data,

$$\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}) \approx -\frac{n}{|S_t|} \nabla_{\boldsymbol{\theta}} \left(\sum_{i \in S_t} \log p(y_i|\boldsymbol{\theta}, \boldsymbol{x}_i) + \log p(\boldsymbol{\theta}) \right).$$

These so-called SG-MCMC methods, in combination with pre-conditioners (Li et al., 2016) and cyclical learning rates (Zhang et al., 2020), offer a powerful and scalable approach to Bayesian inference, as shown in (Wenzel et al., 2020).

Cold Posteriors. Despite their benefits, the adoption of BNNs is not widespread: inference procedures are usually slower and reported to even be outperformed by SGD in certain settings (Wenzel et al., 2020). At the same time however, Wenzel et al. (2020) show that the performance issue can be resolved by re-scaling the posterior with a temperature parameter $T > 0$:

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})^{\frac{1}{T}} \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X})^{\frac{1}{T}} p(\boldsymbol{\theta})^{\frac{1}{T}}. \tag{5}$$

The optimal temperature is found to be consistently smaller than one, i.e. $T << 1$, across several models and datasets. Thus, the term "cold posterior effect" was coined. As discussed in Sec. 1, exactly pin-pointing the origin of this effect is complicated and a number of hypotheses have been put forth in the literature. In this work, we focus on the role of data augmentation since empirically it is observed to be the main driver of the CPE (Noci et al., 2021; Izmailov et al., 2021). Finally, we also consider the variant where only the likelihood is tempered, and not the prior:

$$p_T(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X})^{\frac{1}{T}} p(\boldsymbol{\theta}). \tag{6}$$

We will show both theoretically and experimentally that tempering the likelihood is sufficient to account for the misspecification introduced by data augmentation.

Data Augmentation. A common technique in deep learning to foster invariance of a model is given by data augmentation. Given an example $(\boldsymbol{x}, y) \sim \mathcal{D}$, one produces several (random) augmentations of the input $\boldsymbol{x}$, that by design, should preserve the label information, i.e. $\tilde{y} = y$. More formally, depending on the domain, the practitioner designs a (parametrized) augmentation function $R_{\boldsymbol{\eta}} : \mathbb{R}^d \to \mathbb{R}^d$ that takes an input $\boldsymbol{x}$ and a parameter $\boldsymbol{\eta} \sim p(\boldsymbol{\eta})$ and produces an augmented example $\tilde{\boldsymbol{x}} := R_{\boldsymbol{\eta}}(\boldsymbol{x})$. Usually, $R_{\boldsymbol{\eta}}$ is chosen in a way that the resulting augmentation $\tilde{\boldsymbol{x}}$ preserves the label, i.e. an annotator would assign the same label to $\tilde{\boldsymbol{x}}$ as $\boldsymbol{x}$. In a computer vision context, $R_{\boldsymbol{\eta}}$ would for instance correspond to the composition of randomly rotating, flipping and translating the input image, intuitively leaving the associated label invariant. The same input $\boldsymbol{x}$ is usually augmented several times, leading to a set of augmentations $\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_B$. For instance, in stochastic gradient descent a fresh augmentation of $\boldsymbol{x}$ is produced at every epoch of the optimization. Data augmentation is a standard component of most deep learning pipelines and an almost necessary ingredient for state-of-the-art results. For instance, the top 5 leaders in *ImageNet* accuracy[1] (Dai et al., 2021; Zhai et al., 2021; Pham et al., 2021; Liu et al., 2021; Yuan et al., 2021), all use some form of data augmentation.

Group Convolutions. In order to experimentally verify the predicted relationship between optimal temperature and model invariance, we make use of group-equivariant convolutions (Cohen & Welling, 2016), which extend the translation equivariance property of standard convolutions to richer classes of transformations G that form finite symmetry groups. Given a feature map $\phi(\boldsymbol{x}) \in \mathbb{R}^d$ and a transformation $g : \mathbb{R}^d \to \mathbb{R}^d$ for $g \in G$, we say that $\phi(\boldsymbol{x})$ is equivariant with respect to G if $\exists\ g' \in G$ s.t. $\phi(g(\boldsymbol{x})) = g'(\phi(\boldsymbol{x}))$. Note that invariance is the special case in which g' is the identity function. The mathematical machinery of symmetry groups and equivariance can be combined to design group-equivariant convolutional layers. More concretely, given a feature map ϕ and K filters f_k, where both the filters and feature map are functions on G for all but the first layer, a G-convolutional layer can be defined for $h \in G$, as:

$$[f * \phi](h) = \sum_{g \in G} \sum_{k=1}^{K} f_k(g) \phi(h^{-1}(g)). \tag{7}$$

For the first layer, K is the number of channels of the input image, f_k is the k-th channel of the input image, and the first sum is over its pixel values.

Invariant architectures with respect to these groups can be built by using G-convolutional layers followed by a global average pooling layer (GAP) (Cohen & Welling, 2016; Veeling et al., 2018). However, one has to to be careful when designing such architectures, as some commonly used features break such invariances. In particular, it was shown that strided convolutions (Mouton et al., 2021), pooling layers (subsampling) (Bulusu et al., 2021) and padding (Kayhan & Gemert, 2020) break translational equivariance in standard convolutions.

3. Data Augmentation in Bayesian Models

Correlations. Data augmentation in the context of Bayesian inference is a delicate matter. Consider the three

[1] https://paperswithcode.com/sota/

Figure 1: Illustration of augmentations, the original image $\boldsymbol{x}$ (left), and two random augmentation $R_{\boldsymbol{\eta}_1}(\boldsymbol{x})$ (middle) and $R_{\boldsymbol{\eta}_2}(\boldsymbol{x})$ (right). We apply a composition of random rotations, crops and flips.

images in Fig. 1, where the left image corresponds to the original sample $\boldsymbol{x}$ and the other two images $R_{\boldsymbol{\eta}_1}(\boldsymbol{x}), R_{\boldsymbol{\eta}_2}(\boldsymbol{x})$ are two random augmentations. The resulting label for the augmented samples is still "otter", i.e. $\tilde{y}_1 = \tilde{y}_2 = y$ as the augmentations are chosen in a way to preserve that information. Very intuitively however, the two augmented examples share significant correlation. More concretely, considering $B \in \mathbb{N}$ augmentations of $\boldsymbol{x}$, forming the set $\tilde{\mathcal{S}} := \{(\tilde{\boldsymbol{x}}_1, y), \dots, (\tilde{\boldsymbol{x}}_B, y)\}$, it is evident that the set exhibits significant correlation, leading to an effective sample size that lies between 1 and B. However, the inference methods employed in the BNN literature implicitly assume i.i.d. data (Welling & Teh, 2011; Graves, 2011a; Blundell et al., 2015; Wenzel et al., 2020), (essentially, boiling down to Eq. 4), leading to a misspecified model when used in conjunction with data augmentation.

A Simple Example. We illustrate the arising correlations and the benefits of tempering through a very simple example. We study the classic textbook problem, where we aim to estimate the mean of a given set of samples $\{x_1, \dots, x_n\} \subset \mathbb{R}$. We will show in the next section how the intuition translates to a regression setting. We make the modeling assumption

$$x|\mu \sim \mathcal{N}(\mu, \sigma^2),$$

where $\sigma > 0$ is known but $\mu \in \mathbb{R}$ is unknown. Moreover, we set the prior $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ for $\sigma_0 > 0$ and $\mu_0 \in \mathbb{R}$. Given a sample x, we can augment it as $\tilde{x} = R_\eta(x) = x + \eta$ where $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$ is independent from x, leaving the distribution invariant as evidently, $\tilde{x} \sim \mathcal{N}(\mu, \sigma^2 + \sigma_\eta^2)$. On the other hand, a correlation structure among augmented samples emerges, as for two augmentations $\tilde{x}_i = x + \eta_i$ and $\tilde{x}_j = x + \eta_j$ with independent η_i, η_j, it holds

$$\operatorname{cov}(\tilde{x}_i, \tilde{x}_j) = \sigma^2.$$

If we consider $B \in \mathbb{N}$ augmentations per sample x_i and collect them into a vector $\tilde{\boldsymbol{x}} \in \mathbb{R}^{Bn}$, it holds

$$\tilde{\boldsymbol{x}} \sim \mathcal{N}(\mu \mathbf{1}_{Bn}, \tilde{\boldsymbol{\Sigma}}),$$

where $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{Bn \times Bn}$ is block-diagonal with blocks $\boldsymbol{\Sigma} = \sigma_\eta^2 \mathbb{1} + \sigma^2 \mathbf{1}_B \mathbf{1}_B^T \in \mathbb{R}^{B \times B}$ and $\mathbf{1}_m \in \mathbb{R}^m$ denotes the all-ones vector. On the other hand, we can choose to ignore the correlation structure, treating all the samples as i.i.d. realizations, leading to the diagonal covariance $(\sigma^2 + \sigma_\eta^2)\mathbb{1}$. Denote by $p(\mu|\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{\Sigma}})$ the posterior incorporating correlation and by $p_T(\mu|\tilde{\boldsymbol{x}}; \mathbb{1})$ the tempered posterior under the i.i.d. assumption. We can prove the following:

Theorem 3.1. *$p_T(\mu|\tilde{\boldsymbol{x}}; \mathbb{1})$ and $p(\mu|\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{\Sigma}})$ exactly match for the choice of temperature*

$$T^*(\sigma_\eta; B) = \frac{\sigma_\eta^2 + B\sigma^2}{\sigma_\eta^2 + \sigma^2}.$$

We postpone the proof of Thm. 3.1 to Appendix A.1. Thm. 3.1 shows that tempering completely fixes the misspecification implied by treating augmentations as i.i.d. samples! Let us comment on a few characteristics of the ideal temperature $T^*(\sigma_\eta; B)$. First, it holds that $T^* \geq 1$, i.e. we always require a hot posterior. Moreover, T^* is an increasing function in B, i.e. the more you augment, the hotter the posterior needs to become. This is intuitive, as we should rely less on the data since the sample size is artificially inflated. If a single augmentation is used, i.e. $B = 1$ we recover $T^* = 1$ as expected, since data points indeed become independent. Finally T^* is a decreasing function in σ_η^2, when augmentations become less diverse, i.e. $\sigma_\eta \to 0$, we converge to $T^* = B$. Data augmentation is usually associated with cold posteriors in the literature, which is in stark contrast to our result. We explain this discrepancy in detail in Sec. 4.

Regression Setting. Let us illustrate how a regression setting changes the implied model. Consider the data generating process

$$y = f_{\boldsymbol{\theta}_*}(\boldsymbol{x}) + \epsilon,$$

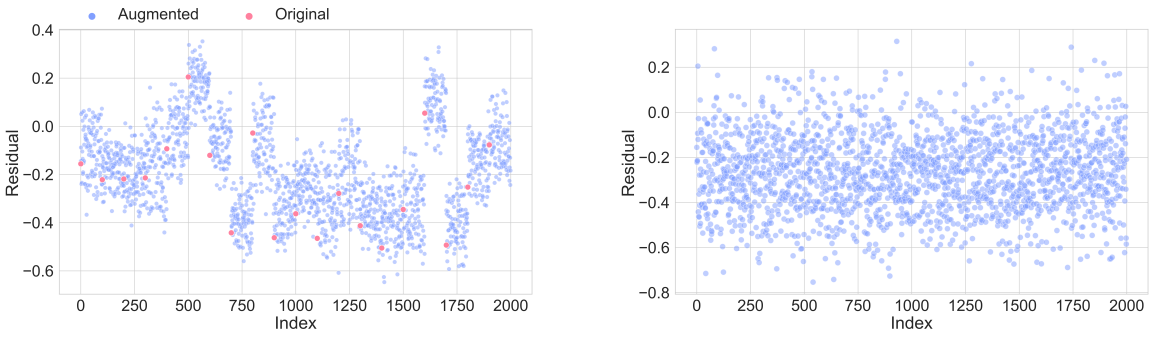
where $f_{\boldsymbol{\theta}}$ defines a family of functions, parametrized by $\boldsymbol{\theta}$ with true but unknown configuration $\boldsymbol{\theta}_* \in \mathbb{R}^p$. Moreover, $\boldsymbol{x} \in \mathbb{R}^d$ denotes the covariate and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the error, inherent to the process. We place a prior $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ on the parameters and assume that we have n independent realizations of the process, $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$. We consider general augmentations $R_{\boldsymbol{\eta}}(\boldsymbol{x})$ where $\boldsymbol{\eta} \sim p(\boldsymbol{\eta})$ governs the randomness. Similarly as in practice, since we cannot access $\boldsymbol{\theta}_*$, we postulate that the augmentation $\tilde{\boldsymbol{x}} := R_{\boldsymbol{\eta}}(\boldsymbol{x})$ shares the same response value, i.e. we form the sample $(\tilde{\boldsymbol{x}}, y)$. Imposing a response induces an error $\tilde{\epsilon}$ which we can calculate, due to the relation

$$y = f_{\boldsymbol{\theta}_*}(\boldsymbol{x}) + \epsilon \overset{!}{=} f_{\boldsymbol{\theta}_*}(R_{\boldsymbol{\eta}}(\boldsymbol{x})) + \tilde{\epsilon},$$

which after re-arranging, leads to the following:

$$\tilde{\epsilon} = \epsilon + \delta_{\boldsymbol{\eta}},$$

where we define $\delta_{\boldsymbol{\eta}} := f_{\boldsymbol{\theta}_*}(\boldsymbol{x}) - f_{\boldsymbol{\theta}_*}(R_{\boldsymbol{\eta}}(\boldsymbol{x}))$. Intuitively, $\delta_{\boldsymbol{\eta}}$ measures the degree of invariance of the true model



(a) Augmented Residuals

(b) Unaugmented Residuals

Figure 2: **Residual vs. Order Plot:** Residuals of an untrained *ResNet18* on *Dogs vs Cats* dataset. (a) 20 independent samples (red dots) of label "cat" with 100 augmentations each. (b) 2000 independent samples of label "cat". Augmented residuals display a strong dependence (forming clusters) while unaugmented residuals show no trend.

$\boldsymbol{\theta}_*$. We make the natural assumption that $\mathbb{E}[\delta_{\boldsymbol{\eta}}] = 0$, implying that $\tilde{\epsilon}$ remains a centered random variable as the augmentation is not inducing any bias. In practice, the same data point $\boldsymbol{x}$ is augmented multiple times, i.e. we produce $R_{\boldsymbol{\eta}_i}(\boldsymbol{x})$ where $\boldsymbol{\eta}_i \overset{i.i.d.}{\sim} p(\boldsymbol{\eta})$ for $i = 1, \dots, B$, which we associate with the same response y. This induces a series of errors $\tilde{\epsilon}_i$, which give rise to a correlation structure

$$\operatorname{cor}(\tilde{\epsilon_i}, \tilde{\epsilon_j}) = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \operatorname{var}(\delta_{\boldsymbol{\eta}})}. \tag{8}$$

We detail the calculation in Appendix A.2. This structure depends on the original error variance σ_ϵ^2 and $\operatorname{var}(\delta_{\boldsymbol{\eta}})$, which in turn depends on the augmentation scheme R and the true model $f_{\boldsymbol{\theta}_*}$. As a consequence, the likelihood **cannot** be factorized, i.e.

$$p(\tilde{y}_1, \dots, \tilde{y}_B | \boldsymbol{x}, \boldsymbol{\theta}) \neq \prod_{i=1}^{B} p(\tilde{y}_i | \boldsymbol{x}, \boldsymbol{\theta}).$$

We empirically demonstrate the correlated nature of the errors in Fig. 2 by resorting to the classic tool of the *Residual vs. Order* plot, often employed in least squares. We approximate the true (but unobservable) errors by the resulting residuals of a random *ResNet18* for both augmented and non-augmented data. In the augmentation case, we observe a strong correlation when the residuals are plotted against the sample index, while without augmentations, they exhibit no structure. Notice that already a random (i.e. untrained) model displays this pattern, hinting at the fact that such invariances (and hence correlations) are built into the architecture.

Finally our results also extend to the classification setting, we refer the reader to Appendix A.3.

Tempering as Effective Sample Size. On the contrary, if errors become perfectly correlated, the likelihood even degenerates,

$$p(\tilde{y}_1, \dots, \tilde{y}_B | \boldsymbol{x}, \boldsymbol{\theta}) = p(\tilde{y} | \boldsymbol{x}, \boldsymbol{\theta})$$

We see that **wrongly** factorizing the likelihood in this case can be fixed through tempering with $T > 0$,

$$\begin{aligned} p(\tilde{y}_1, \dots, \tilde{y}_B | \boldsymbol{x}, \boldsymbol{\theta})^{\frac{1}{T}} &\overset{\text{wrong}}{=} \prod_{i=1}^{B} p(\tilde{y}_i | \boldsymbol{x}, \boldsymbol{\theta})^{\frac{1}{T}} \\ &= p(\tilde{y} | \boldsymbol{x}, \boldsymbol{\theta})^{\frac{B}{T}} \end{aligned}$$

i.e. setting $T = B$ recovers the correct likelihood. Interpreting Eq. 8 as a measure of "factorizability" of the likelihood, we conjecture that $T^* \in [1, B]$, depending on the degree of invariance of the model $\boldsymbol{\theta}_*$. In this sense, $\frac{B}{T}$ measures the "effective sample size" of augmentations. As a consequence, there should ideally be a separate temperature T_i for every datapoint $\boldsymbol{x}_i$. However, we conjecture that in realistic settings, augmentations of different datapoints should exhibit similar correlations, hence a single global temperature T provides a good approximation. For intermediate values of correlations, tempering can reduce but not perfectly fix the misspecification. We refer to Appendix A.5 for details.

Linear Regression. We can illustrate the general framework with the simpler case of linear regression, i.e. $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x}$ under additive augmentations $R_{\boldsymbol{\eta}}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$. We can write

$$\delta_{\boldsymbol{\eta}} = \boldsymbol{\theta}_*^T \boldsymbol{x} - \boldsymbol{\theta}_*^T R_{\boldsymbol{\eta}}(\boldsymbol{x}) = -\boldsymbol{\eta}^T \boldsymbol{\theta}_*$$

Here, we can directly see that $\mathbb{E}[\delta_{\boldsymbol{\eta}}] = 0$. Moreover, the correlation of the errors can be computed in closed form as well (detailed in Appendix A.4),

$$\operatorname{cor}(\tilde{\epsilon_i}, \tilde{\epsilon_j}) = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \boldsymbol{\theta}_*^T \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \boldsymbol{\theta}_*}$$

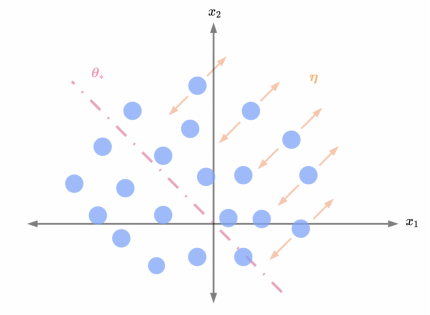


Figure 3: Illustration of additive augmentations $\boldsymbol{\eta}$ (orange) constrained to a line with an invariant model $\boldsymbol{\theta}_*$ (pink).

We notice the following: If the true model $\boldsymbol{\theta}_*$ is contained in the null space of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, the errors exhibit perfect correlation, i.e. $\mathrm{cor}(\tilde{\epsilon_i}, \tilde{\epsilon_j}) = 1$. This is very intuitive and illustrated in Fig. 3 for $d = 2$. If $\boldsymbol{\eta}$ is degenerate and only operates in a subspace, here along the $45°$-axis e.g. $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \begin{pmatrix} \sigma_\eta^2 & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 \end{pmatrix}$, any orthogonal model $\boldsymbol{\theta}_*$ will be invariant to the augmentation, leading to perfectly correlated errors. On the other hand, once the augmentation noise σ_η^2 overwhelms the signal (i.e. $\sigma_\eta^2 \to \infty$) we recover independent (but useless) samples.

Conditioning or Not Conditioning. In the previous section we have assumed, by conditioning on $\boldsymbol{x}$, that the model actually has access to $\boldsymbol{x}$. However, this is not what is done in practice, as the model makes inference relying only on the augmentations and not on the underlying datapoint. Now we derive a general generative model for data augmentation that does not rely on conditioning on the original datapoints $\boldsymbol{x}$, but only on the augmentations, and discuss its implications.

Let $p(\tilde{\boldsymbol{x}}_i|\boldsymbol{x})$ be the likelihood of the augmentation $\tilde{\boldsymbol{x}}_i$ being generated from $\boldsymbol{x}$ through $R_{\boldsymbol{\eta}}(\boldsymbol{x})$. If we do not condition on $\boldsymbol{x}$, the likelihood for the augmentations $\tilde{\mathcal{S}}$ has the following form:

$$p(\tilde{\mathcal{S}}|\boldsymbol{\theta}) = \int p(\{(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{y}}_i)\}_{i=1}^B, \boldsymbol{x}, |\boldsymbol{\theta}) d\boldsymbol{x}$$
$$= \int p(\{\tilde{\boldsymbol{y}}_i\}_{i=1}^B|\boldsymbol{x}, \{\tilde{\boldsymbol{x}}_i\}_{i=1}^B, \boldsymbol{\theta}) \prod_{i=1}^B p(\tilde{\boldsymbol{x}}_i|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}.$$

Note that this intractable setting is what is commonly done in practice, as the model does not rely on $\boldsymbol{x}$, but only on the augmentations to estimate the posterior. This means that we have to marginalize $\boldsymbol{x}$ out in order to account for all possible original datapoints that might have generated $\tilde{\boldsymbol{x}}$. However, more realistically, it is safe to assume that an augmentation $\tilde{\boldsymbol{x}}$ is "sufficiently close" to $\boldsymbol{x}$ that no other $\boldsymbol{x}'$ might have generated it. Therefore conditioning on $\boldsymbol{x}$ does not affect the parameter estimation. Clearly, this final reasoning assumes that the augmentations are not "too wild". In particular, standard augmentations adopted in computer vision, such as rotations, horizontal/vertical flips and crops certainly satisfy this condition.

4. Experiments

To perform approximate inference for BNNs, we use the SG-MCMC sampler described in (Wenzel et al., 2020), which includes cyclical step size (Zhang et al., 2020) and layer-wise preconditioning (we refer the reader to Appendix B for the implementation details).

Hot or Cold Posteriors? To make the approximate inference method match the considered theoretical setting, we adjust the sample size from n to Bn where B is the number of augmentations used for inference, which coincides with the number of training epochs. In order to only temper the likelihood, we apply a simple re-parameterization to the learning rate $\gamma_t = T\alpha_t$ and remove the temperature term from the prior, resulting in the overall update:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \frac{\gamma_t}{2}\left(\frac{Bn}{T|S_t|}\sum_{i\in S_t}\log p(y_i|\boldsymbol{x}_i) + \log p(\boldsymbol{\theta}_t)\right) + \sqrt{\gamma_t}\mathcal{N}(0, \mathbb{1}), \quad (9)$$

From Eq. 9, the role of T in data augmentation is manifest: it rescales the augmented dataset size Bn to adjust for the non i.i.d data. Inspired by our theoretical results, we conjecture that $T^* \in [1, B]$, resulting thus in a *hot* posterior. On the contrary, if one does not explicitly account for the augmentations and considers only n datapoints, then $T' = \frac{T^*}{B}$ is optimal, and *cold* posteriors are obtained instead. In other words, hot and cold posteriors are two sides of the same coin in our setting, they are simply a consequence of the normalization used in SG-MCMC. In our experiment, we will adopt the parametrization of Eq. 9 to remain consistent with our theoretical results. Therefore hot posteriors are to be expected, and $T = B$ is optimal when either the model is invariant with respect to the augmentations, or data augmentation is switched off. In all the following plots, $T = B$ will be highlighted by a vertical dashed line.

As a first experiment, we test whether tempering only the likelihood qualitatively changes the temperature landscape, compared to tempering the posterior. We display the results in Fig. 4 for a *ResNet20* when data augmentation is used (random flips and crops). We observe that the same optimal temperature is achieved by both approaches, strongly suggesting that the likelihood is at the core of the CPE for data augmentation, and not the prior over the parameters, which was shown to play a significant role for small sample sizes (Noci et al., 2021). Moreover, we indeed observe hot posteriors (i.e. $T >> 1$) when employing the update in

Eq. 9, as predicted by our results.

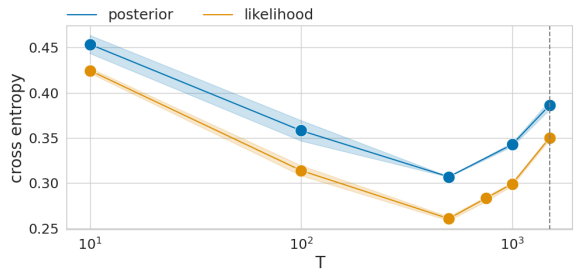


Figure 4: Test cross-entropy as a function of the temperature T for posterior tempering (blue) and likelihood tempering (orange).

G-Convolutions Alleviate the CPE. As detailed in Section 2, the adoption of G-convolutions is not sufficient to obtain an invariant architecture, as standard operations like striding or padding break the invariance. This is concerning, as strided convolutions are heavily used in state-of-the arts architectures like ResNets to enhance performance and reduce the parameter count. In our experiments, we use a *ResNet20* with 2-strided G-convolutions, thus losing its exact invariance but remaining largely insensitive to the group transformations. We will refer to it in the following as *G-ResNet*. In particular, we will use *p4m* equivariant convolutions, enforcing equivariance with respect to compositions of translations, horizontal and vertical flip as well as rotations by multiples of 90°. Finally, note that *p4m* convolutions effectively make the size of the feature maps 8 times larger. Therefore, to have a fair comparison between standard and group convolutions, we follow Cohen & Welling (2016) and reduce the number of filters of G-convolutional layers by $\sqrt{8}$ to roughly have the same number of training parameters.

The results of this comparison are shown in Fig. 5. First, we use only random crops and horizontal flips to augment the dataset (solid lines). Then we repeat the experiment with the additions of multiples of 90° rotations (dashed lines). *G-ResNet* outperforms *ResNet20* at the optimal temperature in all cases, while 90° rotations seems to degrade performance across both architectures. Note when using only flips and crops (solid lines), the optimal temperature is similar. However, when rotations are added, the optimal temperature of *ResNet20* decreases significantly, while for *G-ResNet* - which is more insensitive to such rotations - there is almost no shift in the optimal temperature. In particular, we report a significant temperature shift in *ResNet20* while almost no shift for *G-ResNet*, confirming our hypothesis on the role of invariance. We stress that the *G-ResNet* is not invariant, but only *more* invariant (insensitive), due to the usage of strides and random crops. Therefore we should not expect $T = B$ to be optimal in this case.

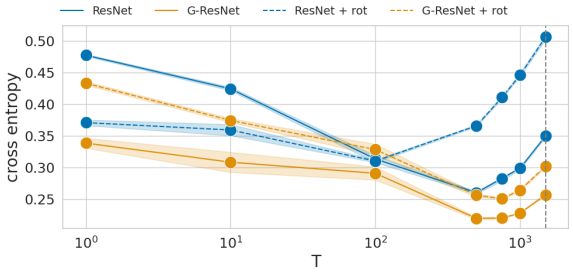


Figure 5: Test cross-entropy as a function of the temperature T for a ResNet with standard convolutions (blue) and G-convolutions (orange), with or without extra 90°rotations (dashed lines).

An Almost Invariant Model. Next, we test whether an almost invariant model (*insensitive*) can achieve an optimal temperature of $T = B$ while employing data augmentation. In our case, such a model can be built by **not** using strides (i.e. stride=1), and use only flips and multiples of 90° rotations as augmentations. We fix the maximum number of augmentations to $B = 150$, and use a burn-in period for SG-MCMC of 150 epochs. In this way, we do not start sampling before the model has visited all the datapoints. We display the results in Fig. 6. As predicted, difference in performance at the optimal temperature and at $T = B = 150$ is minimal. Further evidence for our argument is given by the fact that 2-strided convolutions (i.e. a less invariant model) produce a weaker hot posterior effect. The invariance can be easily destroyed by adding an extra random 10° rotation to the set of transformations, as shown in Fig. 7. In that case we move away from $T = 150$, regardless of the stride.

Finally, to have a quantitative measure of the degree of invariance, we plot the absolute value of the difference of the output probabilities of the model (i.e the total variation) among the augmentations used for training. We use the models with optimal temperatures determined in the previous experiments (Fig. 6-7). Results are shown in Fig. 8. As expected, we find that the total variation monotonically increases with the degree of invariance of model, i.e. the most invariant model (1-strided, no 10° rotations) displays the smallest variation.

5. Related Work

Correlations. Incorporating dependence between samples has a long tradition in machine learning. Correlations emerge naturally in a range of statistical applications, including longitudinal data (Liang & Zeger, 1986), time series (Hamilton, 2020) and clustering (Wakefield, 2013), to name but a few. The most relevant setting to our work is longitu-

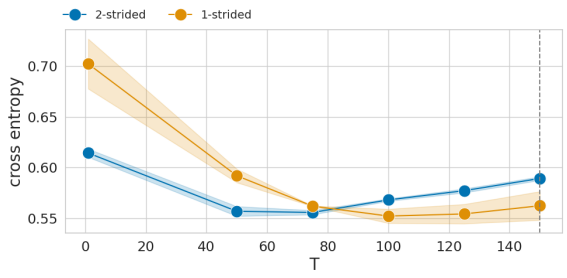


Figure 6: Test cross-entropy as a function of the temperature for 2-strided (blue) and 1-strided (orange) G-convolutional network. Note how the 1-strided model presents a very flat curve toward $T = 150$, indicating its insensitivity with respect to the augmentations.

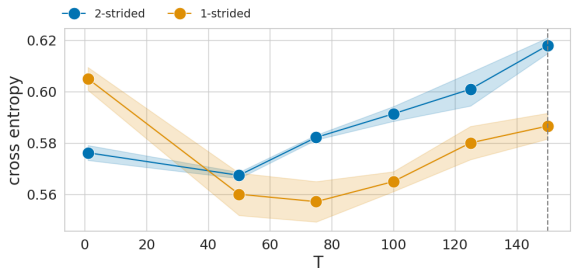


Figure 7: Same setting as in Fig. 6, but this time we perform an extra random rotation of 10°. Note how the loss of invariance shifts the optimal temperature significantly, in both the 2-strided and 1-strided case.

dinal modeling, where multiple measurements are produced from a single source point, a very common scenario in the field of biostatistics. Quantifying the predictive relationship between blood pressure and Diabetes (binary response) is a standard example. A typical dataset consists of n patients, where crucially, for each patient i, repeated blood pressure measurements are performed due to the inherently noisy measurement process. Naturally, these repeated measurements are far from independent. Neglecting these correlations can be detrimental and as a consequence, various approaches have been developed, dating back to fLiang & Zeger (1986); Ware (1985); Laird & Ware (1982). We refer to Verbeke & Molenberghs (2005) for an overview. Many of those approaches build upon the well-known generalized least squares method or the linear mixed model (Robinson, 1991).

Data Augmentation and CPE. While some works have explored data augmentation from a theoretical angle for standard neural networks (Chen et al., 2020; Dao et al., 2019; Wu et al., 2020), to the best of our knowledge only Nabarro et al. (2021) have explored it in the context of the cold posterior effect. They construct an invariant model by marginalizing over the augmented data distribution through averaging the predicted logits/probabilities of the model. They are however not able to explain the cold posterior effect. We argue that this is due to the fact that the augmentations are considered as latent variables that generated the observed (un-augmented) datapoints. Orthogonally to their approach, in this work we argue for the more realistic case in which the augmentations are generated from the original datapoints, and cold/hot posteriors arise from not taking into account the correlations between the errors that this process inevitably generates. This way, infinitely many augmentations can be considered without overwhelming the prior, in contrast to the argument in Nabarro et al. (2021).

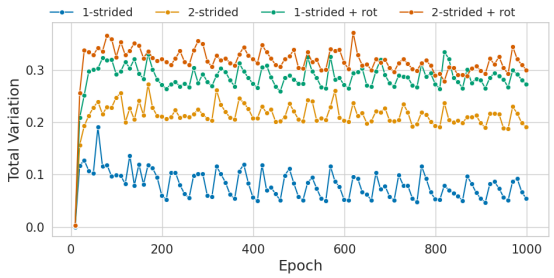


Figure 8: Total variation during learning for the experiments with G-convolutional networks. A lower value of the total variation indicates a greater degree of invariance.

Model Misspecification and Tempering. Our work is inspired by a remarkable series of works on the role of tempering under model misspecification (Grünwald, 2011; 2012; Grünwald et al., 2017). In particular, in Grünwald et al. (2017) it is shown that wrongly modeling heteroscedastic noise errors as homoscedastic, makes Bayesian inference fail, while tempering fixes this model misspecification. In our case, the model misspecification is caused by not modeling the strong correlations arising from the data augmentation process. Finally, although relatively new, the CPE has been analyzed under other perspectives, and its flourishing literature includes other works such as (Zeno et al., 2020; Adlam et al., 2020; Laves et al., 2021).

6. Conclusion

In this work, we showed how data augmentation, in conjunction with invariance, introduces correlations between errors, leading to misspecified models. We demonstrated how tempering can reduce this misspecification by approximating the correct posterior, offering a possible explanation for the CPE. Tempering is thus more principled from a Bayesian perspective than previously assumed. We also identified G-convolutions to be a viable tool for the design of BNNs,

enhancing their invariance and leading to better priors. Our theoretical and empirical results suggest several avenues towards combatting the CPE and improving BNNs in general. From Figure 8, it is manifest that the average invariance, as measured by the total variation, is determined during the burn-in epochs. Therefore, we foresee as a promising future direction to form an estimator of the ideal temperature that could be computed from the data during the burn-in period. Crucially, such an estimator could remove the need for expensive grid searches. Finally, the superior performances achieved by the models employing G-convolutions further motivates the developments of more informed priors over functions that have the desired invariances incorporated into the model. Finally, we want to highlight that the CPE has been shown to arise from other causes as well, which do not involve data augmentation (Fortuin et al., 2021; Noci et al., 2021). It remains exciting future work to complete the understanding of the CPE in these settings.

References

Adlam, B., Snoek, J., and Smith, S. L. Cold posteriors and aleatoric uncertainty. *arXiv preprint arXiv:2008.00029*, 2020.

Aitchison, L. A statistical theory of cold posteriors in deep neural networks. In *International Conference on Learning Representations*, 2021.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pp. 1613–1622, 2015.

Bulusu, S., Favoni, M., Ipp, A., Müller, D. I., and Schuh, D. Generalization capabilities of translationally equivariant neural networks. *arXiv preprint arXiv:2103.14686*, 2021.

Chen, S., Dobriban, E., and Lee, J. A group-theoretic framework for data augmentation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21321–21333. Curran Associates, Inc., 2020.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999, 2016.

Dai, Z., Liu, H., Le, Q. V., and Tan, M. Coatnet: Marrying convolution and attention for all data sizes. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., and Re, C. A kernel theory of modern data augmentation. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1528–1537. PMLR, 09–15 Jun 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 2019.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. Essentially no barriers in neural network energy landscape. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1308–1317. PMLR, 2018.

Fortuin, V., Garriga-Alonso, A., Wenzel, F., Rätsch, G., Turner, R., van der Wilk, M., and Aitchison, L. Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*, 2021.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Graves, A. Practical variational inference for neural networks. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011a. URL https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

Graves, A. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pp. 2348–2356, 2011b.

Grünwald, P. Safe learning: bridging the gap between bayes, mdl and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 397–420. JMLR Workshop and Conference Proceedings, 2011.

Grünwald, P. The safe bayesian. In *International Conference on Algorithmic Learning Theory*, pp. 169–183. Springer, 2012.

Grünwald, P., Van Ommen, T., et al. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4): 1069–1103, 2017.

Hamilton, J. D. *Time Series Analysis*. Princeton University Press, 2020. ISBN 9780691218632. doi: 10.1515/9780691218632.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Hernández-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, pp. 1861–1869, 2015.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. What are bayesian neural network posteriors really like? *Proceedings of the 38th international conference on machine learning (ICML)*, 2021.

Jaynes, E. T. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. doi: 10.1017/CBO9780511790423.

Kayhan, O. S. and Gemert, J. C. v. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14274–14285, 2020.

Kolmogorov, A. N. *Foundations of the Theory of Probability*. Chelsea Pub Co, June 1960.

Laird, N. M. and Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, 38 4:963–74, 1982.

Laves, M.-H., Tölle, M., Schlaefer, A., and Engelhardt, S. Posterior temperature optimization in variational inference for inverse problems. *arXiv preprint arXiv:2106.07533*, 2021.

Li, C., Chen, C., Carlson, D., and Carin, L. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Liang, K.-Y. and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 04 1986. ISSN 0006-3444.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. Swin transformer v2: Scaling up capacity and resolution, 2021.

Mouton, C., Myburgh, J. C., and Davel, M. H. Stride and translation invariance in cnns. In *Southern African Conference for Artificial Intelligence Research*, pp. 267–281. Springer, 2021.

Nabarro, S., Ganev, S., Garriga-Alonso, A., Fortuin, V., van der Wilk, M., and Aitchison, L. Data augmentation in bayesian neural networks and the cold posterior effect, 2021.

Nguyen, A. M., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015.

Noci, L., Roth, K., Bachmann, G., Nowozin, S., and Hofmann, T. Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect. In *Advances in Neural Information Processing Systems*, 2021.

Pham, H., Dai, Z., Xie, Q., Luong, M.-T., and Le, Q. V. Meta pseudo labels. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Robinson, G. That blup is a good thing: The estimation of random effects. *Statistical Science*, 6, 02 1991. doi: 10.1214/ss/1177011926.

Savage, L. J. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. doi: 10.1038/nature16961.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

Verbeke, G. and Molenberghs, G. *Linear Mixed Models For Longitudinal Data*. 01 2005. ISBN 978-1-4419-0299-3. doi: 10.1007/978-1-4419-0300-6.

Wakefield, J. *Bayesian and Frequentist Regression Methods*. 01 2013. ISBN 978-1-4419-0924-4. doi: 10.1007/978-1-4419-0925-1.

Ware, J. H. Linear models for the analysis of longitudinal studies. *The American Statistician*, 39(2):95–101, 1985.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Wenzel, F., Roth, K., Veeling, B., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, pp. 10248–10259. PMLR, 2020.

Wilson, A. G. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

Wu, S., Zhang, H. R., Valiant, G., and Ré, C. On the generalization effects of linear transformations in data augmentation. In *ICML*, 2020.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L., and Zhang, P. Florence: A new foundation model for computer vision, 2021.

Zeno, C., Golan, I., Pakman, A., and Soudry, D. Why cold posteriors? on the suboptimal generalization of optimal bayes estimates. *3rd Symposium on Advances in Approximate Bayesian Inference*, 2020.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers, 2021.

Zhang, R., Li, C., Zhang, J., Chen, C., and Wilson, A. G. Cyclical stochastic gradient MCMC for Bayesian deep learning. In *International Conference on Learning Representations (ICLR 2020)*, 2020.

A. Omitted Proofs

In this section we detail all the calculations and proofs omitted in the main text.

A.1. Proof of Thm. 3.1

We have n independent samples $x_1, \dots, x_n$ and we consider an additive augmentation $\tilde{x}_i^b = R_\eta(x_i) = x_i + \eta_i^b$ where $\eta_i^b \sim \mathcal{N}(0, \sigma_\eta^2)$ for $b = 1, \dots, B$. Per datapoint x_i, we have B augmentations $\tilde{x}_i^1, \dots, \tilde{x}_i^B$ originating from x_i with independent underlying noise $\eta_i^1, \dots, \eta_i^B$. Augmentations of the same input display correlations which we can easily calculate

$$\begin{aligned}\operatorname{cov}(\tilde{x}_i^b, \tilde{x}_i^{b'}) &= \mathbb{E}[\tilde{x}_i^b \tilde{x}_i^{b'}] - \mathbb{E}[\tilde{x}_i^b]\mathbb{E}[\tilde{x}_j^{b'}] = \mathbb{E}[x_i^2] + \mathbb{E}[\eta_i^b x_i] + \mathbb{E}[\eta_i^{b'} x_i] + \mathbb{E}[\eta_i^b \eta_i^{b'}] - \mathbb{E}[x_i + \eta_i^b]\mathbb{E}[x_i + \eta_i^{b'}] \\ &= \mu^2 + \sigma^2 + \mathbb{1}_{\{b=b'\}}\sigma_\eta^2 - \mu^2 \\ &= \sigma^2 + \mathbb{1}_{\{b=b'\}}\sigma_\eta^2\end{aligned}$$

where we used the independence between $\eta_i^b, \eta_i^{b'}$ and x_i. On the other hand, augmentations from different samples remain independent:

$$\begin{aligned}\operatorname{cov}(\tilde{x}_i^b, \tilde{x}_j^{b'}) &= \mathbb{E}[\tilde{x}_i^b \tilde{x}_j^{b'}] - \mathbb{E}[\tilde{x}_i^b]\mathbb{E}[\tilde{x}_j^{b'}] = \mathbb{E}[x_i x_j] + \mathbb{E}[\eta_i^b x_j] + \mathbb{E}[\eta_j^{b'} x_i] + \mathbb{E}[\eta_i^b \eta_j^{b'}] - \mathbb{E}[x_i + \eta_i^b]\mathbb{E}[x_j + \eta_i^{b'}] \\ &= \mu^2 - \mu^2 \\ &= 0\end{aligned}$$

Collecting all x_i^b into one vector $\tilde{\boldsymbol{x}} \in \mathbb{R}^{Bn}$, we can describe the joint distribution as

$$\tilde{\boldsymbol{x}}|\mu \sim \mathcal{N}(\mu \mathbf{1}_{Bn}, \tilde{\boldsymbol{\Sigma}})$$

where $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{Bn \times Bn}$ is block-diagonal with blocks $\boldsymbol{\Sigma} = \sigma_\eta^2 \mathbb{1} + \sigma^2 \mathbf{1}_{Bn}\mathbf{1}_{Bn}^T$. We know by Bayes rule that the posterior is proportional to

$$p(\mu|\tilde{\boldsymbol{x}}) \propto p(\tilde{\boldsymbol{x}}|\mu)p(\mu)$$

Instead of factorizing $p(\tilde{\boldsymbol{x}}|\mu)$ we just work with the joint distribution directly:

$$\begin{aligned}p(\mu|\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{\Sigma}}) &\propto e^{-\frac{1}{2}(\tilde{\boldsymbol{x}} - \mu\mathbf{1}_{Bn})^T \tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\boldsymbol{x}} - \mu\mathbf{1}_{Bn})} e^{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2} = e^{-\frac{1}{2}\mu^2\left(\mathbf{1}_{Bn}^T \tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_{Bn} + \frac{1}{\sigma_0^2}\right) + \mu\left(\mathbf{1}_{Bn}\tilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{x} + \frac{\mu_0}{\sigma_0^2}\right)} \\ &\propto \mathcal{N}\left(\mu_{\text{post}}^{\tilde{\boldsymbol{\Sigma}}}, \left(\sigma_{\text{post}}^{\tilde{\boldsymbol{\Sigma}}}\right)^2\right)\end{aligned}$$

with the posterior mean and variance estimate

$$\mu_{\text{post}}^{\tilde{\boldsymbol{\Sigma}}} = \frac{\mathbf{1}_{Bn}^T \tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{x}} + \frac{\mu_0}{\sigma_0^2}}{\mathbf{1}_{Bn}^T \tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_{Bn} + \frac{1}{\sigma_0^2}} = \frac{\frac{1}{\sigma_\eta^2 + B\sigma^2}\sum_{i=1}^{Bn}\tilde{x}_i + \frac{\mu_0}{\sigma_0^2}}{\frac{Bn}{\sigma_\eta^2 + B\sigma^2} + \frac{1}{\sigma_0^2}} \qquad \left(\sigma_{\text{post}}^{\tilde{\boldsymbol{\Sigma}}}\right)^2 = \frac{1}{\mathbf{1}_{Bn}^T \tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{1}_{Bn} + \frac{1}{\sigma_0^2}} = \frac{1}{\frac{Bn}{\sigma_\eta^2 + B\sigma^2} + \frac{1}{\sigma_0^2}}$$

We can perform the same calculation for tempered likelihoods assuming (wrongly) that the likelihood factorizes, leading to the tempered posterior

$$p_T(\mu|\tilde{\boldsymbol{x}}; \mathbb{1}) = \mathcal{N}(\mu_T, \sigma_T^2)$$

with tempered posterior statistics

$$\mu_T = \left(\frac{Bn}{T(\sigma^2 + \sigma_\eta^2)} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{1}{T(\sigma^2 + \sigma_\eta^2)}\sum_{i=1}^{Bn}\tilde{x}_i\right) \qquad \sigma_T^2 = \left(\frac{Bn}{T(\sigma^2 + \sigma_\eta^2)} + \frac{1}{\sigma_0^2}\right)^{-1}$$

We can check that setting the temperature as

$$T = \frac{\sigma_\eta^2 + B\sigma^2}{\sigma_\eta^2 + \sigma^2}$$

leads to an equality in distribution, i.e. $p_T(\mu|\boldsymbol{x}; \mathbb{1}) = p(\mu|\tilde{\boldsymbol{x}}; \tilde{\boldsymbol{\Sigma}})$

A.2. Correlations in Regression Setting

We can calculate the correlation in the regression setting as follows:

$$\operatorname{cov}(\tilde{\epsilon}_i, \tilde{\epsilon}_j) = \mathbb{E}[\tilde{\epsilon}_i \tilde{\epsilon}_j] = \mathbb{E}[(\epsilon + \delta_{\boldsymbol{\eta}_i})(\epsilon + \delta_{\boldsymbol{\eta}_j})] = \mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$$

where we used that $\tilde{\epsilon}$'s are centered by assumption since δ's are centered. On the other hand, we can compute the respective variance as

$$\operatorname{var}(\tilde{\epsilon}_i) = \mathbb{E}[(\epsilon + \delta_{\boldsymbol{\eta}_i})^2] = \mathbb{E}[\epsilon^2] + \mathbb{E}[\delta_{\boldsymbol{\eta}_i}^2] = \sigma^2 + \operatorname{var}(\delta_{\boldsymbol{\eta}})$$

The resulting correlation is thus given by

$$\operatorname{cor}(\tilde{\epsilon}_i, \tilde{\epsilon}_j) = \frac{\operatorname{cov}(\tilde{\epsilon}_i, \tilde{\epsilon}_j)}{\operatorname{var}(\tilde{\epsilon}_i)} = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \operatorname{var}(\delta_{\boldsymbol{\eta}})}$$

A.3. Correlations in Logistic Regression

Correlations are not an artifact of the regression setting but also arise in classification tasks. To that end, we rely on the latent variable model of logistic regression, where a latent variable

$$z = f_{\boldsymbol{\theta}_*}(\boldsymbol{x}) + \epsilon$$

is introduced, with $\epsilon \sim \operatorname{Logistic}(0, 1)$ following a centered logistic variable with variance. These latent variables z then give rise to the response through the relation

$$y = \mathbb{1}_{\{z \geq 0\}}.$$

In perfect duality to the regression setting, the errors implied by augmentations of the same covariate $\boldsymbol{x}$ exhibit significant correlations. More concretely, we again augment as $\tilde{\boldsymbol{x}} = R_{\boldsymbol{\eta}}(\boldsymbol{x})$ and set $\tilde{y} = y$. Here we have a latent variable $\tilde{z}$ which might differ from z. Notice that in the case of $\tilde{z} = z$, we are in the same setting as regression. In general, this might however not be true but we know that for two augmentations $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_j$ with errors $\tilde{\epsilon}_i$ and $\tilde{\epsilon}_j$, the following relation holds:

$$\tilde{y}_i = y = \tilde{y}_j \iff \mathbb{1}_{\{f_{\boldsymbol{\theta}_*}(R_{\boldsymbol{\eta}_i}(\boldsymbol{x})) + \tilde{\epsilon}_i \geq 0\}} = \mathbb{1}_{\{f_{\boldsymbol{\theta}_*}(\boldsymbol{x}) + \epsilon \geq 0\}} = \mathbb{1}_{\{f_{\boldsymbol{\theta}_*}(R_{\boldsymbol{\eta}_j}(\boldsymbol{x})) + \tilde{\epsilon}_j \geq 0\}} \quad \text{a.s.}$$

Hence, the two errors are related through the equation

$$\mathbb{1}_{\{f_{\boldsymbol{\theta}_*}(R_{\boldsymbol{\eta}_i}(\boldsymbol{x})) + \tilde{\epsilon}_i \geq 0\}} = \mathbb{1}_{\{f_{\boldsymbol{\theta}_*}(R_{\boldsymbol{\eta}_j}(\boldsymbol{x})) + \tilde{\epsilon}_j \geq 0\}} \quad \text{a.s.}$$

and hence $\tilde{\epsilon}_i$ and $\tilde{\epsilon}_j$ share significant correlation. In case of perfect invariance, i.e. $f_{\boldsymbol{\theta}_*}(R_{\boldsymbol{\eta}_i}(\boldsymbol{x})) = f_{\boldsymbol{\theta}_*}(R_{\boldsymbol{\eta}_j}(\boldsymbol{x})) = f_{\boldsymbol{\theta}_*}(\boldsymbol{x})$, we find that

$$\mathbb{1}_{\{\tilde{\epsilon}_i \geq -f_{\boldsymbol{\theta}_*}(\boldsymbol{x})\}} = \mathbb{1}_{\{\tilde{\epsilon}_j \geq -f_{\boldsymbol{\theta}_*}(\boldsymbol{x})\}} \quad \text{a.s.}$$

which, for infinitely supported random variables such as the logistic distribution, can only hold if $\tilde{\epsilon}_i = \tilde{\epsilon}_j$ a.s., hence leading to perfectly correlated errors.

A.4. Correlation in Linear Regression

As seen in the main text, it holds that

$$\tilde{\epsilon} = \epsilon + \delta_{\boldsymbol{\eta}}$$

where $\delta_{\boldsymbol{\eta}} = -\boldsymbol{\eta}^T \boldsymbol{x}$. Since $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$, it is evident that $\mathbb{E}[\delta_{\boldsymbol{\eta}}] = 0$ since $\mathbb{E}[\boldsymbol{\eta}] = \mathbf{0}$. Moreover

$$\operatorname{var}(\tilde{\epsilon}) = \sigma_\epsilon^2 + \mathbb{E}[\boldsymbol{x}^T \boldsymbol{\eta} \boldsymbol{\eta}^T \boldsymbol{x}] = \sigma_\epsilon^2 + \boldsymbol{x}^T \boldsymbol{\Sigma}_\eta \boldsymbol{x}$$

Finally, for two augmentations $\boldsymbol{\eta}_i$, $\boldsymbol{\eta}_j$, it holds for the respective errors that $\tilde{\epsilon}_i, \tilde{\epsilon}_j$ that

$$\operatorname{cov}(\tilde{\epsilon}_i, \tilde{\epsilon}_j) = \mathbb{E}[\tilde{\epsilon}_i \tilde{\epsilon}_j] = \mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$$

We can hence conclude that the correlation is given by

$$\operatorname{cor}(\tilde{\epsilon}_i, \tilde{\epsilon}_j) = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \boldsymbol{x}^T \boldsymbol{\Sigma}_\eta \boldsymbol{x}}$$

A.5. Intermediate Values for Correlation

In this section we study to what degree tempering with T can reduce the misspecification stemming from ignoring the correlation between errors. The posteriors in both the i.i.d. and correlation setting are intractable for general models $f_{\boldsymbol{\theta}}$ but the two likelihoods take a simple form. We show that in general, no temperature T can be found to make the two likelihoods match, i.e. their KL-divergence is not 0. This in turn implies that the posteriors cannot match either (both models use the same prior).

For simplicity we again assume a linear regression setting with one sample $(\boldsymbol{x}, y)$ and additive augmentations, i.e. $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^T\boldsymbol{x}$ and $G_{\boldsymbol{\eta}}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\eta}$ for $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2\mathbb{1})$. As a result, we find that the augmentation errors $\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_B$ still follow a joint Gaussian distribution with

$$\tilde{\boldsymbol{\epsilon}} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}})$$

with $\tilde{\Sigma}_{ii} = \sigma_\epsilon^2 + \sigma_\eta^2$ and $\tilde{\Sigma}_{ij} = \sigma_\epsilon^2$. On the other hand, we can also decide to ignore the correlations and temper, i.e. just work with $\frac{\sigma_\epsilon^2+\sigma_\eta^2}{T}\mathbb{1}$, and thus factorize the likelihood. In both cases, we have a Gaussian likelihood of the form

$$p_T(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}; \mathbb{1}) \sim \mathcal{N}(\tilde{\boldsymbol{X}}\boldsymbol{\theta}, (\sigma_\epsilon^2 + \sigma_\eta^2)\mathbb{1}) \qquad p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}; \tilde{\boldsymbol{\Sigma}}), \sim \mathcal{N}(\tilde{\boldsymbol{X}}\boldsymbol{\theta}, \tilde{\boldsymbol{\Sigma}}))$$

where we define the matrix of augmentations $\tilde{\boldsymbol{X}} \in \mathbb{R}^{B\times d}$.

In general, we can express the KL-divergence between two Gaussians $\rho_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\rho_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ as

$$\mathrm{KL}(\rho_1 \mid\mid \rho_2) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|} - B + Tr(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]. \tag{10}$$

In our case, let $\rho_1 \sim \mathcal{N}(\tilde{\boldsymbol{X}}\boldsymbol{\theta}, \frac{\sigma_\epsilon^2+\sigma_\eta^2}{T}\mathbb{1})$ be the tempered likelihood and $\rho_2 \sim \mathcal{N}(\tilde{\boldsymbol{X}}\boldsymbol{\theta}, \tilde{\boldsymbol{\Sigma}})$ the correlated likelihood. We have that $|\frac{\sigma_\epsilon^2+\sigma_\eta^2}{T}\mathbb{1}| = \left(\frac{\sigma_\epsilon^2+\sigma_\eta^2}{T}\right)^B$ and therefore the derivative of its logarithm w.r.t T is:

$$\frac{\partial}{\partial T} B\log\frac{\sigma_\epsilon^2 + \sigma_\eta^2}{T} = -\frac{B}{T}. \tag{11}$$

For the trace term we have that:

$$\frac{\partial}{\partial T} Tr\left(\frac{(\sigma_\epsilon^2 + \sigma_\eta^2)\tilde{\boldsymbol{\Sigma}}^{-1}}{T}\right) = -\frac{\sigma_\epsilon^2 + \sigma_\eta^2}{T^2} Tr(\tilde{\boldsymbol{\Sigma}}^{-1}) \tag{12}$$

So, by setting the derivative of the KL to zero and solving for T:

$$T^* = (\sigma_\epsilon^2 + \sigma_\eta^2)\frac{Tr(\tilde{\boldsymbol{\Sigma}}^{-1})}{B} \tag{13}$$

Plugging-in the optimal value T^* into the KL-divergence gives

$$2\,\mathrm{KL}\left(p_{T^*}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}; \mathbb{1}) \mid\mid p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}; \tilde{\boldsymbol{\Sigma}})\right) = \log(|\tilde{\boldsymbol{\Sigma}}|) - B\log\left(\frac{B}{\mathrm{Tr}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\right)}\right) \overset{!}{=} 0$$

or equivalently

$$\log(|\tilde{\boldsymbol{\Sigma}}|) \overset{!}{=} B\log\left(\frac{B}{\mathrm{Tr}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\right)}\right) \iff |\tilde{\boldsymbol{\Sigma}}| \overset{!}{=} \left(\frac{B}{\mathrm{Tr}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\right)}\right)^B$$

We can plug-in the concrete $\tilde{\boldsymbol{\Sigma}} = \sigma_\eta^2\mathbb{1} + \sigma_\epsilon^2\mathbf{1}_B\mathbf{1}_B^T$ to find

$$(\sigma_\eta^2 + B\sigma_\epsilon^2)\sigma_\eta^{2(B-1)} \overset{!}{=} \left(\frac{\sigma_\eta^2(\sigma_\eta^2 + \sigma_\epsilon^2 B)}{\sigma_\eta^2 + \sigma_\epsilon^2(B-1)}\right)^B$$

We observe that if $B > 1$ this equation cannot hoald for a general σ_η^2. On the other hand, $\sigma_\eta^2 6$ governs the correlation coefficient , which was shown in Appendix A.4 to be

$$\text{cor}(\tilde{\epsilon}_i, \tilde{\epsilon}_j) = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\eta^2}$$

Tempering can hence not match the likelihoods exactly for a general correlation (i.e. general σ_η^2) and hence the posteriors cannot match either. 1 and B !

B. Experimental Details

In this section we describe the experimental setup, including the architectural details and the SG-MCMC hyperparameters. For the SG-MCMC sampler, we adapted the code from Wenzel et al. (2020)[2]. For the implementation of the group equivariant layers, we used the code from Veeling et al. (2018)[3]

B.1. SG-MCMC

For the experiments with ResNet20 and G-ResNets on CIFAR-10, we have the following hyperparameters:

- initial learning rate: 0.1
- burn-in period: 150 epochs
- cycle length: 50 epochs
- total training time: 1500 epochs

For the experiments with convolutional and G-convolutional networks on CIFAR-10, we have the following hyperparameters:

- initial learning rate: 0.5
- burn-in period: 150 epochs
- cycle length: 50 epochs
- total running time: 1000 epochs

Note that for the experiments where only the likelihood is tempered, the reparamterization of the SG-MCMC updates explained in Section 4 force one to adapt the learning rate to the temperature: if we increase the temperature by 10 times, we should make the learning 10 times larger. Therefore the learning rate specified above refers to the learning rate at $T = 1$. We stress that we adapt the learning rate for a fair comparison of runs with different temperature, but changing the learning rate does *not* affect the posterior distribution, while the temperature does. We refer to Aitchison (2021) for the details of a similar argument in the different context of dataset curation.

Finally, the experiments are executed on Nvidia DGX-1 GPU nodes equipped with 4 20-core Xeon E5-2698v4 processors, 512 GB of memory and 8 Nvidia V100 GPUs.

B.2. Neural Network Architectures

For the SG-MCMC experiments, we use a 20-layer architecture with residual layers (He et al., 2016) and batch normalization. The architecture with only convolutional layers is composed of 4 convolutional layers, the first two with 32 filters and the last two with 64 filters. We experiments both with 2-strided and 1-strided convolutions, as detailed in Section 4. The batch size is 128 across all experiments. The group equivariant architectures (G-ResNet and G-Conv) are designed by replacing the convolutional layers and pooling layer with the Group equivariant counterpart, as detailed in (Cohen & Welling, 2016). To build more invariant architectures, also G-batch normalization layers should replace batch norm. However, due to a code incompatibility between the two code repositories (SG-MCMC and Group equivariant layers) we are not able to use G-batch norms.

[2] https://github.com/google-research/google-research/tree/master/cold_posterior_bnn
[3] https://github.com/basveeling/keras-gcnn

B.3. Data Augmentation

The datapoints are always transformed in an online fashion and not precomputed, as it is commonly done in deep learning for memory efficiency. This means that when a batch of original datapoints S_t is processed at epoch t, then each datapoint is preprocessed with the augmentation function and then propagated through the network. This means that at every epoch we potentially have new augmented datapoints. Therefore the total the number of augmentationd per datapoint is equal to the number of epochs. On the contrary, for the experiment with fixed number of augmentations, we fix a sequence of B random seeds, which is then repeated every B epochs, guaranteeing that the total number of augmentations is fixed.

C. Further Results

Here we show the figures of the plot for all the experiments of Section 4.

Hot or Cold Posteriors? In Fig. 10, we plot the accuracy results for *ResNet20* for the tempered posterior and tempered likelihood cases.

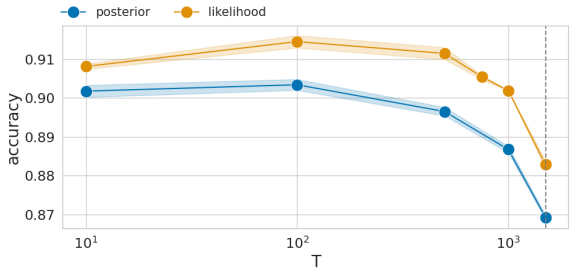


Figure 9: Test accuracy as a function of the temperature T for posterior tempering (blue) and likelihood tempering (orange).

G-Convolutions alleviates the CPE In Fig. 10, we plot the accuracy results for *ResNet20* and *G-ResNet20* trained with and without additional 90° rotations. Additionally, in Fig. 11, we plot the evolution of the invariance measure (total variation) for the *G-ResNet* trained with and without additional 90° rotation.

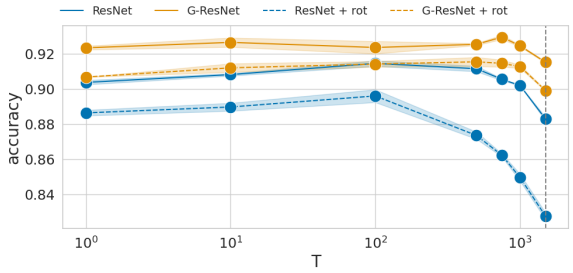


Figure 10: Test accuracy as a function of the temperature T for a ResNet with standard convolutions (blue) and G-convolutions (orange), with or without extra 90° rotations (dashed lines).

An Almost Invariant Model Finally, in Fig. 12 and Fig. 13 we plot the accuracy for the G-Convolutional networks trained with flip and 90° rotations and with additional random 10° rotations.

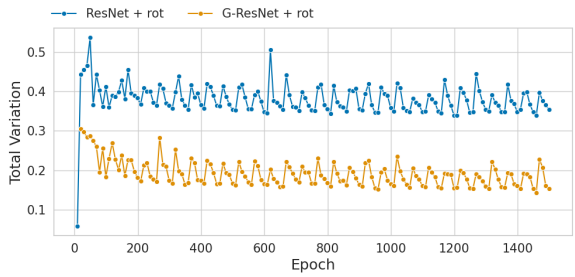


Figure 11: Total variation during learning for the experiments with G-ResNet20. A lower value of the total variation indicates a greater degree of invariance.

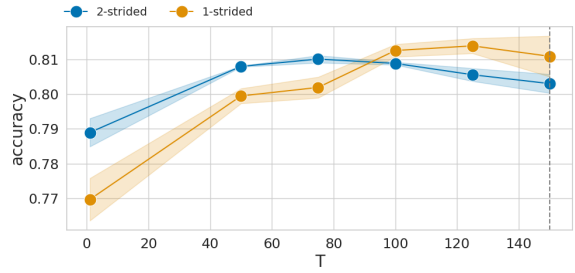


Figure 12: Test accuracy as a function of the temperature for a model trained 2-strided (blue) and 1-strided (orange) G-convolutional network.

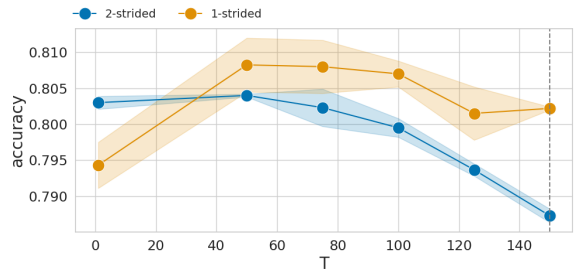


Figure 13: Same setting as in Fig. 12, but this time we perform an extra random rotation of 10°. Note how the loss of invariance shifts the optimal temperature significantly, in both the 2-strided and 1-strided case.