The Cold Posterior Effect Indicates Underfitting, and Cold Posteriors Represent a Fully Bayesian Method to Mitigate It

Anonymous authors
Paper under double-blind review

Abstract

The cold posterior effect (CPE) (Wenzel et al., 2020) in Bayesian deep learning shows that, for posteriors with a temperature $T < 1$, the resulting posterior predictive could have better performance than the Bayesian posterior ($T = 1$). As the Bayesian posterior is known to be optimal under perfect model specification, many recent works have studied the presence of CPE as a model misspecification problem, arising from the prior and/or from the likelihood. In this work, we provide a more nuanced understanding of the CPE as we show that *misspecification leads to CPE only when the resulting Bayesian posterior underfits*. In fact, we theoretically show that if there is no underfitting, there is no CPE. Furthermore, we show that these *tempered posteriors* with ($T < 1$) are indeed proper Bayesian posteriors with a different combination of likelihood and prior parameterized by T. Within the *empirical Bayes* framework, this observation validates the adjustment of the temperature hyperparameter T as a straightforward approach to mitigate underfitting in the Bayesian posterior. In essence, we show that by fine-tuning the temperature T we implicitly utilize alternative Bayesian posteriors, albeit with less misspecified likelihood and prior distributions.

1 Introduction

In Bayesian deep learning, the cold posterior effect (CPE) (Wenzel et al., 2020) refers to the phenomenon in which if we artificially "temper" the posterior by either $p(\boldsymbol{\theta}|D) \propto (p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}))^{1/T}$ or $p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^{1/T}p(\boldsymbol{\theta})$ with a temperature $T < 1$, the resulting posterior enjoys better predictive performance than the standard Bayesian posterior (with $T = 1$). The discovery of the CPE has sparked debates in the community about its potential contributing factors.

If the prior and likelihood are properly specified, the Bayesian solution (i.e., $T = 1$) should be optimal (Gelman et al., 2013), assuming approximate inference is properly working. Hence, the presence of the CPE implies either the prior (Wenzel et al., 2020; Fortuin et al., 2022), the likelihood (Aitchison, 2021; Kapoor et al., 2022), or both are misspecified. This has been, so far, the main argument of many works trying to explain the CPE.

One line of research examines the impact of the prior misspecification on the CPE (Wenzel et al., 2020; Fortuin et al., 2022). The priors of modern Bayesian neural networks are often selected for tractability. Consequently, the quality of the selected priors in relation to the CPE is a natural concern. Previous research has revealed that while adjusting priors can help alleviate the CPE in certain cases, there are instances where the effect persists despite such adjustments (Fortuin et al., 2022). Some studies even show that the role of priors may not be critical (Izmailov et al., 2021). Therefore, the impact of priors on the CPE remains an open question.

Furthermore, the influence of likelihood misspecification on CPE has also been investigated (Aitchison, 2021; Noci et al., 2021; Kapoor et al., 2022; Fortuin et al., 2022), and has been identified to be particularly relevant in curated datasets (Aitchison, 2021; Kapoor et al., 2022). Several studies have proposed alternative likelihood functions to address this issue and successfully mitigate the CPE (Nabarro et al., 2022; Kapoor et al., 2022). However, the underlying relation between the likelihood and CPE remains a partially unresolved question. Notably, the CPE usually emerges when data augmentation (DA) techniques are used (Wenzel

et al., 2020; Izmailov et al., 2021; Fortuin et al., 2022; Noci et al., 2021; Nabarro et al., 2022; Kapoor et al., 2022). A popular hypothesis is that using DA implies the introduction of a randomly perturbed log-likelihood, which lacks a clear interpretation as a valid likelihood function (Wenzel et al., 2020; Izmailov et al., 2021). However, Nabarro et al. (2022) demonstrates that the CPE persists even when a proper likelihood function incorporating DA is defined. Therefore, further investigation is needed to fully understand their relationship.

Other works argued that CPE could mainly be an artifact of inaccurate approximate inference methods, especially in the context of neural networks, where the posteriors are extremely high dimensional and complex (Izmailov et al., 2021). However, many of the previously mentioned works have also found setups where the CPE either disappears or is significantly alleviated through the adoption of better priors and/or better likelihoods with approximate inference methods. In these studies, the same approximate inference methods were used to illustrate, for example, how using a standard likelihood function leads to the observation of CPE and how using an alternative likelihood function removes it (Aitchison, 2021; Noci et al., 2021; Kapoor et al., 2022). In other instances, under the same approximate inference scheme, CPE is observed when using certain types of priors but it is strongly alleviated when an alternative class of priors is utilized (Wenzel et al., 2020; Fortuin et al., 2022). Therefore, there is compelling evidence suggesting that approximate methods are not, at least, a necessary condition for the CPE.

This study, both theoretically and empirically, demonstrates that the presence of the cold posterior effect (CPE) implies the existence of underfitting; in other words, *if there is no underfitting, there is no CPE*. Integrating this perspective with previous findings suggesting that CPE indicates misspecified likelihood, prior, or both (Gelman et al., 2013), we conclude that CPE implies both misspecification and underfitting. Consequently, mitigating CPE necessitates addressing both aspects. Notably, simplifying the issue by solely focusing on misspecification is insufficient, as misspecification can lead Bayesian methods to both underfitting and overfitting (Domingos, 2000; Immer et al., 2021; Kapoor et al., 2022); CPE only arises when underfitting occurs. This study thus offers a nuanced perspective on the factors contributing to CPE. Additionally, we demonstrate that tempered posteriors represent proper Bayesian posteriors under different likelihood and prior distributions, jointly parameterized by the temperature parameter T. Consequently, by adjusting T, we effectively identify Bayesian posteriors with less misspecified likelihood and prior distributions, leading to a more accurate representation of the training data and improved generalization performance. Furthermore, we delve into the relationship between prior/likelihood misspecification, data augmentation, approximate inference, and CPE, offering insights into potential strategies for addressing these issues.

Contributions **(i)** We theoretically demonstrate that the presence of the CPE implies the Bayesian posterior is underfitting in Section 3. **(ii)** We show that any tempered posterior is a proper Bayesian posterior with an alternative likelihood and prior distribution in Section 4. **(iii)** We show in Section 5 that likelihood misspecification and prior misspecification result in CPE only if they also induce underfitting. Furthermore, the tempered posteriors offer an effective and well-founded Bayesian mechanism to address the underfitting problem. **(iv)** Finally, we show that data augmentation results in stronger CPE because it induces a stronger underfitting of the Bayesian posterior in Section 6. In conclusion, our theoretical analysis reveals that the occurrence of the CPE signifies underfitting of the Bayesian posterior. Also, fine-tuning the temperature in tempered posteriors offers a well-founded and effective Bayesian approach to mitigate the issue. Furthermore, our work aims to settle the debate surrounding CPE and its implications for Bayesian principles, specifically within the context of deep learning.

2 Background

2.1 Notation

Let us start by introducing basic notation. Consider a supervised learning problem with the sample space $\mathcal{Y}\times\mathcal{X}$. Let $D = \{(\boldsymbol{y}_i, \boldsymbol{x}_i)\}_{i=1}^n$ denote the training data, which we assume to be generated from an unknown data-generating distribution ν on $\mathcal{Y}\times\mathcal{X}$. We also assume we have a family of probabilistic models parameterized by $\boldsymbol{\Theta}$, where each $\boldsymbol{\theta}$ defines a conditional probability distribution denoted by $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$. The standard metric to measure the quality of a probabilistic model $\boldsymbol{\theta}$ on a sample $(\boldsymbol{y},\boldsymbol{x})$ is the (negative) log-loss $-\ln p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$. The expected (or population) loss of a probabilistic model $\boldsymbol{\theta}$ is defined as $L(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{y},\boldsymbol{x})\sim\nu}[-\ln p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})]$, and the

empirical loss of the model $\boldsymbol{\theta}$ on the data D is defined as $\hat{L}(D, \boldsymbol{\theta}) = -\frac{1}{n}\sum_{i\in[n]} \ln p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = -\frac{1}{n}\ln p(D|\boldsymbol{\theta})$. We might interchange the loss expression, $\hat{L}(D, \boldsymbol{\theta})$, and the negative log-likelihood expression, $-\frac{1}{n}\ln p(D|\boldsymbol{\theta})$, in the paper for presentation. Also, if it induces no ambiguity, we use $\mathbb{E}_\nu[\cdot]$ as a shorthand for $\mathbb{E}_{(\boldsymbol{y},\boldsymbol{x})\sim\nu}[\cdot]$.

2.2 (Generalized) Bayesian Learning

In Bayesian learning, we learn a probability distribution $\rho(\boldsymbol{\theta}|D)$, often called a posterior, over the parameter space $\boldsymbol{\Theta}$ from the training data D. Given a new input $\boldsymbol{x}$, the posterior ρ makes the prediction about $\boldsymbol{y}$ through (an approximation of) *Bayesian model averaging (BMA)* $p(\boldsymbol{y}|\boldsymbol{x}, \rho) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}[p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})]$, where the posterior ρ is used to combine the predictions of the models. Again, if it induces no ambiguity, we use $\mathbb{E}_\rho[\cdot]$ as a shorthand for $\mathbb{E}_{\boldsymbol{\theta}\sim\rho}[\cdot]$. The predictive performance of such BMA is usually measured by the Bayes loss, defined by

$$B(\rho) = \mathbb{E}_\nu[-\ln \mathbb{E}_\rho[p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})]]\,. \tag{1}$$

For some $\lambda > 0$ and a prior $p(\boldsymbol{\theta})$, the so-called *tempered posteriors* (or the generalized Bayes posterior) (Barron & Cover, 1991; Zhang, 2006; Bissiri et al., 2016; Grünwald & van Ommen, 2017), are defined as a probability distribution

$$p^\lambda(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^\lambda p(\boldsymbol{\theta})\,, \tag{2}$$

where $p(D|\boldsymbol{\theta})^\lambda = \prod_i p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})^\lambda$. Note that, when $\lambda \neq 1$, $\int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})^\lambda d\boldsymbol{y}$ might not be 1 in general.

Even though many works on CPE use the parameter $T = 1/\lambda$ instead, we adopt λ in the rest of the work for the convenience of derivations. Therefore, the CPE ($T < 1$) corresponds to when $\lambda > 1$. We also note that while some works study CPE with a full-tempering posterior, where the prior is also tempered, many works also find CPE for likelihood-tempering posterior (see (Wenzel et al., 2020) and the references therein). Also, with some widely chosen priors (e.g., zero-centered Gaussian priors), the likelihood-tempering posteriors are equivalent to full-tempering posteriors with rescaled prior variances (Aitchison, 2021; Bachmann et al., 2022).

When $\lambda = 1$, the tempered posterior equals the (standard) Bayesian posterior. The tempered posterior can be obtained by optimizing a generalization of the so-called (generalized) ELBO objective (Alquier et al., 2016; Higgins et al., 2017), which, for convenience, we write as follows:

$$p^\lambda(\boldsymbol{\theta}|D) = \arg\min_\rho \mathbb{E}_\rho[-\ln p(D|\boldsymbol{\theta})] + \frac{1}{\lambda}\mathrm{KL}(\rho(\boldsymbol{\theta}|D), p(\boldsymbol{\theta}))\,. \tag{3}$$

The first term is known as the (un-normalized) *reconstruction error* or the empirical Gibbs loss of the posterior ρ on the data D, denoted as $\hat{G}(\rho, D) = \mathbb{E}_\rho[-\frac{1}{n}\ln p(D|\boldsymbol{\theta})]$, which further equals to $\mathbb{E}_\rho[\hat{L}(D, \boldsymbol{\theta})]$. Therefore, it is often used as the *training loss* in Bayesian learning (Morningstar et al., 2022). The second term is a Kullback-Leibler divergence between the posterior $\rho(\boldsymbol{\theta}|D)$ and the prior $p(\boldsymbol{\theta})$ scaled by a hyper-parameter λ.

If it induces no ambiguity, we will use p^λ as a shorthand for $p^\lambda(\boldsymbol{\theta}|D)$. So, for example, $B(p^\lambda)$ would refer to the expected Bayes loss of the tempered-posterior $p^\lambda(\boldsymbol{\theta}|D)$. In the rest of this work, we will interpret the CPE as how changes in the parameter λ affect the *test error* and the *training error* of p^λ or, equivalently, the Bayes loss $B(p^\lambda)$ and the empirical Gibbs loss $\hat{G}(p^\lambda, D)$.

3 The presence of the CPE implies underfitting

A standard understanding for underfitting refers to a situation when the trained model cannot properly capture the relationship between input and output in the data-generating process, resulting in high errors on both the training data and testing data. In the context of highly flexible model classes such as neural networks, underfitting refers to a scenario where the trained model exhibits (much) higher training and testing losses compared to what is achievable. Essentially, it means that there exists another model in the model class that achieves lower training and testing losses simultaneously. In the context of Bayesian inference, we argue that the Bayesian posterior is underfitting if there exists another posterior distribution with lower empirical Gibbs and Bayes losses at the same time. In fact, we will show later in Section 4 that such a posterior is essentially another *Bayesian posterior but with a different prior and likelihood function*. Before delving into that, we focus on characterizing the cold posterior effect (CPE) and its connection to underfitting.

As previously discussed, the CPE describes the phenomenon of getting better predictive performance when we make the parameter of the tempered posterior, λ, higher than 1. The next definition introduces a formal characterization. *We do not claim this is the best possible formal characterization.* However, through the rest of the paper, we will show that this simple characterization is enough to understand the relationship between CPE and underfitting.

Definition 1. *We say there is a CPE for Bayes loss if and only if the gradient of the Bayes loss of the posterior p^λ, $B(p^\lambda)$, evaluated at $\lambda = 1$ is negative. That is,*

$$\nabla_\lambda B(p^\lambda)_{|\lambda=1} < 0\,, \tag{4}$$

where the magnitude of the gradient $\nabla_\lambda B(p^\lambda)\big|_{\lambda=1}$ defines the strength of the CPE.

According to the above definition, a (relatively large) negative gradient $\nabla_\lambda B(p^\lambda)_{|\lambda=1}$ implies that by making λ slightly greater than 1, we will have a (relatively large) reduction in the Bayes loss with respect to the Bayesian posterior. Note that if the gradient $\nabla_\lambda B(p^\lambda)_{|\lambda=1}$ is not relatively large and negative, then we can not expect a relatively large reduction in the Bayes loss and, in consequence, the CPE will not be significant. Obviously, this formal definition could also be extended to other specific λ values different from 1, or even consider some aggregation over different $\lambda > 1$ values. We will stick to this definition because it is simpler, and the insights and conclusions extracted here can be easily extrapolated to other similar definitions involving the gradient of the Bayes loss.

Next, we present another critical observation. We postpone the proofs in this section to Appendix A.

Proposition 2. *The gradient of the empirical Gibbs loss of the tempered posterior p^λ satisfies*

$$\forall \lambda \geq 0 \quad \nabla_\lambda \hat{G}(p^\lambda, D) = -\mathbb{V}_{p^\lambda}\big(\ln p(D|\boldsymbol{\theta})\big) \leq 0\,, \tag{5}$$

where $\mathbb{V}(\cdot)$ denotes the variance.

As shown in Proposition 6 in Appendix A, to achieve $\mathbb{V}_{p^\lambda}\big(\ln p(D|\boldsymbol{\theta})\big) = 0$, we need $p^\lambda(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta})$, implying that the data has no influence on the posterior. In consequence, in practical scenarios, $\mathbb{V}_{p^\lambda}\big(\ln p(D|\boldsymbol{\theta})\big)$ will always be greater than zero. Thus, increasing λ will monotonically reduce the empirical Gibbs loss $\hat{G}(p^\lambda, D)$ (i.e., the *train error*) of p^λ. The next result also shows that the empirical Gibbs loss of the Bayesian posterior $\hat{G}(p^{\lambda=1})$ cannot reach its *floor* to observe the CPE.

Proposition 3. *A necessary condition for the presence of the CPE, as defined in Definition 1, is that*

$$\hat{G}(p^{\lambda=1}, D) > \min_{\boldsymbol{\theta}} -\ln p(D|\boldsymbol{\theta})\,.$$

Insight 1. *Definition 1 in combination with Proposition 2 state that if the CPE is present, by making $\lambda > 1$, the test loss $B(p^\lambda)$ and the empirical Gibbs loss $\hat{G}(p^\lambda, D)$ will be reduced at the same time. Furthermore, Proposition 3 states that the Bayesian posterior still has room to fit the training data further (e.g., by placing more probability mass on the maximum likelihood estimator). From here, we can deduce that the presence of CPE implies that the original Bayesian posterior ($\lambda = 1$) is experiencing underfitting. This conclusion arises because there exists another Bayesian posterior (i.e, $p^\lambda(\boldsymbol{\theta}|D)$ with $\lambda > 1$) that has lower training (Proposition 3) and testing (Definition 1) loss at the same time. Further elaboration on the nature of $p^\lambda(\boldsymbol{\theta}|D)$ as another Bayesian posterior will be provided later in Section 4. In short, if there is CPE, the original Bayesian posterior is underfitting. Or, equivalently, if the original Bayesian posterior does not underfit, there is no CPE.*

However, a final question arises: when is $\lambda = 1$ (the original Bayesian posterior of interest) *optimal*? More precisely, when does the gradient of the Bayes loss with respect to λ evaluated at $\lambda = 1$ become zero ($\nabla_\lambda B(p^\lambda)_{|\lambda=1} = 0$)? This would imply that neither (infinitesimally) increasing nor decreasing λ changes the predictive performance. We will see that this condition is closely related to the situation that updating such a Bayesian posterior with more data does not enhance its fit to the original training data better. In other words, when such a Bayesian posterior contains more information about the data-generating distribution, it continues to *fit the originally provided training data in a similar manner.*

We start by denoting $\tilde{p}^{\lambda}(\boldsymbol{\theta}|D,(\boldsymbol{y},\boldsymbol{x}))$ as the distribution obtained by updating the posterior $p^{\lambda}(\boldsymbol{\theta}|D)$ with one new sample $(\boldsymbol{y},\boldsymbol{x})$, i.e., $\tilde{p}^{\lambda}(\boldsymbol{\theta}|D,(\boldsymbol{y},\boldsymbol{x})) \propto p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})p^{\lambda}(\boldsymbol{\theta}|D)$. And we also denote $\bar{p}^{\lambda}$ as the distribution resulting from averaging $\tilde{p}^{\lambda}(\boldsymbol{\theta}|D,(\boldsymbol{y},\boldsymbol{x}))$ over different *unseen* samples from the data-generating distribution $(\boldsymbol{y},\boldsymbol{x}) \sim \nu(\boldsymbol{y},\boldsymbol{x})$:

$$\bar{p}^{\lambda}(\boldsymbol{\theta}|D) = \mathbb{E}_{\nu}\left[\tilde{p}^{\lambda}(\boldsymbol{\theta}|D,(\boldsymbol{y},\boldsymbol{x}))\right]. \tag{6}$$

In this sense, $\bar{p}^{\lambda}$ represents how the posterior p^{λ} would be, on average, after being updated with a new sample from the data-generating distribution. This updated posterior contains a bit more information about the data-generating distribution, compared to p^{λ}. Using the updated posterior $\bar{p}^{\lambda}$, the following result introduces a characterization of the *optimality* of the original Bayesian posterior.

Theorem 4. *The gradient of the Bayes loss at $\lambda = 1$ is null, i.e., $\nabla_{\lambda} B(p^{\lambda})_{|\lambda=1} = 0$, if and only if,*

$$\hat{G}(p^{\lambda=1}, D) = \hat{G}(\bar{p}^{\lambda=1}, D).$$

Insight 2. *The original Bayesian posterior of interest is optimal if after updating it using the procedure described in Eq. 6, or in other words, after exposing the Bayesian posterior to more data from the data-generating distribution, the empirical Gibbs loss over the initial training data remains unchanged.*

We will then show that the tempered posterior $p^{\lambda}(\boldsymbol{\theta}|D)$ is actually yet another Bayesian posterior.

4 Tempered Posteriors are Bayesian Posteriors

As previously discussed, the CPE phenomenon involves achieving improved predictive accuracy by employing a tempered posterior. A potential criticism is that this tempered posterior does not strictly adhere to the principles of a proper Bayesian posterior because the tempered likelihood, $P(D|\boldsymbol{\theta})^{\lambda}$ fails to meet the criteria of a proper likelihood function when $\lambda \neq 1$ (i.e., $\int P(D|\boldsymbol{\theta})^{\lambda} dD \neq 1$ when $\lambda \neq 1$). In this section, we aim to demonstrate that this tempered posterior effectively serves as a *proper Bayesian posterior* when considering a new combination of *likelihood and prior functions*. This strengthens our understanding of the CPE as a consequence of underfitting, resulting from poorly specified likelihood and prior functions. We will show that the presence of the CPE implies the existence of an alternative set of likelihood and prior functions, which are better specified in the sense that they define a Bayesian posterior that exhibits superior performance in both training and testing loss metrics.

Before delving into the description of the new likelihood and prior functions, it is essential to acknowledge a fundamental aspect. Given a labeled dataset $D = (\boldsymbol{X}, \boldsymbol{Y})$ and the conditional likelihood associated to a classification model, the application of Bayes' theorem naturally results in the following Bayesian posterior: $p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})$, where the prior over $\boldsymbol{\theta}$ is a *conditional prior* (Marek et al., 2024) that depends on the unlabelled training data $\boldsymbol{X}$. However, specifying $p(\boldsymbol{\theta}|\boldsymbol{X})$ for a complex model, like a deep neural network, poses a significant challenge. We are not aware of any work defining this kind of prior in the context of Bayesian deep learning. Therefore, for practical purposes, nearly all existing works (Wenzel et al., 2020; Fortuin et al., 2022) assume $\boldsymbol{\theta}$ to be independent of $\boldsymbol{X}$, resulting in the simplified expression $p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{\theta})$, where the prior over $\boldsymbol{\theta}$ is now an *unconditional prior*. The next result shows that the tempered posteriors are indeed proper Bayesian posteriors with different priors and posteriors.

Proposition 5. *For any given dataset $D = (\boldsymbol{X}, \boldsymbol{Y})$ and $\lambda > 0$, the tempered posterior defined in Equation 2 can be expressed as a Bayesian posterior with a new prior and likelihood function as follows:*

$$p^{\lambda}(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y}) \propto q(\boldsymbol{\theta}|\boldsymbol{X},\lambda) \prod_{(\boldsymbol{y},\boldsymbol{x})\in D} q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda), \tag{7}$$

where the new prior distribution $q(\boldsymbol{\theta}|\boldsymbol{X},\lambda)$ and likelihood function $q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda)$ are defined as:

$$q(\boldsymbol{\theta}|\boldsymbol{X},\lambda) \propto p(\boldsymbol{\theta}) \prod_{\boldsymbol{x}\in\boldsymbol{X}} \int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})^{\lambda} d\boldsymbol{y}, \qquad q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda) = \frac{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})^{\lambda}}{\int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})^{\lambda} d\boldsymbol{y}}. \tag{8}$$

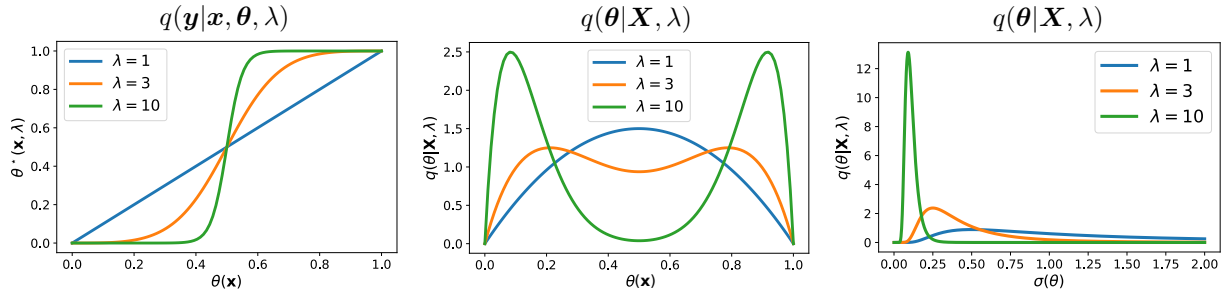


Figure 1: Illustration of the new likelihood $q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda)$ and priors $q(\boldsymbol{\theta}|\boldsymbol{X},\lambda)$, respectively, with the original likelihood in the form of Bernoulli distribution (left and middle); and the new prior obtained from an inverse-gamma prior and Gaussian likelihood (right) with a single observation. The transformation from $\theta(\boldsymbol{x})$ to $\theta^\star(\boldsymbol{x},\lambda) := \frac{\theta(\boldsymbol{x})^\lambda}{\theta(\boldsymbol{x})^\lambda+(1-\theta(\boldsymbol{x}))^\lambda}$ is shown at the left. The transformation of a Beta prior in a coin-flipping Beta-Binomial example with a single observation is shown in the middle.

Note that the new conditional likelihood and the new prior are both parametrized by the same $\lambda > 0$, and note that the prior only depends on the unlabelled training data $\boldsymbol{X}$ as in the general case described above. We now elaborate on how the new likelihood and prior distributions depend on λ. The next inequality, proved in Appendix B.2, shows that higher λ values induce likelihood distributions with lower aleatoric uncertainty or, equivalently, lower entropy, denoted as $H(q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda))$:

$$\nabla_\lambda H(q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda)) \leq 0 \quad \forall \lambda > 0\,. \tag{9}$$

We give two concrete examples with Gaussian and Bernoulli conditional likelihoods to illustrate the proposition. We also show that higher λ values result in likelihood distributions with lower aleatoric uncertainty. Let's first have a look at the new likelihood distributions.

Likelihood Examples *Consider the case where the original likelihood is Gaussian, defined as* $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \mathcal{N}(\mu(\boldsymbol{x},\boldsymbol{\theta}),\sigma(\boldsymbol{\theta})^2)$*, where the variance is input-independent, as typically seen in many regression problems. Then, following Equation 8, the new likelihood corresponds to a scaling in the variance, given by* $q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda) = \mathcal{N}(\mu(\boldsymbol{x},\boldsymbol{\theta}),\frac{\sigma(\boldsymbol{\theta})^2}{\lambda^2})$*. Thus, as* λ *increases, the tempered likelihood* $q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda)$ *induces a proper Gaussian likelihood with reduced variance, i.e., a new likelihood with lower aleatoric uncertainty.*

Consider the case of a binary classification problem where the original conditional likelihood is Bernoulli, defined as $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \theta(\boldsymbol{x})^y(1-\theta(\boldsymbol{x}))^{1-y}$ *with* $y \in \{0,1\}$ *and the input-dependent parameter function* $\theta(\boldsymbol{x}) \in [0,1]$*, which is usually implemented by a neural network with a softmax activation function in the last layer. Then, following Equation 8, the new conditional likelihood* $q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\lambda) = \theta^*(\boldsymbol{x},\lambda)^y(1-\theta^*(\boldsymbol{x},\lambda))^{1-y}$ *also follows a Bernoulli distribution with a different parameter function* $\theta^*(\boldsymbol{x},\lambda) = \frac{\theta(\boldsymbol{x})^\lambda}{\theta(\boldsymbol{x})^\lambda+(1-\theta(\boldsymbol{x}))^\lambda} \in [0,1]$*. The function* $\theta^*(\boldsymbol{x},\lambda)$ *is displayed in Figure 1 (left). When* λ *increases, the parameter function that defines the new Bernoulli likelihood becomes more extreme, resulting in a new likelihood with lower aleatoric uncertainty.*

On the other hand, according to Proposition 5, using the tempered posteriors implies implicitly using the prior $q(\boldsymbol{\theta}|\boldsymbol{X},\lambda)$. Such prior depends not only on the *unlabelled training data* $\boldsymbol{X}$, but also on the likelihood function defined by the probabilistic model family through the term $\int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})^\lambda d\boldsymbol{y}$ for $\boldsymbol{x} \in \boldsymbol{X}$. Thus, for models $\boldsymbol{\theta}$ that yield a large value for this term across most of the training data $\boldsymbol{x} \in \boldsymbol{X}$, the new prior will assign larger probability mass accordingly. We will showcase this effect in both binary classification and regression problems. Moreover, we will see how this new prior $q(\boldsymbol{\theta}|\boldsymbol{X},\lambda)$ with $\lambda > 1$ *favors those models within the model class that yield likelihoods with lower aleatoric uncertainty on the training data* X.

Prior Examples *Consider the case where the original likelihood is Gaussian, defined as* $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \mathcal{N}(\mu(\boldsymbol{x},\boldsymbol{\theta}),\sigma^2(\boldsymbol{\theta}))$*, where the variance is input-independent, as typically seen in many regression problems. A common parametrization in this case is to have* $\boldsymbol{\theta} = (\boldsymbol{w},\gamma)$*, where* $\boldsymbol{w}$ *refer to the weights of the neural network defining the function* $\mu(\boldsymbol{x},\boldsymbol{\theta})$ *and* $\gamma > 0$ *is a free parameter encoding the variance of the Gaussian likelihood,*

$\sigma^2(\boldsymbol{\theta}) = \gamma$. *And* $p(\boldsymbol{\theta})$ *is then defined as* $p(\boldsymbol{\theta}) = p(\boldsymbol{w})p(\gamma)$, *where* $p(\boldsymbol{w})$ *is usually a Gaussian distribution with a diagonal covariance matrix, and* $p(\gamma)$ *is usually defined in terms of a inverse-gamma distribution. Following Equation 8, the new prior would be then expressed as* $q(\boldsymbol{\theta}|\boldsymbol{X}, \lambda) = q(\boldsymbol{w}|\boldsymbol{X}, \lambda)q(\gamma|\boldsymbol{X}, \lambda)$, *where each term is defined:*

$$q(\boldsymbol{w}|\boldsymbol{X}, \lambda) = p(\boldsymbol{w}) \qquad q(\gamma|\boldsymbol{X}, \lambda) \propto p(\gamma)/\gamma^{n(\lambda-1)} \,.$$

Figure 1 (right) plots the density of $q(\gamma|\boldsymbol{X}, \lambda)$ *for* $n = 1$ *and several* $\lambda > 1$ *values when* $p(\gamma)$ *is an inverse-gamma prior. For larger* λ *values, this new prior will assign more probability mass to models defining a likelihood with smaller variance or, equivalently, smaller aleatoric uncertainty.*

Consider another case where the original conditional likelihood is Bernoulli, defined as $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \theta(\boldsymbol{x})^y(1 - \theta(\boldsymbol{x}))^{1-y}$ *with* $y \in \{0, 1\}$ *and* $\theta(\boldsymbol{x}) \in [0, 1]$, *as commonly used in binary classification problems. Then, following Equation 8, the new prior is expressed as*

$$q(\boldsymbol{\theta}|\boldsymbol{X}, \lambda) \propto p(\boldsymbol{\theta}) \prod_{\boldsymbol{x} \in \boldsymbol{X}} \left(\theta(\boldsymbol{x})^\lambda + (1 - \theta(\boldsymbol{x}))^\lambda \right) .$$

Figure 1 (middle) illustrates this prior for a Beta-Binomial model under $\lambda \geq 1$, *assuming that there is a single training sample. This figure shows that as* λ *increases, the new prior assigns more probability mass to models where* $\theta(\boldsymbol{x})$ *is close to either* 1 *or* 0. *In other words, this new prior assigns more probability mass to models that assign more extreme probabilities to the training data (i.e., models with lower aleatoric uncertainty). Note that the prior does not consider how accurately these models classify the training data, but only the extremity of the probabilities assigned to the training data.*

In this context, employing tempered posteriors seamlessly fits within a Bayesian framework. Tuning the hyperparameter λ resembles an empirical Bayesian technique. This method streamlines and enriches the utilization of diverse likelihood and prior functions. Furthermore, using likelihood/priors with $\lambda > 1$, the tempered Bayesian posterior enhances the fit to training data by allocating more probability to models with lower aleatoric uncertainty. Consequently, tempered posteriors provide a simple, computationally efficient, and theoretically sound approach to mitigate the underfitting problem commonly encountered in contemporary Bayesian deep learning methods.

Insight 3. *If, in comparison to the original Bayesian posterior, we observe a decrease in both training and test loss when using tempered posteriors with* $\lambda > 1$, *it implies that the new likelihood and priors implicitly defined in Equation 8 are better specified. This alignment with the underlying data distribution enables the tempered posterior to better capture the data-generating distribution, leading to enhanced model performance on both training and unseen test data.*

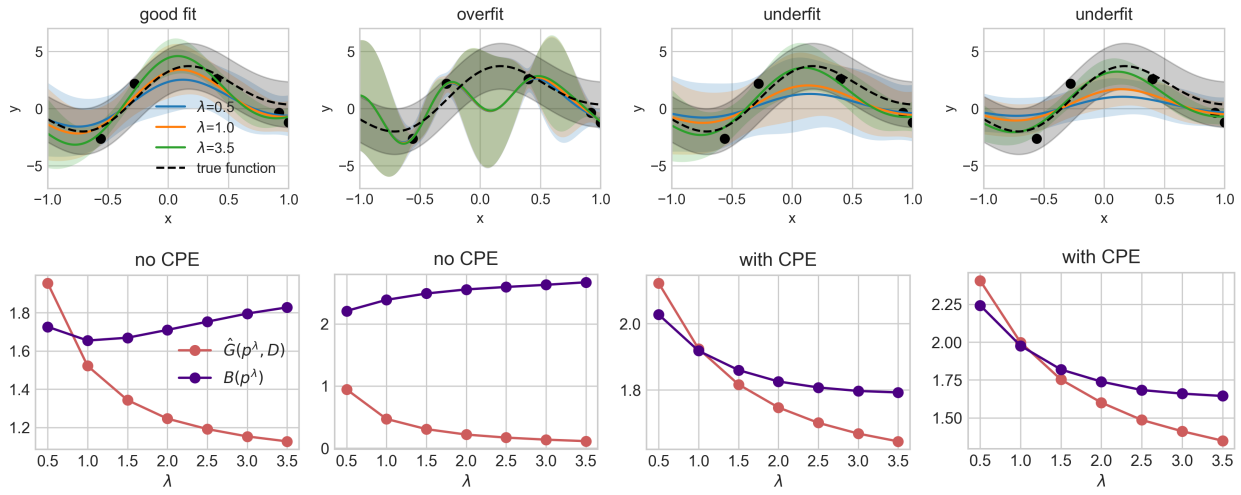
Generalized ELBOs are also proper ELBOs Generalized ELBOs, characterized by scaling the KL divergence term using a hyper-parameter λ, have found widespread application in many studies (Wenzel et al., 2020). This popularity stems from the demonstrated ability of adjusting λ to improve the predictive accuracy of variational approximations:

$$q_\lambda^\star := \arg\min_{r \in \Pi} \mathbb{E}_r[-\ln p(D|\boldsymbol{\theta})] + \frac{1}{\lambda} \mathrm{KL}(r(\boldsymbol{\theta}), p(\boldsymbol{\theta})) \,, \tag{10}$$

where Π defines the variational family. Critics have pointed out a flaw in the above generalized ELBO when λ deviates from 1, as it no longer functions as a true lower bound for the marginal likelihood. But Proposition 5 can be used to justify that such a variational posterior $q_\lambda^\star$ still emerges from minimizing a valid ELBO. Specifically, it is constructed based on the revised likelihood and prior functions as follows:

$$q_\lambda^\star = \arg\min_{r \in \Pi} \mathbb{E}_r[-\ln q(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}, \lambda)] + \mathrm{KL}(r(\boldsymbol{\theta}), q(\boldsymbol{\theta}|\boldsymbol{X}, \lambda)) \,. \tag{11}$$

Consequently, this analysis shows that using generalized ELBOs as Equation 10 perfectly adheres to variational and Bayesian principles.



(a) No misspecification (b) Misspecified likelihood I (c) Misspecified likelihood II (d) Misspecified prior

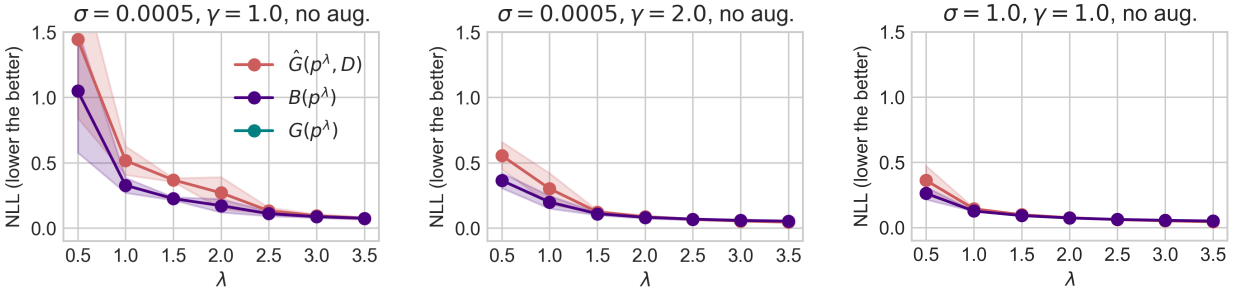
Figure 2: **1. The CPE occurs in Bayesian linear regression with exact inference. 2. Model misspecification can lead to overfitting and to a "warm" posterior effect (WPE).** Every column displays a specific setting, as indicated in the caption. The first row shows exact Bayesian posterior predictive fits for three different values of the tempering parameter λ. The second row shows the Gibbs loss $\hat{G}(p^\lambda, D)$ (aka training loss) and the Bayes loss $B(p^\lambda)$ (aka testing loss) with respect to λ. The experimental details are given in Appendix D.

5 Likelihood Misspecification, Prior Misspecification and the CPE

In light of the theoretical characterization of the CPE given above in terms of underfitting, we will revisit the main arguments by previous works in relation to CPE, and we will show how we can provide a new and more nuanced perspective on the underlying implications of the presence of the CPE. For now, we set aside data augmentation, which will be specifically treated in the next section.

CPE, approximate inference, and NNs: As mentioned in the introduction, several works have discussed that CPE is an artifact of inappropriate approximate inference methods, especially in the context of the highly complex posterior that emerge from neural networks (Wenzel et al., 2020). The main reasoning is that if the approximate inference method is accurate enough, the CPE disappears (Izmailov et al., 2021). However, Proposition 2 shows that when λ is made larger than 1, the *training loss* of the exact Bayesian posterior decreases; if the *test loss* decreases too, the exact Bayesian posterior underfits. It means that even if the inference method is accurate, we can still observe the CPE due to underfitting. In fact, Figure 2 shows examples of a Bayesian linear regression model learned on synthetic data. Here, the exact Bayesian posterior can be computed, and it is clear from Figures 2c and 2d that the CPE can occur in Bayesian linear regression with exact inference. Although simple, the setting is articulated specifically to mimic the classification tasks using BNNs where CPE was observed. In particular, the linear model has more parameters than observations (i.e. it's overparameterized).

Model misspecification, CPE, and underfitting: Prior and/or likelihood misspecification can lead Bayesian methods to both underfitting and overfitting, as widely discussed in the literature (Domingos, 2000; Immer et al., 2021; Kapoor et al., 2022). We illustrate this using a Bayesian linear regression model: Figures 2c and 2d show how the Bayesian posterior underfits due to likelihood and prior misspecification, respectively. On the other hand, Figure 2b showcases a scenario where likelihood misspecification can perfectly lead to overfitting as well, giving rise to what we term a "warm" posterior effect (WPE), i.e., there exists other posteriors (p^λ with $\lambda < 1$) with lower testing loss, which, at the same time, also have higher training loss due to Proposition 2. As a result, to describe CPE merely as a model misspecification issue without acknowledging underfitting offers a narrow interpretation of the problem.

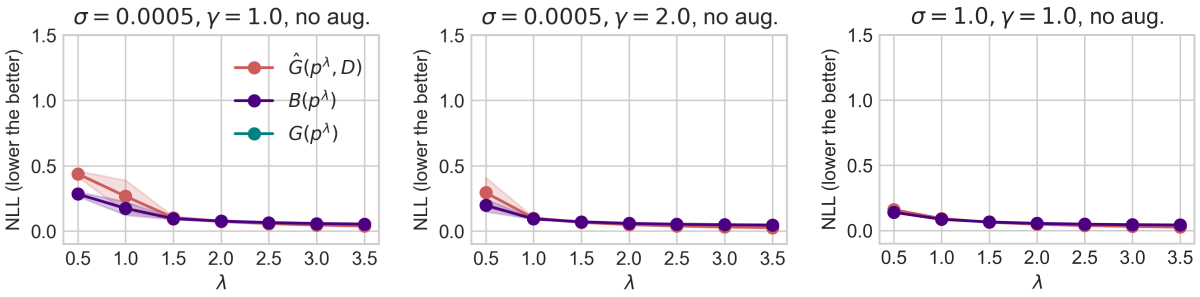


(a) Narrow prior and standard softmax (b) Narrow prior and tempered softmax (c) Standard prior and standard softmax

Figure 3: **Experimental illustrations for the arguments in Section 5 using small CNN via SGLD on MNIST. We show similar results on Fashion-MNIST with small CNN and CIFAR-10(0) with ResNet-18 in Appendix E.** Figures 3a to 3c illustrate the arguments in Section 5. Figure 3c uses the standard prior ($\sigma = 1$) and the standard softmax ($\gamma = 1$) for the likelihood without applying DA. Figure 3a follows a similar setup except for using a narrow prior. Figure 3b uses a narrow prior as in Figure 3a but with a tempered softmax that results in a lower aleatoric uncertainty. We report the training loss $\hat{G}(p^\lambda, D)$ and the testing losses $B(p^\lambda)$ and $G(p^\lambda)$ from 10 samples of the small Convolutional neural network (CNN) via Stochastic Gradient Langevin Dynamics (SGLD). We show the mean and standard error across three different seeds. For additional experimental details, please refer to Appendix E.

Note that the above examples, corresponding to the Gaussian case in the likelihood example and prior example, provide a perfect and intuitive demonstration and explanation to our Proposition 5: when CPE shows up, tuning λ is equivalent to finding a Bayesian posterior with a less misspecified likelihood and prior. We first expand the discussion on the likelihood and then the prior. For the likelihood, we go to Figure 2c, where the Gaussian likelihood model has a larger variance than the true data-generating process. Here, the CPE is taking place since increasing λ results in a likelihood model with lower aleatoric uncertainty (Equation 9), bringing a less misspecified model (the Gaussian case in the likelihood example) and, thus, better performance. More precisely, the new likelihood model has a smaller variance (divided by λ), and is, accordingly, much closer to the true data-generating process. The opposite can be seen for Figure 2b, where the Gaussian likelihood model has a lower variance than the true data-generating process and the WPE occurs. This reveals the nature of CPE under likelihood misspecification: the performance improves because the tempered posteriors implicitly assume a better specified likelihood. For the prior, we compare Figure 2a and 2d. Since our likelihood variance is lower than 1, increasing λ is the same as putting more probability mass in likelihood models with lower uncertainty (shown in the Gaussian case in the prior example). Since both the likelihood and prior are well-specified in Figure 2a, the "new prior" causes misspecification and CPE does not occur. Conversely, this "new prior" places more mass on the less uncertain likelihood in Figure 2d. In consequence, it mitigates the misspecification and brings the CPE.

The likelihood misspecification argument: Likelihood misspecification has also been identified as a cause of CPE, especially in cases where the dataset has been *curated* (Aitchison, 2021; Kapoor et al., 2022). Data curation often involves carefully selecting samples and labels to improve the quality of the dataset. As a result, the curated data-generating distribution typically presents very low aleatoric uncertainty, meaning that $\nu(\boldsymbol{y}|\boldsymbol{x})$ usually takes values very close to either 1 or 0. However, the standard likelihoods used in deep learning, like softmax or sigmoid, implicitly assume a higher level of aleatoric uncertainty in the data (Aitchison, 2021; Kapoor et al., 2022). Therefore, their use in curated datasets, that exhibit low uncertainty, made them misspecified (Kapoor et al., 2022; Fortuin et al., 2022). To address this issue, alternative likelihood functions like the Noisy-Dirichlet model (Kapoor et al., 2022, Section 4) have been proposed, which better align with the characteristics of the curated data. On the other hand, introducing noise labels also alleviates the CPE, as demonstrated in Aitchison (2021, Figure 7). By introducing noise labels, we intentionally increase aleatoric uncertainty in the data-generating distribution, which aligns better with the high aleatoric uncertainty assumed by the standard Bayesian deep networks (Kapoor et al., 2022). Consequently, according to these works, the CPE can be strongly alleviated when the likelihood misspecification is addressed.



(a) Narrow prior and standard softmax (b) Narrow prior and tempered softmax (c) Standard prior and standard softmax

Figure 4: **Experimental illustrations for the arguments in Section 5 using large CNN via SGLD on MNIST. We show similar results on Fashion-MNIST with large CNN and CIFAR-10(0) with ResNet-50 in Appendix E.** The experiment setup is similar to the setups in Figure 3 but with a large CNN. Please refer to Appendix E for further details on the model.

Our theoretical analysis aligns with these findings. However, we can also add that the key underlying cause of CPE under data curation is the underfitting induced by likelihood misspecification. Because fitting low aleatoric uncertainty data-generating distributions, e.g., $\nu(y|\boldsymbol{x}) \in \{0.01, 0.99\}$, with high aleatoric uncertainty likelihood functions e.g., $p(y|\boldsymbol{x}, \boldsymbol{\theta}) \in [0.2, 0.8]$, induces underfitting. The presence of underfitting is not mentioned at all by any of these previous works (Aitchison, 2021; Kapoor et al., 2022). On top of that, using Proposition 5 and Equation 9, our work explains why the likelihood implicitly used by the tempered posterior with $\lambda > 1$ provides better generalizaton performance. Because, in this case, we are using a likelihood $q(\boldsymbol{\theta}|\boldsymbol{X}, \lambda)$ (Equation 8) with lower aleatoric-uncertainty, which better aligns with the low aleatoric-uncertainty data-generating distribution induced by curated datasets, thus reducing the degree of model misspecification.

Figures 3a and 3b, along with Figures 4a and 4b, illustrate this point through a regular multi-class classification task on a curated benchmark dataset. Both scenarios utilize the same narrow prior. The distinction in Figure 3b lies in the adoption of a tempered softmax likelihood, defined as $p(y|x, \theta) = (1 + \exp(-\gamma \operatorname{logits}(x, \theta)))^{-1}$, with $\gamma = 2$, compared to $\gamma = 1$ in Figure 3a. This tempered softmax likelihood, more closely aligned with the dataset's low aleatoric uncertainty as outlined by (Guo et al., 2017), leads to a reduced incidence of CPE in Figure 3b compared to Figure 3a. From the perspective of Proposition 5 and specifically Equation 9, the intrinsic lower aleatoric uncertainty of the likelihood used in Figure 3b (softmax with $\gamma = 2$) makes the potential for improvement through increasing λ somewhat limited, resulting in a less pronounced CPE compared to Figure 3a. It is, however, important to highlight the critical interaction between the likelihood and the prior, as we dicuss next.

The prior misspecification argument: As highlighted in previous works, such as in Wenzel et al. (2020); Fortuin et al. (2022), isotropic Gaussian priors are commonly chosen in modern Bayesian neural networks for the sake of tractability in approximate Bayesian inference rather than chosen based on their alignment with our actual beliefs. Given that the presence of the CPE implies that either the likelihood and/or the prior are misspecified, and given that neural networks define highly flexible likelihood functions, there are strong reasons for thinking these commonly used priors are misspecified. Notably, the experiments conducted by Fortuin et al. (2022) demonstrate that the CPE can be mitigated in fully connected neural networks when using heavy-tailed prior distributions that better capture the weight characteristics typically observed in such networks. However, such priors were found to be ineffective in addressing the CPE in convolutional neural networks (Fortuin et al., 2022), indicating the challenges involved in designing effective Bayesian priors within this context.

Our theoretical analysis provides a deeper insight into these observations. As mentioned in Section 3, the absence of underfitting means the absence of CPE. This implies that, assuming that large neural networks defines a sufficiently flexible likelihood function, underfitting occurs due to the prior's failure to allocate enough probability to models that both fit the training data well and exhibit good generalization capabilities, essentially due to excessive regularization by the prior $p(\boldsymbol{\theta})$. As detailed in Section 4, employing tempered

posteriors with $\lambda > 1$ effectively defines a conditional prior $q(\boldsymbol{\theta}|\boldsymbol{X}, \lambda)$ (see Equation 8 and the prior example) that favors models with lower aleatoric-uncertainty. Such models aligns better with the training data. Hence, the conditional prior $q(\boldsymbol{\theta}|\boldsymbol{X}, \lambda)$ with $\lambda > 1$ can be considered better specified that the original prior $p(\boldsymbol{\theta})$ because it leads to a tempered Bayesian posterior that not only more accurately represents the training data but also improves generalization.

Figures 3a and 3c exemplify this situation: when the prior is too narrow ($\sigma = 0.0005$) and induces a very strong regularization, the resulting posterior severely underfits the training data and leads to a high empirical Gibbs loss that deviates significantly from zero (Figure 3a). We also observe a strong CPE in this case, i.e., the Bayes loss $B(p^\lambda)$ significantly decreases when $\lambda > 1$. However, by using a flatter prior (Figure 3c) there is less underfitting, and the CPE is considerably diminished. Using Proposition 5 and the above discussion, we can see that in the former case, the new prior $q(\boldsymbol{\theta}|\boldsymbol{X}, \lambda)$ with $\lambda > 1$ would place much more probability mass in models with lower-aleatoric uncertainty than the narrow prior and, hence, strongly alleviating underfitting. In the second case, since the flatter prior already has a more diffuse probability mass over the model class, the probability mass it can transfer to those models with lower aleatoric uncertainty will be lesser than that from a narrower prior, hence resulting in a milder CPE.

Model size, CPE, and underfitting: Larger models have the capacity to fit data more effectively, while smaller models are more likely to underfit. As we have argued that if there is no underfitting, there is no CPE, we expect that the size of the model has an impact on the strength of CPE as well, as demonstrated in Figure 3 and Figure 4. Specifically, in our experiments presented in Figure 3, we utilize a relatively small convolutional neural network (CNN), which has a more pronounced underfitting behavior, and this indeed corresponds to a stronger CPE. On the other hand, we employ a larger CNN in Figure 4, which has less underfitting, and we see the CPE is strongly alleviated.

6 Data Augmentation (DA) and the CPE

Machine learning is applied to many different fields and problems, and in many of them, the data-generating distribution is known to have properties that can be exploited to artificially generate new data samples (Shorten & Khoshgoftaar, 2019). This is commonly known as *data augmentation (DA)* and relies on the property that for a given set of transformations T, the data-generating distribution satisfies $\nu(\boldsymbol{y}|\boldsymbol{x}) = \nu(\boldsymbol{y}|t(\boldsymbol{x}))$ for all $t \in T$. In practice, not all the transformations are applied to every single data. Instead, a probability distribution (usually uniform) μ_T is defined over T, and augmented samples are drawn accordingly. As argued in Nabarro et al. (2022), the use of data augmentation when training Bayesian neural networks implicitly targets the following (pseudo) log-likelihood, denoted $\hat{L}_{\text{DA}}(D, \boldsymbol{\theta})$ and defined as

$$\hat{L}_{\text{DA}}(D, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in [n]} \mathbb{E}_{t \sim \mu_T} \left[-\ln p(\boldsymbol{y}_i | t(\boldsymbol{x}_i), \boldsymbol{\theta}) \right] , \tag{12}$$

where data augmentation provides unbiased estimates of the expectation under the set of transformations using *Monte Carlo samples* (i.e., random data augmentations).

Although some argue that this data-augmented *(pseudo) log-likelihood* "does not have a clean interpretation as a valid likelihood function" (Wenzel et al., 2020; Izmailov et al., 2021), we do not need to enter into this discussion to understand why the CPE emerges when using the generalized Bayes posterior (Bissiri et al., 2016) associated to this *(pseudo) log-likelihood*, which is the main goal of this section. We call this posterior the DA-tempered posterior and is denoted by $p^\lambda_{\text{DA}}(\boldsymbol{\theta}|D)$. The DA-tempered posterior can be expressed as the global minimizer of the following learning objective,

$$p^\lambda_{\text{DA}}(\boldsymbol{\theta}|D) = \arg\min_{\rho} \mathbb{E}_\rho[n\hat{L}_{\text{DA}}(D, \boldsymbol{\theta})] + \frac{1}{\lambda} \text{KL}(\rho(\boldsymbol{\theta}|D), p(\boldsymbol{\theta})) . \tag{13}$$

This is similar to Eq. 3 but now using $\hat{L}_{\text{DA}}(D, \boldsymbol{\theta})$ instead of $\hat{L}(D, \boldsymbol{\theta})$, where we recall the notation $\hat{L}(D, \boldsymbol{\theta}) = -\frac{1}{n} \ln p(D|\boldsymbol{\theta})$. Hence, the resulting DA-tempered posterior is given by $p^\lambda_{\text{DA}}(\boldsymbol{\theta}|D) \propto e^{-n\lambda \hat{L}_{\text{DA}}(D, \boldsymbol{\theta})} p(\boldsymbol{\theta})$. In comparison, the tempered posterior $p^\lambda(\boldsymbol{\theta}|D)$ in Eq. 2 can be similarly expressed as $e^{-n\lambda \hat{L}(D, \boldsymbol{\theta})} p(\boldsymbol{\theta})$.

There is large empirical evidence that DA induces a stronger CPE (Wenzel et al., 2020; Izmailov et al., 2021; Fortuin et al., 2022). Indeed, many of these studies show that if CPE is not present in our Bayesian learning settings, using DA makes it appear. According to our previous analysis, this means that the use of DA induces a stronger underfitting. To understand why this is case, we will take a step back and begin analyzing the impact of DA in the so-called Gibbs loss of the DA-Bayesian posterior $p_{\text{DA}}^{\lambda=1}$ rather than the Bayes loss, as this will help us in understanding this puzzling phenomenon.

6.1 Data Augmentation and CPE on the Gibbs loss

The expected Gibbs loss of a given posterior ρ, denoted $G(\rho)$, is a commonly used metric in the theoretical analysis of the *generalization performance* of Bayesian methods (Germain et al., 2016; Masegosa, 2020). The Gibbs loss represents the average of the expected log-loss of individual models under the posterior ρ, that is,

$$G(\rho) = \mathbb{E}_{\rho}[L(\boldsymbol{\theta})] = \mathbb{E}_{\rho}[\mathbb{E}_{\nu}[-\ln[p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})]]\,.$$

In fact, Jensen's inequality confirms that the expected Gibbs loss serves as an upper bound for the Bayes loss, i.e., $G(\rho) \geq B(\rho)$. This property supports the expected Gibbs loss to act as a proxy of the Bayes loss, which justifies its usage in gaining insights into how DA impacts the CPE.

We will now study whether data augmentation can cause a CPE on the Gibb loss. In other words, we will examine whether increasing the parameter λ of the DA-tempered posterior leads to a reduction in the Gibbs loss. This can be formalized by extending Definition 1 to the expected Gibbs loss by considering its gradient $\nabla_{\lambda} G(p^{\lambda})$ at $\lambda = 1$, which can be represented as follows:

$$\nabla_{\lambda} G(p^{\lambda})_{|\lambda=1} = -\text{COV}_{p^{\lambda=1}}\left(n\hat{L}(D,\boldsymbol{\theta}), L(\boldsymbol{\theta})\right). \tag{14}$$

Where $\text{COV}(X,Y)$ denotes the covariance of X and Y. Again, due to the page limit, we postpone the necessary proofs in this section to Appendix C.

With this extended definition, if Eq. 14 is negative, we can infer the presence of CPE for the Gibbs loss as well. Based on this, we say that DA induces a stronger CPE if the gradient of the expected Gibbs loss for the DA-tempered posterior exhibits a more negative trend at $\lambda = 1$, i.e., if $\nabla_{\lambda} G(p_{\text{DA}}^{\lambda})_{|\lambda=1} < \nabla_{\lambda} G(p^{\lambda})_{|\lambda=1}$.This condition can be equivalently stated as

$$\text{COV}_{p_{\text{DA}}^{\lambda=1}}\left(n\hat{L}_{\text{DA}}(D,\boldsymbol{\theta}), L(\boldsymbol{\theta})\right) > \text{COV}_{p^{\lambda=1}}\left(n\hat{L}(D,\boldsymbol{\theta}), L(\boldsymbol{\theta})\right) > 0\,. \tag{15}$$

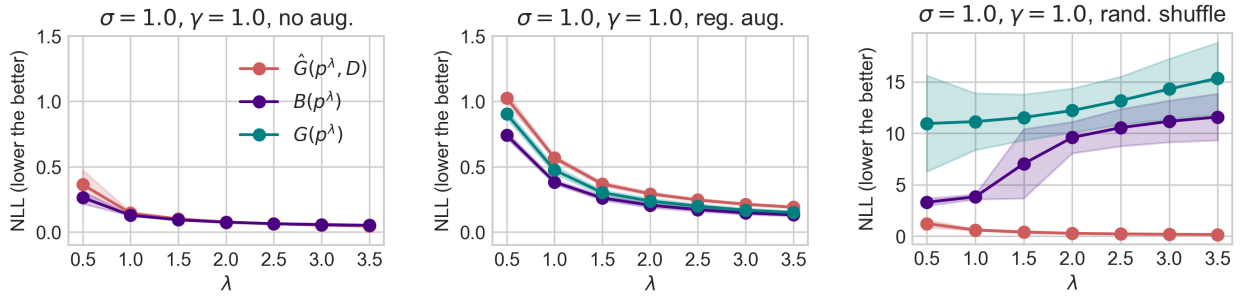
The inequality presented above helps characterize and understand the occurrence of a stronger CPE when using DA. A stronger CPE arises if the expected Gibbs loss of a model $L(\boldsymbol{\theta})$ is more *correlated* with the empirical Gibbs loss of this model on the augmented training dataset $\hat{L}_{\text{DA}}(D,\boldsymbol{\theta})$ than on the non-augmented dataset $\hat{L}(D,\boldsymbol{\theta})$. This observation suggests that, if we empirically observe that the CPE is stronger when using an augmented dataset, the set of transformations $\mathcal{T}$ used to generate the augmented dataset are introducing *valuable information* about the data-generating process.

Figure 5 clearly illustrates such situations. Figure 5b shows that, compared to Figure 5a, the standard DA, which makes use of the invariances inherent in the data-generating distribution, induces a CPE on the Gibbs loss. Thus, the condition in Eq. 15 holds by definition. On the other hand, Figure 5c uses a fabricated DA, where the same permutation is applied to the pixels of the images in the training dataset, which destroys low-level features present in the data-generating distribution. In this case, the gradient of the Gibb loss is positive, and Eq. 15 holds in the opposite direction. These findings align perfectly with the explanations provided above, showing that DA induces a stronger underfitting.

6.2 Data Augmentation and CPE on the Bayes loss

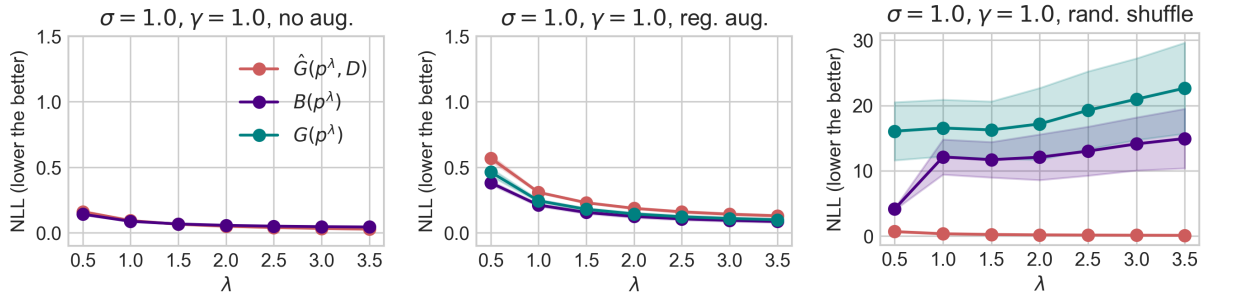
Now, we step aside of the Gibbs loss and focus back to the Bayes loss. The gradient of the Bayes loss at $\lambda = 1$ can also be written as,

$$\nabla_{\lambda} B(p^{\lambda})_{|\lambda=1} = -\text{COV}_{p^{\lambda=1}}\left(n\hat{L}(D,\boldsymbol{\theta}), S_{p^{\lambda=1}}(\boldsymbol{\theta})\right), \tag{16}$$



(a) Standard prior and standard softmax (b) Random crop and horizontal flip (c) Pixels randomly shuffled

Figure 5: **Experimental illustrations for the arguments in Section 6 using small CNN via SGLD on MNIST. We show similar results on Fashion-MNIST with small CNN and CIFAR-10(0) with ResNet-18 in Appendix E.** Figures 5a to 5c illustrate the arguments in Section 6. Figure 5a uses the standard prior ($\sigma = 1$) and the standard softmax ($\gamma = 1$) for the likelihood without applying DA. Figure 5b follows the setup as in Figure 5a but with standard DA applied, while Figure 5c uses fabricated DA. We report the training loss $\hat{G}(p^\lambda, D)$ and the testing losses $B(p^\lambda)$ and $G(p^\lambda)$ from 10 samples of the small Convolutional neural network (CNN) via Stochastic Gradient Langevin Dynamics (SGLD). We show the mean and standard error across three different seeds. For additional experimental details, please refer to Appendix E.



(a) Standard prior and standard softmax (b) Random crop and horizontal flip (c) Pixels randomly shuffled

Figure 6: **Experimental illustrations for the arguments in Section 6 using large CNN via SGLD on MNIST. We show similar results on Fashion-MNIST with large CNN and CIFAR-10(0) with ResNet-50 in Appendix E.** The experiment setup is similar to the setups in Figure 5 but with a large CNN. Please refer to Appendix E for further details on the model.

where for any posterior ρ, $S_\rho(\boldsymbol{\theta})$ is a (negative) performance measure defined as

$$S_\rho(\boldsymbol{\theta}) = -\mathbb{E}_\nu \left[\frac{p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})}{\mathbb{E}_\rho[p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})]} \right] . \tag{17}$$

This function measures the relative performance of a model parameterized by $\boldsymbol{\theta}$ compared to the average performance of the models weighted by ρ. Such measure is conducted on samples from the data-generating distribution $\nu(\boldsymbol{y}, \boldsymbol{x})$. Specifically, if the model $\boldsymbol{\theta}$ outperforms the average, we have $S_\rho(\boldsymbol{\theta}) < -1$, and if the model performs worse than the average, we have $S_p(\boldsymbol{\theta}) > -1$ (i.e., the lower the better). The derivations of the above equations are given in Appendix C.

According to Definition 1 and Eq. 16, DA will induce a stronger CPE if and only if the following condition is satisfied:

$$\mathrm{COV}_{p_{\mathrm{DA}}^{\lambda=1}} \left(n\hat{L}_{\mathrm{DA}}(D, \boldsymbol{\theta}), S_{p_{\mathrm{DA}}^{\lambda=1}}(\boldsymbol{\theta}) \right) > \mathrm{COV}_{p^{\lambda=1}} \left(n\hat{L}(D, \boldsymbol{\theta}), S_{p^{\lambda=1}}(\boldsymbol{\theta}) \right) . \tag{18}$$

The previous analysis on the Gibbs loss remains applicable in this context, with the use of $S_\rho(\boldsymbol{\theta})$ as a metric for the expected performance on the true data-generating distribution instead of $L(\boldsymbol{\theta})$. While these metrics

are slightly different, it is reasonable to assume that the same arguments we presented to explain the CPE under data augmentation for the Gibbs loss also apply here. The theoretical analysis aligns with the behavior of the Bayes loss as depicted in Figure 5.

Finally, comparing Figure 5b with Figure 6b, we also notice that using a larger neural network enables us to mitigate the CPE because we reduce the underfitting introduced by DA.

Related work of the data augmentation argument. The relation between data augmentation and CPE is an active topic of discussion (Wenzel et al., 2020; Izmailov et al., 2021; Noci et al., 2021; Nabarro et al., 2022). Some studies suggest that CPE is an artifact of DA because turning off data augmentation is enough to eliminate the CPE (Izmailov et al., 2021; Fortuin et al., 2022). Our study shows that this is *much more* than an artifact, as also argued in Nabarro et al. (2022). As discussed, the (pseudo) log-likelihood induced by standard DA is a better proxy of the expected log-loss, in the precise sense given by Eq. 15 and Eq. 18.

Other works argue that, when using DA, we are not using a proper likelihood function (Izmailov et al., 2021), and that could be problem. Recent works (Nabarro et al., 2022) have developed principle likelihood functions that integrate DA-based approaches, hoping that this will remove CPE. But they find that CPE still persist. Another widely accepted viewpoint regarding the interplay between the CPE and DA is that DA increases the effective sample size (Izmailov et al., 2021; Noci et al., 2021), "intuitively, data augmentation increases the amount of data observed by the model, and should lead to higher posterior contraction" (Izmailov et al., 2021).

Our analysis provides a more nuance understanding of this interplay between CPE and DA. First, we show that, when the augmented data provide extra information about the data-generating process, there is a stronger CPE, as shown in Equations 15 and 18. This, in turn, leads to higher posterior concentration. But, we also show that higher posterior concentration in the context of non-meaningful DA does not improve performance; as discussed before, Figure 5c illustrates this situation. Using the analysis given in Section 4, we can also add that tempering the posterior under DA is again a way to define alternative Bayesian posteriors that addresses this stronger underfitting, i.e., they better fit the training data and improve generalization.

7 Conclusions

Our research makes several contributions toward understanding the cold posterior effect (CPE) and its implications for Bayesian deep learning. Firstly, we theoretically demonstrate that the presence of the CPE implies that the Bayesian posterior is underfitting. And, secondly, we show that any tempered posterior can be considered as a proper Bayesian posterior with an alternative likelihood and prior distribution jointly parametrized by T. Hence, this work shows that fine-tuning the temperature parameter T serves as an effective and theoretically sound mechanism to address the underfitting of the Bayesian posterior. Finally, our analysis in Section 6 unveils that data augmentation exacerbates the cold posterior effect (CPE) by intensifying the degree of underfitting. This is attributed to fact that the augmented data supplies richer and more reliable information, thereby enhancing the capacity for fitting.

Overall, our theoretical analysis underscores the significance of the CPE as an indicator of underfitting within the Bayesian framework and promotes the fine-tuning of the temperature T in tempered posteriors as a principled approach to mitigate this issue. Furthermore, by dissecting the nature of the CPE and its effect with Bayesian principle, our work aims to resolve ongoing debates and clarify the role of cold posteriors in enhancing the predictive performance of Bayesian deep learning models.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 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. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

Andrew Barron and Thomas Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 1991.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

François Chollet et al. Keras. https://keras.io, 2015.

Joshua V Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A Saurous. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

Pedro Domingos. Bayesian averaging of classifiers and the overfitting problem. In *ICML*, 2000.

Vincent Fortuin, Adrià Garriga-Alonso, Sebastian W Ober, Florian Wenzel, Gunnar Ratsch, Richard E Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pp. 1884–1892, 2016.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 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*, pp. 703–711. PMLR, 2021.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Sanyam Kapoor, Wesley J Maddox, Pavel Izmailov, and Andrew G Wilson. On uncertainty, tempering, and data augmentation in bayesian classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Martin Marek, Brooks Paige, and Pavel Izmailov. Can a confident prior replace a cold posterior? *arXiv preprint arXiv:2403.01272*, 2024.

Andrés R Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 2020.

Warren R. Morningstar, Alex Alemi, and Joshua V. Dillon. Pacm-bayes: Narrowing the empirical risk gap in the misspecified bayesian regime. In *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8270–8298. PMLR, 2022.

Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. Data augmentation in bayesian neural networks and the cold posterior effect. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

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 (NeurIPS)*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, pp. 681–688. Omnipress, 2011.

Florian Wenzel, Kevin Roth, Bastiaan 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? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. In *International Conference on Learning Representations*, 2019.

Tong Zhang. From ε-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *Annals of Statistics*, 2006.

Yijie Zhang and Eric Nalisnick. On the inconsistency of bayesian inference for misspecified neural networks. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021. URL https://openreview.net/forum?id=vfZtrgabCr.