



Ph.D. Thesis

Yijie Zhang

On Cold Posteriors of Probabilistic Neural Networks

Understanding the Cold Posterior Effect and A New Way to Learn Cold Posteriors with Tight Generalization Guarantees

Advisors: Christian Igel and Sadegh Talebi

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen on September 3rd 2024.

Abstract

Bayesian inference provides a principled probabilistic framework for quantifying uncertainty by updating beliefs based on prior knowledge and observed data through Bayes’ theorem. In Bayesian deep learning, neural network weights are treated as random variables with prior distributions, allowing for a probabilistic interpretation and quantification of predictive uncertainty. However, Bayesian methods lack theoretical generalization guarantees for unseen data. PAC-Bayesian analysis addresses this limitation by offering a frequentist framework to derive generalization bounds for randomized predictors, thereby certifying the reliability of Bayesian methods in machine learning.

Temperature T , or inverse-temperature $\lambda = \frac{1}{T}$, originally from statistical mechanics in physics, naturally arises in various areas of statistical inference, including Bayesian inference and PAC-Bayesian analysis. In Bayesian inference, when $T < 1$ (“cold” posteriors), the likelihood is up-weighted, resulting in a sharper posterior distribution. Conversely, when $T > 1$ (“warm” posteriors), the likelihood is down-weighted, leading to a more diffuse posterior distribution. By balancing the influence of observed data and prior regularization, temperature adjustments can address issues of underfitting or overfitting in Bayesian models, bringing improved predictive performance.

We begin by investigating the cold posterior effect (CPE) in Bayesian deep learning. We demonstrate that misspecification leads to CPE only when the Bayesian posterior underfits. Additionally, we show that tempered posteriors are valid Bayesian posteriors corresponding to different combinations of likelihoods and priors parameterized by temperature T . Fine-tuning T thus allows for the selection of alternative Bayesian posteriors with less misspecified likelihood and prior distributions.

Next, we introduce an effective PAC-Bayesian procedure, Recursive PAC-Bayes (RPB), that enables sequential posterior updates without information loss. This method is based on a novel decomposition of the expected loss of randomized classifiers, which reinterprets the posterior loss as an excess loss relative to a scaled-down prior loss, with the latter being recursively bounded. We show empirically that RPB significantly outperforms prior works and achieves the best generalization guarantees.

We then explore the connections between Recursive PAC-Bayes, cold posteriors ($T < 1$), and KL-annealing (where T increases from 0 to 1 during

optimization), showing how RPB’s update rules align with these practical techniques and providing new insights into RPB’s effectiveness.

Finally, we present a novel evidence lower bound (ELBO) decomposition for mean-field variational global latent variable models, which could enable finer control of the temperature T . This decomposition could be valuable for future research, such as understanding the training dynamics of probabilistic neural networks.

Resumé

Bayesiansk inferens giver en principiel, probabilistisk ramme for at kvantificere usikkerhed ved at opdatere overbevisninger baseret på tidligere viden og observerede data gennem Bayes' sætning. I bayesiansk dyb læring behandles neurale netværkssvægte som stokastiske variable med fordeling før dataindsamlingen, hvilket muliggør en probabilistisk fortolkning og kvantificering af prædiktiv usikkerhed. Dog mangler bayesianske metoder teoretiske generaliseringsgarantier for uobserverede data. PAC-bayesiansk analyse imødegår denne begrænsning ved at tilbyde en frekventistisk ramme til at udlede generaliseringsgrænser for randomiserede prædiktorer og dermed sikre pålideligheden af bayesianske metoder i maskinlæring.

Temperatur T , eller omvendt temperatur $\lambda = \frac{1}{T}$, stammer oprindeligt fra statistisk mekanik i fysik og opstår naturligt i forskellige områder af statistisk inferens, herunder bayesiansk inferens og PAC-bayesiansk analyse. I bayesiansk inferens, når $T < 1$ ("kolde" posteriors), opvejes sandsynligheden, hvilket resulterer i en skarpere posteriorfordeling. Omvendt, når $T > 1$ ("varme" posteriors), nedvejes sandsynligheden, hvilket fører til en mere diffus posteriorfordeling. Ved at balancere indflydelsen fra observerede data og forudgående regulering kan temperaturjusteringer afhjælpe problemer med under- eller overtilpasning i bayesianske modeller og dermed forbedre prædiktiv ydeevne.

Vi begynder med at undersøge det kolde posteriorfænomen (CPE) i bayesiansk dyb læring. Vi demonstrerer, at fejl i modellen fører til CPE kun, når den bayesianske posterior undertilpasser. Derudover viser vi, at tempererede posteriors er gyldige bayesianske posteriors, der svarer til forskellige kombinationer af sandsynligheder og forudgående distributioner parameteriseret ved temperatur T . Finjustering af T tillader således valg af alternative bayesianske posteriors med mindre fejl i sandsynlighed og forudgående distributioner.

Derefter introducerer vi en effektiv PAC-bayesiansk procedure, Recursive PAC-Bayes (RPB), der muliggør sekventielle posterioropdateringer uden informations tab. Denne metode er baseret på en ny opdeling af det forventede tab for randomiserede klassifikatorer, som fortolker posterior-tabet som et overskudstab i forhold til et nedskaleret forudgående tab, hvor sidstnævnte rekursivt begrænses. Vi viser empirisk, at RPB signifikant overgår tidligere værker og opnår de bedste generaliseringsgarantier.

Vi undersøger derefter forbindelserne mellem Recursive PAC-Bayes, kolde posteriors ($T < 1$), og KL-annealing (hvor T stiger fra 0 til 1 under optimering), hvilket viser, hvordan RPB's opdateringsregler stemmer overens med disse praktiske teknikker og giver nye indsigter i RPB's effektivitet.

Endelig præsenterer vi en ny dekomposition af evidens-lower-bound (ELBO) for mean-field variational globale latente variabelmodeller, som kunne muliggøre finere kontrol af temperaturen T . Denne dekomposition kan være værdifuld for fremtidig forskning, såsom forståelse af træningsdynamikken i probabilistiske neurale netværk.

Acknowledgements

Compared to some of my dearest friends and colleagues, I had almost never envisioned embarking on the journey of a Ph.D., let alone becoming a Ph.D. I still remember the first time I considered the possibility of this path: it was in Amsterdam, where I completed my master's thesis, a project I found incredibly interesting and fulfilling. It was the first time I truly experienced the joy of research. I also remember receiving the offer from Christian for this position. Although I felt happy and excited, the fear of the unknown journey ahead was even more predominant. The journey of a Ph.D. is never easy. It means spending years learning and exploring the boundaries of a particular area of human knowledge, hoping to make a small discovery or breakthrough. It means overcoming the fear of the unknown and believing in the significance of your work.

Looking back, I am reminded of Steve Jobs' 2005 Stanford commencement speech about connecting the dots and doing what you love. On countless nights, when I worried about not having enough publications to graduate or feared my research themes lacked coherence, I recalled these points. As a result, I always followed my intuition to find or engage in projects that truly captivated me. Ultimately, this approach proved to be crucial. Furthermore, without the support, companionship, and encouragement of the incredible people around me, I would never have been able to complete this journey. Now, I would like to express my gratitude to them.

First, I would like to thank my supervisors, Christian and Sadegh. They granted me tremendous freedom and encouraged me to pursue topics that genuinely interested me. Whenever I encountered difficulties, they consistently provided valuable advice and support, both in my work and personal life.

Second, I want to thank my wonderful collaborators. Each of you has taught me something unique, and our collaboration has been a lifelong benefit. Yi-Shan, I am immensely grateful for your help with both my work and life. Your diligence and rigor have deeply inspired me. The quote on your mug, "Every piece of work reflects oneself; anything handled by me will be a masterpiece" perfectly encapsulates you. Our travels, meals, coffee tastings, and alcohol appreciation (though I am far from matching your expertise) are among the

most cherished memories of my Ph.D. life. Andrés, you are always full of brilliant and wild ideas, and you always manage to transform them into beautiful mathematics. Your research introduced me to a fascinating field that became the most important part of my thesis. Yevgeny, although my first impression of you was someone particularly meticulous and zen-like, as the project progressed, I discovered your keen observation of experimental data and your creative thinking in new directions. Your writing is also incredibly clear and elegant. Luis and Badr, although we have never met in person, our remote working relationship has been exceptionally smooth, and our communications have benefited me greatly.

Third, I want to thank my colleagues and friends at DIKU, especially those in the DeLTA Group: Yi-Shan, Hippolyte, Chloé, Saeed, Ola, Yuen, Yunlian, Arthur, Aymeric, Shaojie, and Oliver. Our memories span across the entire European continent, and I never imagined having such an exciting life.

Fourth, I want to thank Novo Nordisk A/S for funding my Ph.D. project and all my colleagues there for their help. I especially want to thank Jakob, David, Merete, Jens, and Haocheng. You taught me the Danish way of working and living, and how to focus on long-term goals in a fun and balanced way.

Fifth, I would like to thank my colleagues and mentors at the State Key Laboratory of Intelligent Technology and Systems at Tsinghua University, especially Professor Jun Zhu, Ziyu, Yuhao, Kaiwen, Tianjiao, Rosie, and Julian. I am very fortunate to have had the opportunity to visit and exchange ideas at the university I dreamed of as a child. The taste, innovation, and high standards of your work have deeply influenced me. I also want to thank Melih, Nicklas, Abdullah, and Bahareh for their help during my visit to the University of Southern Denmark.

Lastly, I want to express my deepest gratitude to my family and friends, with a special mention to my parents. Your unwavering love, support, and encouragement, particularly during my most challenging moments, have been my anchor. I am incredibly fortunate to have you in my life, and you are the most invaluable part of it.

Table of Contents

Abstract	ii
Resumé	iv
Acknowledgements	vi
1 Introduction	1
1.1 Outline of the Thesis	4
1.2 Main Contributions	5
2 The Cold Posterior Effect Indicates Underfitting, and Cold Posteriors Represent a Fully Bayesian Method to Mitigate It	8
2.1 Introduction	9
2.2 Background	12
2.3 The presence of the CPE implies underfitting	14
2.4 Tempered posteriors are Bayesian posteriors	17
2.5 Likelihood misspecification, prior misspecification and the CPE	24
2.6 Data augmentation (DA) and the CPE	32
2.7 Conclusions and limitations	38
2.8 Appendix	38
3 Recursive PAC-Bayes: A Frequentist Approach to Sequential Prior Updates	57
3.1 Introduction	58
3.2 The evolution of data-informed priors and the idea of recursive PAC-Bayes	60
3.3 Split-kl and PAC-Bayes-split-kl inequalities for discrete random variables	64
3.4 Recursive PAC-Bayes bound	67
3.5 Experiments	69
3.6 Discussion	74

3.7	Appendix	75
4	Another Perspective on Recursive PAC-Bayes: Connections to Cold Posteriors and KL-Annealing	85
4.1	Introduction	86
4.2	Recursive PAC-Bayes and cold posteriors	87
4.3	Recursive PAC-Bayes and KL-annealing	89
4.4	Conclusion	90
5	(Generalized) ELBO Revisited: Another Decomposition for Mean-Field Variational Global Latent Variable Models	92
5.1	Introduction	93
5.2	Rewriting the (generalized) ELBO for mean-field variational global latent variable models	94
5.3	Conclusion	96
6	Summary and Discussion	97
	List of Publications	99
	Bibliography	100

Chapter 1

Introduction

Bayesian inference (Bayes, 1763; Bishop, 2007; Ghahramani, 2015) is a cornerstone of statistical inference, providing a flexible and robust framework for updating the probability of a hypothesis as new data is acquired. Utilizing Bayes' theorem, it combines prior knowledge with observed data to produce a posterior distribution that quantifies updated beliefs. This approach allows for a full probabilistic understanding of uncertainty, offering a nuanced perspective beyond point estimates typical of classical methods. Bayesian inference is particularly powerful in complex and evolving datasets, continuously refining predictions as new data becomes available. Its ability to handle hierarchical models and incorporate various levels of uncertainty makes it highly suitable for real-world applications. Additionally, Bayesian methods facilitate model comparison and selection through marginal likelihoods and Bayes factors (Lotfi et al., 2022), ensuring all sources of uncertainty are properly accounted for. Widely used across diverse fields such as biology (Way and Greene, 2018), economics (Geweke et al., 2011), and artificial intelligence (Gelman et al., 2013b), Bayesian inference integrates prior knowledge with empirical data, enhancing both exploratory data analysis and predictive modeling, making it an essential tool for scientific research and practical decision-making.

Bayesian neural networks (BNNs) (Wilson, 2020; Gal and Ghahramani, 2016; Wang and Yeung, 2016) extend Bayesian inference principles to deep learning, treating weights and biases as random variables with prior distributions that update to posterior beliefs based on data. This probabilistic approach contrasts with traditional neural networks, where parameters are fixed point estimates, enhancing model robustness and uncertainty quantification. BNNs incorporate prior knowledge, improve learning efficiency, and

provide accurate uncertainty estimates, crucial for high-stakes applications like medical diagnosis (Abdullah et al., 2022), autonomous driving (McAllister et al., 2017), and financial forecasting (Zhang et al., 2019c). Recent advances, such as variational inference (Blei et al., 2017; Jordan et al., 1999; Wainwright and Jordan, 2008) and Monte Carlo methods (Brooks et al., 2011; Welling and Teh, 2011; Zhang et al., 2019b), have made BNNs more computationally feasible, allowing for efficient approximation of intractable posterior distributions. These techniques have expanded the practical applications of BNNs, enabling their use in large-scale and complex models. BNNs excel in various fields, including computer vision (Gustafsson et al., 2020), natural language processing (Cohen, 2019), and reinforcement learning (Osband et al., 2023), where they handle ambiguous inputs, language variability, and model uncertainty in value functions. The Bayesian approach in deep learning through BNNs improves predictive performance, robustness, interpretability, and reliability, making it a valuable tool in developing and applying deep learning technologies.

On the other hand, probabilistic neural networks (PNNs) (Pérez-Ortiz et al., 2021; Dziugaite and Roy, 2017a; Rivasplata et al., 2019a), or randomized neural networks, generalize the concept of Bayesian neural networks. Unlike Bayesian neural networks, which update distributions using Bayes' rule, PNNs can employ various methods for updating these distributions, offering greater flexibility in training and inference. Methods for obtaining probabilistic neural networks, such as standard Bayes (Bayes, 1763; Bishop, 2007; Ghahramani, 2015), Gibbs Bayes (Ghosh and Basu, 2016; Bissiri et al., 2016a; Chérif-Abdellatif and Alquier, 2019), and power likelihood Bayes (tempered posteriors) (Holmes and Walker, 2017; Grünwald and van Ommen, 2017; Miller and Dunson, 2019), can all be unified under the following optimization problem (Knoblauch et al., 2022):

$$q^*(\boldsymbol{\theta}) = \arg \min_{q \in \Pi} \left\{ \mathbb{E}_{q(\boldsymbol{\theta})} \left[\sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i) \right] + \underbrace{\frac{1}{\lambda}}_{\text{temperature } T} D(q \parallel \pi) \right\} = P(\ell, D, \Pi), \quad (1.1)$$

where ℓ is the loss function, D is the divergence measure, and Π is the set of feasible posteriors. Most prominently, the temperature T , or inverse-temperature $\lambda = \frac{1}{T}$, which originated from statistical mechanics in physics, naturally arises in this context. Adjusting the temperature T provides flexi-

bility in probabilistic models, allowing for better control over the trade-off between fitting the data and maintaining model complexity. This flexibility is particularly beneficial when dealing with complex real-world data, where the true data-generating process is unknown and model misspecification is likely to occur. Since the optimality of Bayesian inference relies on the assumption of perfect model specification (Bochkina and Green, 2022), tuning the temperature to values other than 1 (standard Bayes) often leads to improved performance in practice (Grünwald and van Ommen, 2017; Holmes and Walker, 2017; Wenzel et al., 2020).

Despite probabilistic neural networks’ strong empirical success, the theoretical generalization guarantee on unseen data is still missing. This gap has led to the development of PAC-Bayesian (Probably Approximately Correct Bayesian) analysis (Shawe-Taylor and Williamson, 1997; McAllester, 1998), which provides a robust framework for deriving generalization bounds for probabilistic models. PAC-Bayesian analysis combines the principles of Bayesian inference with PAC learning theory, offering probabilistic bounds on a model’s generalization error based on its empirical error and the complexity of its hypothesis class. This approach has been particularly useful in understanding and improving the generalization capabilities of neural networks (Dziugaite and Roy, 2017a; Zhou et al., 2019; Negrea et al., 2019), where PAC-Bayesian bounds can guide the design of training algorithms that minimize these bounds (Pérez-Ortiz et al., 2021; Rivasplata et al., 2019a), thus enhancing the models’ robustness and predictive accuracy on new data. Interestingly, many PAC-Bayesian bounds, when treated as training objectives, can be represented as special cases of Equation 1.1, containing again the temperature parameter $T = \frac{1}{\lambda}$. By choosing an appropriate temperature to balance empirical performance and model complexity, PAC-Bayesian analysis offers a powerful tool for developing more reliable and theoretically sound machine learning models.

As previously discussed, the temperature parameter T plays a pivotal role in various aspects of statistical inference. Its importance has drawn considerable interest from the research community, particularly in the context of overparameterized neural networks where cold posteriors, characterized by temperatures $T < 1$, are employed. In this work, we aim to provide a comprehensive and unified study of recent advances in the understanding and application of cold posteriors within probabilistic neural networks. Our focus will be on dissecting the cold posterior effect and introducing a novel

approach for learning cold posteriors with tight generalization guarantees. Additionally, we will examine the broader implications of varying temperature settings across different scenarios, illuminating its connections to other key topics within the field.

1.1 Outline of the Thesis

This thesis is organized into the following chapters, where Chapter 2 and Chapter 3 contain the papers Zhang et al. (2024) and Wu et al. (2024) respectively.

In Chapter 2 we study the problem of cold posterior effect (CPE) (Wenzel et al., 2020). The cold posterior effect (CPE) (Wenzel et al., 2020) in Bayesian deep learning demonstrates that posteriors with a temperature $T < 1$ can yield posterior predictives that outperform the standard Bayesian posterior ($T = 1$). Since the Bayesian posterior is considered optimal under perfect model specification, recent studies have explored CPE as a consequence of model misspecification, originating from either the prior or the likelihood. In this research, we refine the understanding of CPE by showing that misspecification induces CPE only when the resulting Bayesian posterior underfits. We theoretically establish that without underfitting, CPE does not occur. Furthermore, we demonstrate that these tempered posteriors with $T < 1$ are indeed valid Bayesian posteriors, corresponding to different combinations of likelihoods and priors parameterized by T . This insight justifies adjusting the temperature hyperparameter T as an effective strategy to mitigate underfitting in the Bayesian posterior. Essentially, we reveal that fine-tuning the temperature T allows for the implicit utilization of alternative Bayesian posteriors, characterized by less misspecified likelihood and prior distributions.

In Chapter 3, we investigate the problem of sequentially updating posteriors in PAC-Bayesian analysis without losing confidence information. PAC-Bayesian analysis, inspired by Bayesian learning, traditionally struggles with maintaining confidence information during sequential updates, as the final confidence intervals only depend on data not used for prior construction, losing information related to the size of the initial dataset. This limitation hinders the benefits of sequential updates. We introduce a novel and surprisingly simple PAC-Bayesian procedure, Recursive PAC-Bayes (RPB), that overcomes this issue by decomposing the expected loss of randomized classifiers into an excess loss relative to a downscaled prior loss plus the downscaled prior loss,

allowing recursive bounding. Additionally, we generalize the split-kl and PAC-Bayes-split-kl inequalities to discrete random variables for bounding excess losses. Our empirical evaluations show that this new procedure significantly outperforms state-of-the-art methods.

In Chapter 4, we explore the connections between Recursive PAC-Bayes (RPB) and two commonly used techniques in probabilistic modeling. The first technique, cold posteriors, characterized by a temperature scaling factor less than one, significantly improves the performance of Bayesian neural networks. The second technique, KL-annealing, gradually increases the weight of the KL divergence term during training, balancing the fit to empirical data with maintaining proximity between prior and posterior distributions. By linking these methodologies with Recursive PAC-Bayes, we provide new insights into its success and practical implications.

In Chapter 5, we introduce a novel decomposition of the (generalized) evidence lower bound (ELBO) objective, which is a special case of Equation 1.1, for (generalized) mean-field variational global latent variable models. This decomposition breaks the (generalized) ELBO objective into interpretable components, with the potential to allow for finer control of the temperature T for optimization and provide valuable insights into the training dynamics of the variational posteriors.

We conclude this work with a discussion of these results in Chapter 6.

1.2 Main Contributions

The main contributions of this work are:

- We theoretically demonstrate that the presence of the CPE implies the Bayesian posterior is underfitting. In fact, we show that if there is no underfitting, there is no CPE.
- We show in a more general case that any tempered posterior is a proper Bayesian posterior with an alternative likelihood and prior distribution, extending Zeno et al. (2021) in the case of classification. We also provide illustrative examples in both regression and classification settings to offer an intuitive and visual understanding of our findings.
- We demonstrate that likelihood misspecification and prior misspecification result in the cold posterior effect (CPE) only if they also induce

underfitting. Furthermore, we provide comprehensive empirical evidence in both linear regression and neural network settings, where approximate inference is applied. Additionally, we discuss the implications of model size and sample size in relation to CPE and underfitting.

- We show that data augmentation results in stronger CPE because it induces a stronger underfitting of the Bayesian posterior.
- We show that fine-tuning the temperature T in tempered posteriors offers a well-founded and effective Bayesian approach to mitigate underfitting: By fine-tuning the temperature T , we implicitly utilize alternative Bayesian posteriors resulting from less misspecified likelihood and prior distributions.
- We present a novel PAC-Bayesian bound, Recursive PAC-Bayes (RPB), that allows for meaningful and beneficial sequential data processing and sequential posterior updates, without loss of confidence information.
- We introduce a new and simple way to decompose the loss of a randomized classifier defined by the posterior into an excess loss relative to a downscaled loss of the classifier defined by the prior plus the downscaled loss of the prior.
- Importantly, this excess loss can be bounded using PAC-Bayes-Empirical-Bernstein-style inequalities, and the loss of the prior can be bounded recursively, preserving confidence information on the prior.
- We provide a generalization of the split-kl and PAC-Bayes-split-kl inequalities from ternary to general discrete random variables, based on a novel representation of discrete random variables as a superposition of Bernoulli random variables.
- We develop a practical and efficient optimization procedure based on the Recursive PAC-Bayes bound, resulting in posteriors that outperform state-of-the-art methods on benchmark datasets.
- We explore the connections between Recursive PAC-Bayes, cold posteriors, and KL-annealing, providing new insights into Recursive PAC-Bayes's effectiveness and empirical success.

- We introduce a novel decomposition of the (generalized) ELBO objective for mean-field variational global latent variable models, where latent variables (typically model or network weights) are shared across all data points. By breaking down the (generalized) ELBO into interpretable components, this approach might allow for finer control of the temperature T during optimization and provide a deeper understanding of the training dynamics of variational posteriors.

Chapter 2

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

The work presented in this chapter is based on a paper that has been published as:

Yijie Zhang, Yi-Shan Wu, Luis A. Ortega, and Andres R Masegosa. The cold posterior effect indicates underfitting, and cold posteriors represent a fully bayesian method to mitigate it. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=GZORXGxHHT>.

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 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 likelihoods and priors parameterized by T . 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. The code for replicating the experiments can be found at <https://github.com/pyijiezhang/cpe-underfit>.

2.1 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., 2013a), 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 adjusting priors can help alleviate the CPE in certain cases (Adlam et al., 2020; Zeno et al., 2021), 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., 2013a), 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, by building on Zeno et al. (2021), we show how 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 2.3. (ii) We show in a more general case that any tempered posterior is a proper Bayesian posterior with an alternative likelihood and prior distribution in Section 2.4, extending Zeno et al. (2021) in the case of classification. (iii) We show in Section 2.5 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 2.6. 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.

2.2 Background

2.2.1 Notation and assumptions

Let us start by introducing basic notation. Consider a supervised learning problem with the sample space $\mathcal{Y} \times \mathcal{X}$. In this work, we consider two cases: when \mathcal{Y} is a finite set, corresponding to a supervised classification problem, and when \mathcal{Y} is a subset of \mathbb{R} , corresponding to a regression problem. For simplicity, we also assume that \mathcal{X} is a subset of \mathbb{R}^d . Let the training set $D = (\mathbf{Y}, \mathbf{X})$, where \mathbf{Y} denotes the set of output entries and \mathbf{X} denotes the set of input entries. If D consists n pairs of samples, we denote $D = \{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^n$.

We assume a family of probabilistic models parameterized by Θ , where each $\theta \in \Theta$ defines a conditional probability distribution for a sample (\mathbf{y}, \mathbf{x}) , denoted by $p(\mathbf{y}|\mathbf{x}, \theta)$. The standard metric to measure the quality of a probabilistic model θ on a sample (\mathbf{y}, \mathbf{x}) is the (negative) log-loss $-\ln p(\mathbf{y}|\mathbf{x}, \theta)$. The expected (or population) loss of a probabilistic model θ is defined as $L(\theta) = \mathbb{E}_{(\mathbf{y}, \mathbf{x}) \sim \nu}[-\ln p(\mathbf{y}|\mathbf{x}, \theta)]$, where ν denotes the unknown data-generating distribution ν on $\mathcal{Y} \times \mathcal{X}$. The empirical loss of the model θ on the data D is defined as $\hat{L}(D, \theta) = -\frac{1}{n} \ln p(\mathbf{Y}|\mathbf{X}, \theta) = -\frac{1}{n} \sum_{i \in [n]} \ln p(\mathbf{y}_i|\mathbf{x}_i, \theta)$. In this work, we assume that the likelihood function fully factorizes, i.e., $p(\mathbf{Y}|\mathbf{X}, \theta) = \prod_{(\mathbf{y}, \mathbf{x}) \in D} p(\mathbf{y}|\mathbf{x}, \theta)$. We might use the notation $p(D|\theta)$ for $p(\mathbf{Y}|\mathbf{X}, \theta)$ in the presentation when the roles of input/output in the samples are not important in the context. Also, if it induces no ambiguity, we use $\mathbb{E}_\nu[\cdot]$ as a shorthand for $\mathbb{E}_{(\mathbf{y}, \mathbf{x}) \sim \nu}[\cdot]$.

2.2.2 (Generalized) Bayesian learning

In Bayesian learning, we learn a probability distribution $\rho(\theta|D)$, often called a posterior, over the parameter space Θ from the training data D . Given a new input \mathbf{x} , the posterior ρ makes the prediction about \mathbf{y} through (an approximation of) *Bayesian model averaging (BMA)* $p(\mathbf{y}|\mathbf{x}, \rho) = \mathbb{E}_{\theta \sim \rho}[p(\mathbf{y}|\mathbf{x}, \theta)]$, 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 $\mathbb{E}_{\theta \sim \rho}[\cdot]$. The predictive performance of such BMA is usually measured by the Bayes loss, defined by

$$B(\rho) = \mathbb{E}_\nu[-\ln \mathbb{E}_\rho[p(\mathbf{y}|\mathbf{x}, \theta)]] . \quad (2.1)$$

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

$$p_\lambda(\boldsymbol{\theta}|D) \propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda p(\boldsymbol{\theta}). \quad (2.2)$$

Note that when $\lambda \neq 1$, $\int p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda d\mathbf{Y}$ might not be 1 in general. An implicit assumption is that $p_\lambda(\boldsymbol{\theta}|D)$ is a *proper distribution*, meaning the normalization constant is finite. In supervised classification problems, this is always the case because $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \leq 1$. Consequently, for any $\lambda > 0$, we have $1 = \int p(\boldsymbol{\theta}) d\boldsymbol{\theta} > \int p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Thus, the tempered posteriors are always a proper distribution in supervised classification problems.

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:

$$p_\lambda(\boldsymbol{\theta}|D) = \arg \min_{\rho} \mathbb{E}_{\rho}[-\ln p(D|\boldsymbol{\theta})] + \frac{1}{\lambda} \text{KL}(\rho(\boldsymbol{\theta}|D), p(\boldsymbol{\theta})). \quad (2.3)$$

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 .

As 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 $\hat{G}(p_\lambda, D)$.

2.3 The presence of the CPE implies underfitting

Section Overview

We present a definition of the Cold Posterior Effect (CPE) (Definition 2.1) and show that the presence of CPE indicates the Bayesian posterior is underfitting, where both the testing loss (Definition 2.1) and training loss (Proposition 2.1) can be improved at the same time by decreasing the temperature T (increasing λ). We also present the necessary condition of the CPE in Proposition 2.2 and the case when Bayesian posterior is optimal in Theorem 2.3.

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 2.4 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 2.1. *We say there is a CPE for Bayes loss if and only if the derivative of the Bayes loss of the posterior p_λ , $B(p_\lambda)$, evaluated at $\lambda = 1$ is*

negative. That is,

$$\frac{d}{d\lambda} B(p_\lambda)_{|\lambda=1} < 0, \quad (2.4)$$

where the magnitude of the derivative $\frac{d}{d\lambda} B(p_\lambda)_{|\lambda=1}$ defines the strength of the CPE.

According to the above definition, a (relatively large) negative derivative $\frac{d}{d\lambda} B(p_\lambda)_{|\lambda=1}$ 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 derivative $\frac{d}{d\lambda} B(p_\lambda)_{|\lambda=1}$ 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 derivative of the Bayes loss.

Next, we present another critical observation. We postpone the proofs in this section to Appendix 2.8.1.

Proposition 2.1. *The derivative of the empirical Gibbs loss of the tempered posterior p_λ satisfies*

$$\forall \lambda \geq 0 \quad \frac{d}{d\lambda} \hat{G}(p_\lambda, D) = -\mathbb{V}_{p_\lambda}(\ln p(D|\boldsymbol{\theta})) \leq 0, \quad (2.5)$$

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

As shown in Proposition 2.6 in Appendix 2.8.1, to achieve $\mathbb{V}_{p_\lambda}(\ln p(D|\boldsymbol{\theta})) = 0$, 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, $\mathbb{V}_{p_\lambda}(\ln p(D|\boldsymbol{\theta}))$ 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 minimum to observe the CPE.

Proposition 2.2. *A necessary condition for the presence of the CPE, as defined in Definition 2.1, is that*

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

Insight 1. *Definition 2.1 in combination with Proposition 2.1 shows 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 2.2 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). We hence deduce that the presence of CPE implies that the original Bayesian posterior ($\lambda = 1$) underfits. This conclusion arises because there exists another Bayesian posterior (i.e., $p_\lambda(\boldsymbol{\theta}|D)$ with $\lambda > 1$) that has lower training (Proposition 2.2) and testing (Definition 2.1) 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 2.4. 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 derivative of the Bayes loss with respect to λ evaluated at $\lambda = 1$ become zero ($\frac{d}{d\lambda} B(p_\lambda)|_{\lambda=1} = 0$)? 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, the extra information about the data-generation process does not provide the Bayesian posterior with better performance on the originally provided training data.

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

$$\bar{p}_\lambda(\boldsymbol{\theta}|D) = \mathbb{E}_\nu [\tilde{p}_\lambda(\boldsymbol{\theta}|D, (\mathbf{y}, \mathbf{x}))]. \quad (2.6)$$

In this sense, \bar{p}_λ 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}_λ , the following result introduces a characterization of the *optimality* of the original Bayesian posterior.

Theorem 2.3. *The derivative of the Bayes loss at $\lambda = 1$ is null, i.e., $\frac{d}{d\lambda}B(p_\lambda)|_{\lambda=1} = 0$, 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 Equation 2.6, 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 give examples that empirically illustrate Theorem 2.3 and the induced insight later in Section 2.5.5.

2.4 Tempered posteriors are Bayesian posteriors

Section Overview

By extending Zeno et al. (2021) on classification only, we show in general that tempered posteriors are proper Bayesian posteriors with an alternative combination of likelihood and prior functions parameterized by λ . Thus, the occurrence of CPE can be explained within the Bayesian framework.

- We provide two examples to show how λ influences the new likelihoods in Section 2.4.1 and two examples to show how λ influences the new priors in Section 2.4.2.
- We show in Section 2.4.3 that the generalized ELBOs are also proper ELBOs.
- We expand the discussion of the implications in Section 2.4.4.

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 $\lambda \neq 1$). However, as previously discussed by Zeno et al. (2021), this tempered posterior effectively serves as a *proper Bayesian posterior* with a combination of *new likelihood and prior functions*. We extend this result beyond classification to our Proposition 2.4, proved in Appendix 2.8.2.1.

Before delving into the description of the new likelihood and prior functions, it is essential to acknowledge a fundamental aspect. Given a labeled dataset $D = (\mathbf{X}, \mathbf{Y})$ and the conditional likelihood associated to a classification model, the application of Bayes' theorem naturally results in the following Bayesian posterior:

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X}),$$

where the prior over $\boldsymbol{\theta}$ is a *conditional prior* (Marek et al., 2024; Zeno et al., 2021) that depends on the unlabelled training data \mathbf{X} . However, specifying $p(\boldsymbol{\theta}|\mathbf{X})$ for a complex model, like a deep neural network, poses a significant challenge. Therefore, for practical purposes, nearly all existing works (Wenzel et al., 2020; Fortuin et al., 2022) assume $\boldsymbol{\theta}$ to be independent of \mathbf{X} , resulting in the simplified expression $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})$, where the prior over $\boldsymbol{\theta}$ is now an *unconditional prior*.

Proposition 2.4. *For any given dataset $D = (\mathbf{X}, \mathbf{Y})$ such that the likelihood fully factories $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \prod_{(\mathbf{y}, \mathbf{x}) \in D} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, and $\lambda > 0$, the tempered posterior defined in Equation 2.2 can be expressed as a Bayesian posterior with a new prior and likelihood function as follows:*

$$p_\lambda(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto q(\boldsymbol{\theta}|\mathbf{X}, \lambda) \prod_{(\mathbf{y}, \mathbf{x}) \in D} q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda), \quad (2.7)$$

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

$$q(\boldsymbol{\theta}|\mathbf{X}, \lambda) \propto p(\boldsymbol{\theta}) \prod_{\mathbf{x} \in \mathbf{X}} \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y}, \quad q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) = \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda}{\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y}}. \quad (2.8)$$

Note that the new conditional likelihood $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$ and the new prior $q(\boldsymbol{\theta}|\mathbf{X}, \lambda)$ are both parametrized by the same $\lambda > 0$, and note that the prior only depends on the unlabelled training data \mathbf{X} .

Adlam et al. (2020) shows that in the specific scenario of Gaussian process regression (essentially our Bayesian linear regression example in Figures 2.2 and 2.3), any positive temperature aligns with a legitimate posterior under an adjusted unconditional prior. This can be seen as a special case of our argument.

In the rest of the section, we will discuss in Section 2.4.1 how the new likelihoods change with respect to λ , and in Section 2.4.2 how the new priors change with respect to λ . Additionally, besides demonstrating that tempered posteriors are proper Bayesian posteriors, we also show in Section 2.4.3 that the generalized ELBOs are also proper ELBOs. Lastly, we discuss the implications of the results in Section 2.4.4.

2.4.1 How λ influences the new likelihoods

The next result, proved in Appendix 2.8.2.2, shows that in supervised classification settings, higher λ values induce new likelihood distributions with lower aleatoric uncertainty or, equivalently, lower Shannon entropy, denoted as $H(q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) := -\sum_{\mathbf{y} \in \mathcal{Y}} q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$.

Proposition 2.5. *For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, any $\mathbf{x} \in \mathcal{X}$, and any finite output set \mathcal{Y} , the entropy of the conditional likelihood $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$ monotonically decreases with $\lambda > 0$, i.e.,*

$$\frac{d}{d\lambda} H(q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) \leq 0 \quad \forall \lambda > 0. \quad (2.9)$$

This result also holds for regression settings, where $\mathcal{Y} \subset \mathbb{R}^d$, under the differential entropy, assuming the Leibniz rule holds. See the proof for a detailed discussion on the matter.

We give two concrete examples, one in regression and one in classification, to illustrate the proposition.

Regression example

Consider the case where the original likelihood is Gaussian, defined as $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\mu(\mathbf{x}, \boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta}))$, where the variance is input-independent, as typically seen in many regression problems. Then, following Equation 2.8, the new likelihood corresponds to a scaling in the variance, given by $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) = \mathcal{N}(\mu(\mathbf{x}, \boldsymbol{\theta}), \frac{\sigma^2(\boldsymbol{\theta})}{\lambda^2})$. Thus, as λ increases, the tempered likelihood $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$ induces a proper Gaussian likelihood with reduced variance, i.e., a new likelihood with lower aleatoric uncertainty.

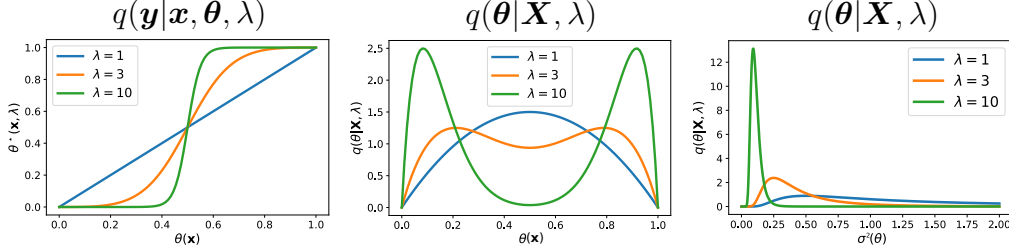


Figure 2.1: Illustration of the new likelihood $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$ and priors $q(\boldsymbol{\theta}|\mathbf{X}, \lambda)$. In the left and middle figures, the original likelihood is in the form of the Bernoulli distribution. The left figure demonstrates the transformation from $\theta(\mathbf{x})$ to $\theta^*(\mathbf{x}, \lambda) := \frac{\theta(\mathbf{x})^\lambda}{\theta(\mathbf{x})^\lambda + (1 - \theta(\mathbf{x}))^\lambda}$. In the middle figure, we display a Beta-Binomial example, where the prior, initialized as a Beta distribution, is updated with a single Bernoulli-distributed sample. In the right figure, we display the new prior, initialized as an inverse-gamma prior and updated with a Gaussian likelihood with a single observation.

Classification example

Consider the case of a binary classification problem where the original conditional likelihood is Bernoulli, defined as $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \theta(\mathbf{x})^y(1 - \theta(\mathbf{x}))^{1-y}$ with $y \in \{0, 1\}$ and the input-dependent parameter function $\theta(\mathbf{x}) \in [0, 1]$, which is usually implemented by a neural network with a softmax activation function in the last layer. Then, following Equation 2.8, the new conditional likelihood $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) = \theta^*(\mathbf{x}, \lambda)^y(1 - \theta^*(\mathbf{x}, \lambda))^{1-y}$ also follows a Bernoulli distribution with a different parameter function $\theta^*(\mathbf{x}, \lambda) = \frac{\theta(\mathbf{x})^\lambda}{\theta(\mathbf{x})^\lambda + (1 - \theta(\mathbf{x}))^\lambda} \in [0, 1]$. The function $\theta^*(\mathbf{x}, \lambda)$ is displayed in Figure 2.1 (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.

In both cases, we see that as suggested by Proposition 2.5, as λ increases, the new conditional likelihoods $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$ have lower entropy, i.e., lower aleatoric uncertainty. In Sections 2.5.2 and 2.5.3, we will further explore the implications of this finding and its connection to existing literature.

2.4.2 How λ influences the new priors

On the other hand, according to Proposition 2.4, using the tempered posteriors implies implicitly using the prior $q(\boldsymbol{\theta}|\mathbf{X}, \lambda)$. Such prior depends on the

unlabelled training data \mathbf{X} . On top of that, the functional form of the likelihood function is defined by the probabilistic model family through the term $\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y}$ for $\mathbf{x} \in \mathbf{X}$. Hence, models $\boldsymbol{\theta}$ that yield a large value for this term across most of the training data $\mathbf{x} \in \mathbf{X}$ will be assigned larger probability mass by the new prior. We will showcase this effect in both regression and binary classification problems. Moreover, we will see how the new prior $q(\boldsymbol{\theta}|\mathbf{X}, \lambda)$ with $\lambda > 1$ *favors those models within the model class that yield likelihoods with lower aleatoric uncertainty on the training data \mathbf{X} .*

Regression example

Consider the case where the original likelihood is Gaussian, defined as $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\mu(\mathbf{x}, \boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta}))$, where the variance is input-independent, as typically seen in many regression problems. A common parametrization involves $\boldsymbol{\theta} = (\mathbf{w}, \gamma)$, where \mathbf{w} refer to the weights of the neural network defining the function $\mu(\mathbf{x}, \boldsymbol{\theta})$ and $\gamma > 0$ is a parameter encoding the variance of the Gaussian likelihood such that $\sigma^2(\boldsymbol{\theta}) = \gamma$. The prior $p(\boldsymbol{\theta})$ is then defined as $p(\boldsymbol{\theta}) = p(\mathbf{w})p(\gamma)$, where $p(\mathbf{w})$ is usually a Gaussian distribution with a diagonal covariance matrix, and $p(\gamma)$ is usually defined in terms of an inverse-gamma distribution. Following Equation 2.8, the new prior would be expressed as $q(\boldsymbol{\theta}|\mathbf{X}, \lambda) = q(\mathbf{w}|\mathbf{X}, \lambda)q(\gamma|\mathbf{X}, \lambda)$, where each term

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

Figure 2.1 (right) plots the density of $q(\gamma|\mathbf{X}, \lambda)$ when only one data is observed, with various $\lambda > 1$ 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.

Classification example

Consider another case where the original conditional likelihood is Bernoulli, defined as $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = \theta(\mathbf{x})^y(1 - \theta(\mathbf{x}))^{1-y}$ with $y \in \{0, 1\}$ and $\theta(\mathbf{x}) \in [0, 1]$, as commonly used in binary classification problems. Also, take any prior $p(\boldsymbol{\theta})$. Then, following Equation 2.8, the new prior is expressed as

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

Figure 2.1 (middle) illustrates the transformation of the prior for a Beta-Binomial model with a single training sample. Initially, the prior $p(\boldsymbol{\theta})$ is taken as a Beta distribution, while the likelihood of this single data is Bernoulli. As $\lambda \geq 1$ increases, the new prior assigns more probability mass to models where $\theta(\mathbf{x})$ 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.

We will discuss in Section 2.5.4 further implications of this finding and how it relates to the literature.

2.4.3 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 to adjust λ to improve the predictive accuracy of variational approximations:

$$q_\lambda^\star := \arg \min_{r \in \Pi} \mathbb{E}_r[-\ln p(D|\boldsymbol{\theta})] + \frac{1}{\lambda} \text{KL}(r(\boldsymbol{\theta}), p(\boldsymbol{\theta})), \quad (2.10)$$

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. However, Proposition 2.4 can be used to justify that such a variational posterior q_λ^\star still emerges from minimizing a valid ELBO. Specifically, it is constructed based on the revised likelihood and prior functions as follows:

$$q_\lambda^\star = \arg \min_{r \in \Pi} \mathbb{E}_r[-\ln q(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \lambda)] + \text{KL}(r(\boldsymbol{\theta}), q(\boldsymbol{\theta}|\mathbf{X}, \lambda)). \quad (2.11)$$

Consequently, this analysis shows that using generalized ELBOs as Equation 2.10 perfectly adheres to variational and Bayesian principles.

2.4.4 Insights and implications from the section

In this section, we show that employing tempered posteriors seamlessly fits within a Bayesian framework, which streamlines and enriches the use of diverse likelihood and prior functions.

With the characterization in Section 2.3, observing CPE implies that the tempered posterior, which implicitly employs the new likelihood and priors defined in Equation 2.8 is better specified in comparison to the original Bayesian posterior. The alignment with the underlying data-generating distribution is easily achieved by tempering. Consequently, tempered posteriors offer a simple, computationally efficient, and theoretically sound approach to mitigate the underfitting problem often encountered in contemporary Bayesian deep learning methods.

Furthermore, as discussed in Section 2.4.1 and Section 2.4.2, increasing λ results in likelihoods with lower aleatoric uncertainty and priors that favor models yielding such likelihoods on the training data X . Therefore, the occurrence of CPE in contemporary Bayesian deep learning indicates that the models currently employed in the field often underfit the data by assuming models with too high aleatoric uncertainty. This strengthens our understanding of the CPE as a consequence of underfitting, resulting from poorly specified likelihood and prior functions.

We will further expand on and discuss how these implications relate to the literature in Section 2.5.

2.5 Likelihood misspecification, prior misspecification and the CPE

Section Overview

We relate our analysis in previous sections to the main arguments of CPE from the literature.

- Section 2.5.1: we demonstrate with Bayesian linear regression that exact inference can also bring CPE, showing CPE is not merely a side effect of approximate inference in NNs.
- Section 2.5.2: using the same regression examples, we show that model misspecification can lead to underfitting or overfitting. CPE arises specifically when misspecified likelihoods or priors cause underfitting, not just from misspecification alone.
- Section 2.5.3: likelihood misspecification is often identified as a source of CPE in practice. We show it is because the standard softmax likelihood (high aleatoric uncertainty) is misspecified and underfits the data-generating process (curated data with low aleatoric uncertainty).
- Section 2.5.4: we show prior misspecification leads to CPE if it induces underfitting. Using tempered posteriors implicitly defines better-specified conditional priors that alleviate it.
- Section 2.5.5: we show that larger models have more flexibility to fit data, thereby mitigating underfitting and CPE. Conversely, with small models and abundant data, the Bayesian posterior may already fit the data optimally, thereby exhibiting minimal underfitting (by our definition) and CPE.

2.5.1 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). There are occasions suggesting that if the approximate inference method is accurate enough, the CPE disappears (Izmailov et al., 2021). However, Proposition 2.1 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, Figures 2.2 and 2.3 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 2.3b and 2.2c 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). We also note that Adlam et al. (2020) presents similar findings and observations for Gaussian process regression.

2.5.2 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 2.3b and 2.2c show how the Bayesian posterior underfits due to likelihood and prior misspecification, respectively. On the other hand, Figure 2.2b 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 exist other posteriors (p_λ with $\lambda < 1$) with lower testing loss, which, at the same time, have higher training loss due to Proposition 2.1. As a result, to describe CPE merely as a model misspecification issue without acknowledging underfitting offers a narrow interpretation of the problem.

The examples presented in Figures 2.2 and 2.3 help illustrate the results of Proposition 2.4 and provide concrete demonstrations of the theoretical insights discussed: when CPE shows up, tuning λ is akin to finding another Bayesian posterior with a less misspecified likelihood and prior. However, we note that in this particular Bayesian linear regression setup, the new prior $q(\boldsymbol{\theta}|\mathbf{X}, \lambda)$ is always equal to initial prior $p(\boldsymbol{\theta})$ because the variance of the likelihood is assumed to be constant. Therefore, the analysis of the regression case in Section 2.4.2 does not directly apply here.

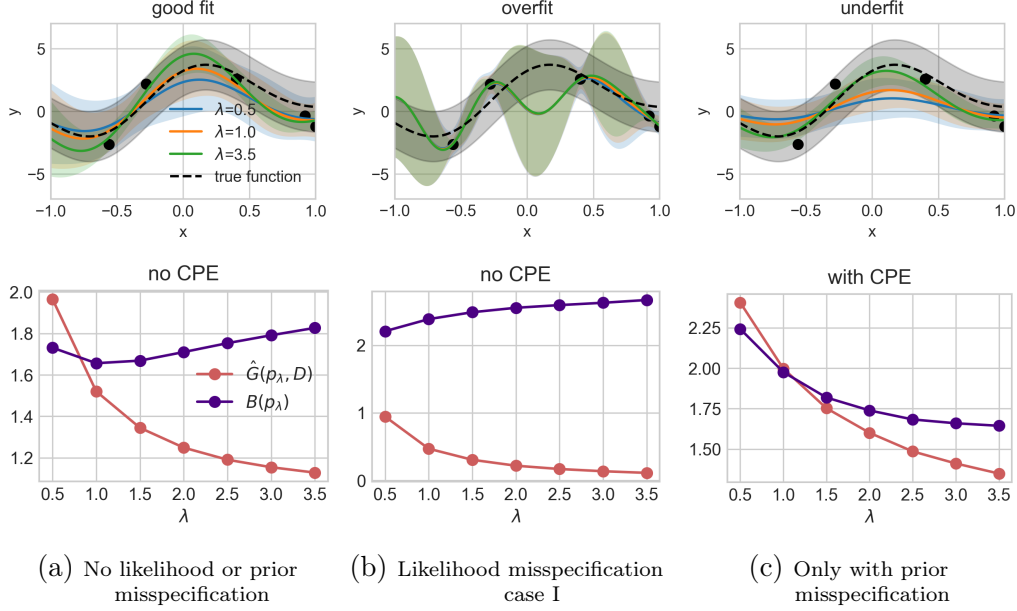


Figure 2.2: **1. The CPE occurs in Bayesian linear regression with exact inference. 2. Model misspecification can lead to overfitting and to a “warm” posterior effect (WPE).** Every column displays a specific setting, as indicated in the caption. The first row shows exact Bayesian posterior predictive 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 2.8.4.

In the discussion regarding likelihood, we refer to the regression example in Section 2.4.1. Let’s first have a look at Figure 2.3b, where the Gaussian likelihood model has a larger variance than the true data-generating process. By increasing λ , we obtain a likelihood model with a smaller variance (divided by λ^2 , as shown in the regression example in Section 2.4.1), i.e., we induce a new likelihood with lower aleatoric uncertainty (Proposition 2.5). Such a new model is closer to the true data-generating distribution and less misspecified, thus enjoying better performance. The opposite can be seen in Figure 2.2b, where the Gaussian likelihood model has a lower variance than the true data-generating distribution and the WPE occurs.

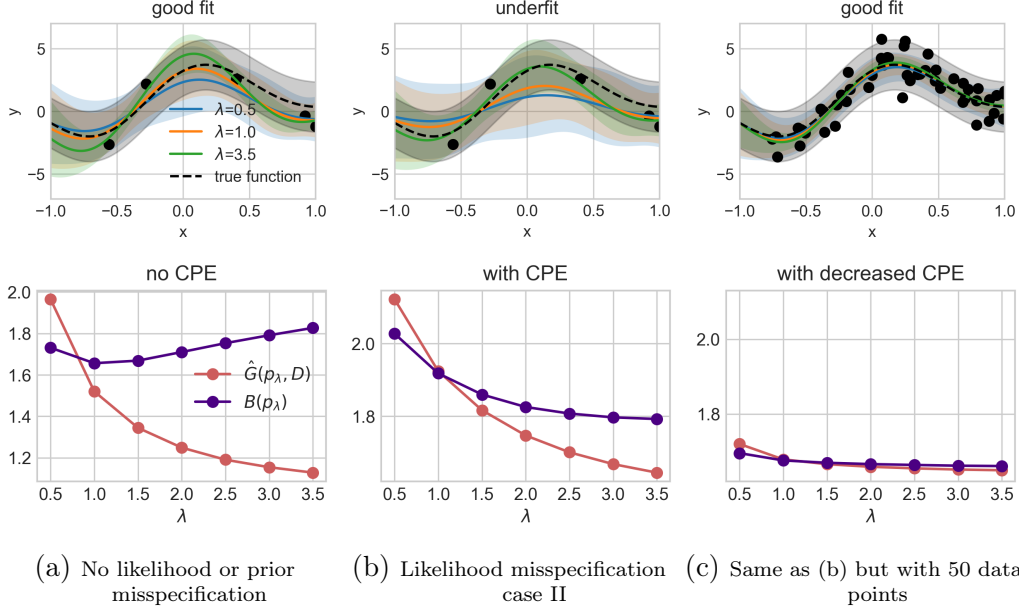


Figure 2.3: **Extended results of Figure 2.2 with more configurations of model misspecification.**

2.5.3 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 $\nu(\mathbf{y}|\mathbf{x})$ usually takes values very close to either 1 or 0. However, as previously discussed in (Aitchison, 2021; Kapoor et al., 2022), the standard likelihoods used in deep learning for image classification, like softmax or sigmoid, tend to allocate more spread-out probabilities to the outcomes, implicitly reflecting a higher level of aleatoric uncertainty. 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.

Our theoretical analysis aligns with these findings in Sections 2.4.1 and 2.4.4: fitting low aleatoric uncertainty data-generating distributions, e.g., $\nu(y|\mathbf{x}) \in \{0.01, 0.99\}$, with high aleatoric uncertainty likelihood functions e.g., $p(y|\mathbf{x}, \boldsymbol{\theta}) \in [0.2, 0.8]$, induces underfitting, and thus, CPE. 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 Propositions 2.4 and 2.5, our work explains why the likelihood implicitly used by the tempered posterior with $\lambda > 1$ provides better generalization performance. Because, in this case, we are using a likelihood $q(\boldsymbol{\theta}|\mathbf{X}, \lambda)$ (Equation 2.8) 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 2.4a and 2.4b, along with Figures 2.5a and 2.5b, 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 2.4b lies in the adoption of a tempered softmax likelihood, defined as $p(y|x, \theta) = (1 + \exp(-\gamma \text{logits}(x, \theta)))^{-1}$, with $\gamma = 2$, compared to $\gamma = 1$ in Figure 2.4a. 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 2.4b compared to Figure 2.4a. From the perspective of Proposition 2.4 and specifically Proposition 2.5, the intrinsic lower aleatoric uncertainty of the likelihood used in Figure 2.4b (softmax with $\gamma = 2$) makes the potential for improvement through increasing λ somewhat limited, resulting in a less pronounced CPE compared to Figure 2.4a. It is, however, important to highlight the critical interaction between the likelihood and the prior, as we discuss next.

2.5.4 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

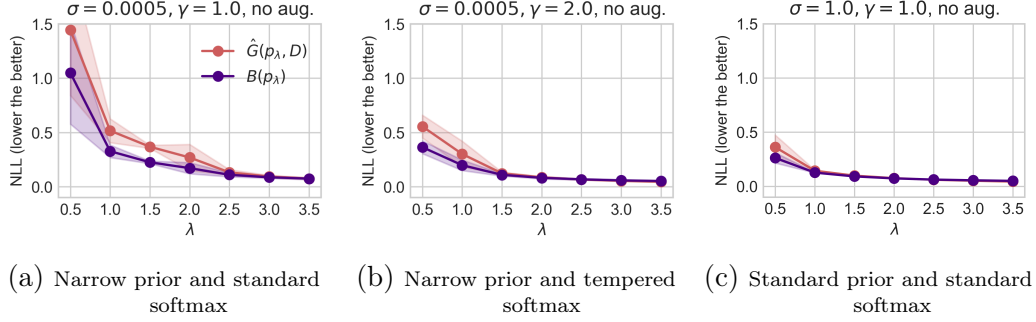


Figure 2.4: **Experimental illustrations for the arguments in Section 2.5 using small CNN via SGLD on MNIST.** We show similar results on Fashion-MNIST with small CNN and CIFAR-10(0) with ResNet-18 in Appendix 2.8.5. Figures 2.4a to 2.4c illustrate the arguments in Section 2.5. Figure 2.4c uses the standard prior ($\sigma = 1$) and the standard softmax ($\gamma = 1$) for the likelihood without applying DA. Figure 2.4a follows a similar setup except for using a narrow prior. Figure 2.4b uses a narrow prior as in Figure 2.4a 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 2.8.5.

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 discussed in Section 2.3, the absence of underfitting means the absence of CPE. This suggests flexible likelihood functions may still result in posteriors that underfit, due to the prior’s tendency to overly regularize. This may incur CPE when the strong prior fails to allocate enough probability to models that both fit the training data well and exhibit good generalization capabilities. As

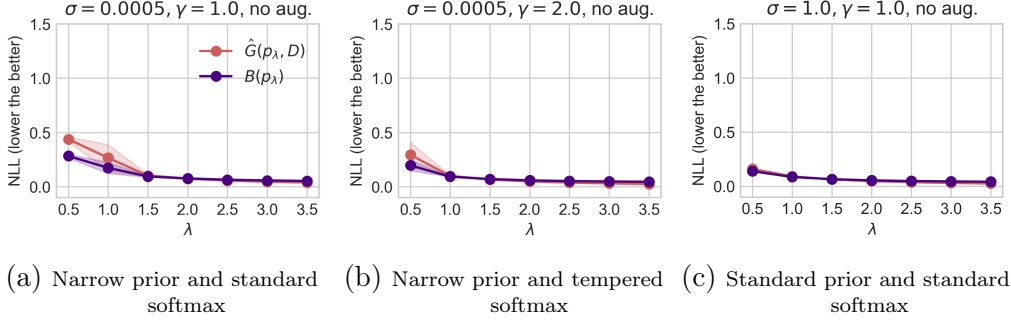


Figure 2.5: **Experimental illustrations for the arguments in Section 2.5 using large CNN via SGLD on MNIST. We show similar results on Fashion-MNIST with large CNN and CIFAR-10(0) with ResNet-50 in Appendix 2.8.5.** The experiment setup is similar to the setups in Figure 2.4 but with a large CNN. Please refer to Appendix 2.8.5 for further details on the model.

detailed in Section 2.4.2, employing tempered posteriors with $\lambda > 1$ effectively defines a conditional prior $q(\theta|\mathbf{X}, \lambda)$ that favors models with lower aleatoric uncertainty. If such models align better with the training data compared to the original prior, then we may observe CPE. Also, the conditional prior $q(\theta|\mathbf{X}, \lambda)$ with $\lambda > 1$ can be considered better specified than the original prior $p(\theta)$.

Figures 2.4a and 2.4c exemplify this situation. The prior in the case of Figures 2.4a is very narrow ($\sigma = 0.0005$), inducing strong regularization. Such a narrow prior results in a posterior that severely underfits the training data, evident from the high empirical Gibbs loss that deviates significantly from zero. Additionally, we observe a strong CPE. On the other hand, with a flatter prior in the case of Figure 2.4c, the CPE is considerably diminished. According to the discussion above and Sections 2.4.2 and 2.4.4, we know that the flatter prior allocates more probability mass to preferred models. Also, such preferred models have lower aleatoric uncertainty than the ones assigned initially by the narrow prior in the former case. To elaborate further, in the former case, the new prior $q(\theta|\mathbf{X}, \lambda)$ with $\lambda > 1$ would place much more probability mass to models with lower aleatoric uncertainty than the narrow prior and strongly alleviating underfitting. In the second case, since the flatter prior already distributes probability mass more broadly across the model class, the room to shift probability mass to models with lower aleatoric uncertainty is more limited than that from a narrower prior, resulting in a

milder CPE.

2.5.5 Model size, sample size in relation to 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. We demonstrate in Figure 2.4 and Figure 2.5. Specifically, in our experiments presented in Figure 2.4, we use 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 2.5, which has less underfitting, and we see the CPE is strongly alleviated. Actually, this effect can be directly inferred from Theorem 2.3. For an extremely flexible model capable of perfectly fitting both the original training samples and new samples, this theorem suggests that a CPE should not be expected, as the model’s fit on the original data remains perfect, even when new examples are introduced.

Theorem 2.3 can also be used to understand why small models in the presence of large training data sets do not exhibit CPE. We empirically illustrate this point in Figures 2.2 and 2.3. In particular, Figure 2.3b and Figure 2.3c use the same (small) regression models and settings where the only difference is that Figure 2.3b uses 5 data points while Figure 2.3c uses 50 data points. In situations where a model possesses limited flexibility and the training set is large, including additional examples should barely affect the fit of the original training data because the Bayesian posterior is highly concentrated and will be barely affected by a single extra sample. Then, as predicted by Theorem 2.3, CPE in Figure 2.3c is much less significant than in Figure 2.3b.

Finally, it’s worth noting that Figure 11 in Wenzel et al. (2020) shows the opposite effect, where larger models exhibit much stronger CPE compared to shallower or narrower versions of the same architectures. However, it’s important to recognize that Wenzel et al. (2020) studied full-tempering, whereas our work focuses on likelihood-tempering. For full-tempering, Proposition 2.1 does not necessarily hold. Intuitively, since λ operates on both the likelihood (data) and the prior (regularization) simultaneously, the effect of increasing λ

is mixed, not necessarily improving the fit on the training data. Consequently, the CPE brought by full-tempering as λ increases does not necessarily coincide with better training loss, as the training loss may not be improvable. As a result, the CPE observed with full-tempering cannot be interpreted solely as underfitting. Therefore, for full-tempering, increasing model capacity may not achieve a lower degree of CPE, unlike the behavior we observed in our case focusing on likelihood-tempering.

2.6 Data augmentation (DA) and the CPE

Section Overview

We show conditions under which data augmentation exacerbates the CPE.

- Section 2.6.1: starting with the Gibbs loss for clarity, we show that data augmentation induces a stronger CPE on the Gibbs loss if the augmented data provides more information about the data-generating process, increasing the correlation between the expected and empirical losses.
- Section 2.6.2: extending the above idea, we show analogous conditions where data augmentation exacerbates CPE on the Bayes loss.

Machine learning is applied to many different fields and problems. In many of them, the data-generating distribution is known to have properties that can be exploited to generate new data samples (Shorten and Khoshgoftaar, 2019) artificially. This is commonly known as *data augmentation (DA)* and relies on the property that for a given set of transformations $H = \{h : \mathcal{X} \rightarrow \mathcal{X}\}$, the data-generating distribution satisfies $\nu(\mathbf{y}|\mathbf{x}) = \nu(\mathbf{y}|h(\mathbf{x}))$ for all $h \in H$. In practice, not all the transformations are applied to every single data. Instead, a probability distribution (usually uniform) μ_H is defined over H , 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

$$\hat{L}_{\text{DA}}(D, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in [n]} \mathbb{E}_{h \sim \mu_H} [-\ln p(\mathbf{y}_i | h(\mathbf{x}_i), \boldsymbol{\theta})] , \quad (2.12)$$

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., 2016b) 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,

$$p_{\lambda}^{\text{DA}}(\boldsymbol{\theta} | D) = \arg \min_{\rho} \mathbb{E}_{\rho} [n \hat{L}_{\text{DA}}(D, \boldsymbol{\theta})] + \frac{1}{\lambda} \text{KL}(\rho(\boldsymbol{\theta} | D), p(\boldsymbol{\theta})) . \quad (2.13)$$

This is similar to Equation 2.3 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 Equation 2.2 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 $p_{\lambda=1}^{\text{DA}}$ rather than the Bayes loss, as this will help us in understanding this puzzling phenomenon.

2.6.1 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,

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

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 2.1 to the expected Gibbs loss by considering its derivative $\frac{d}{d\lambda}G(p_\lambda)$ at $\lambda = 1$, which can be represented as follows:

$$\frac{d}{d\lambda}G(p_\lambda)|_{\lambda=1} = -\text{COV}_{p_{\lambda=1}}(n\hat{L}(D, \boldsymbol{\theta}), L(\boldsymbol{\theta})). \quad (2.14)$$

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 2.8.3.

With this extended definition, if Equation 2.14 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 derivative of the expected Gibbs loss for the DA-tempered posterior exhibits a more negative trend at $\lambda = 1$, i.e., if $\frac{d}{d\lambda}G(p_\lambda^{\text{DA}})|_{\lambda=1} < \frac{d}{d\lambda}G(p_\lambda)|_{\lambda=1}$. This condition can be equivalently stated as

$$\text{COV}_{p_{\lambda=1}^{\text{DA}}}(n\hat{L}_{\text{DA}}(D, \boldsymbol{\theta}), L(\boldsymbol{\theta})) > \text{COV}_{p_{\lambda=1}}(n\hat{L}(D, \boldsymbol{\theta}), L(\boldsymbol{\theta})) > 0. \quad (2.15)$$

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 H used to generate the augmented dataset are introducing *valuable information* about the data-generating process.

Figure 2.6 clearly illustrates such situations. Figure 2.6b shows that, compared to Figure 2.6a, the standard DA, which makes use of the invariances inherent in the data-generating distribution, induces a CPE on the Gibbs loss.

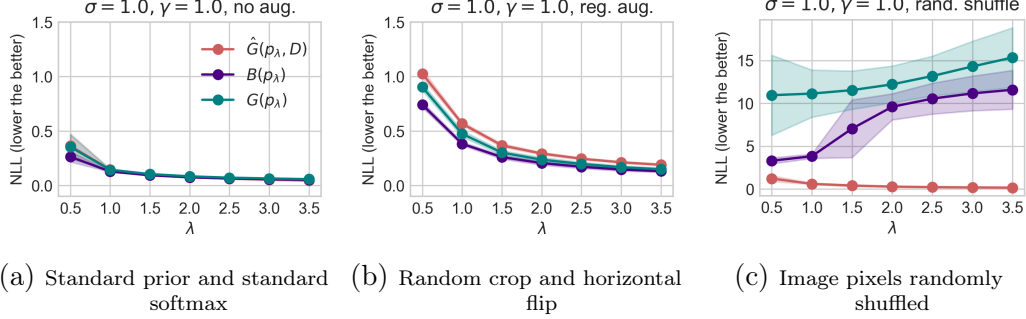


Figure 2.6: **Experimental illustrations for the arguments in Section 2.6 using small CNN via SGLD on MNIST.** We show similar results on Fashion-MNIST with small CNN and CIFAR-10(0) with ResNet-18 in Appendix 2.8.5. Figures 2.6a to 2.6c illustrate the arguments in Section 2.6. Figure 2.6a uses the standard prior ($\sigma = 1$) and the standard softmax ($\gamma = 1$) for the likelihood without applying DA. Figure 2.6b follows the setup as in Figure 2.6a but with standard DA applied, while Figure 2.6c 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 2.8.5.

Thus, the condition in Equation 2.15 holds by definition. On the other hand, Figure 2.6c 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 derivative of the Gibbs loss is positive, and Equation 2.15 holds in the opposite direction. These findings align perfectly with the explanations provided above, showing that DA induces a stronger underfitting.

2.6.2 Data augmentation and CPE on the Bayes loss

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

$$\frac{d}{d\lambda} B(p_\lambda)|_{\lambda=1} = -\text{COV}_{p_{\lambda=1}} \left(n\hat{L}(D, \boldsymbol{\theta}), S_{p_{\lambda=1}}(\boldsymbol{\theta}) \right), \quad (2.16)$$

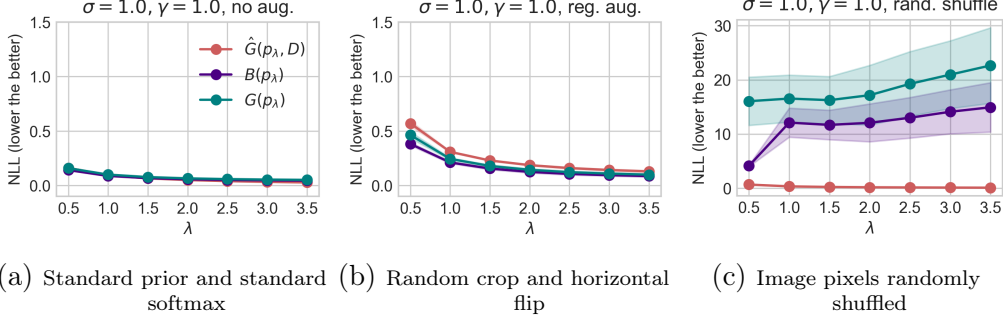


Figure 2.7: **Experimental illustrations for the arguments in Section 2.6 using large CNN via SGLD on MNIST. We show similar results on Fashion-MNIST with large CNN and CIFAR-10(0) with ResNet-50 in Appendix 2.8.5.** The experiment setup is similar to the setups in Figure 2.6 but with a large CNN. Please refer to Appendix 2.8.5 for further details on the model.

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

$$S_\rho(\boldsymbol{\theta}) = -\mathbb{E}_\nu \left[\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\mathbb{E}_\rho[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]} \right]. \quad (2.17)$$

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 $\nu(\mathbf{y}, \mathbf{x})$. 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 $S_\rho(\boldsymbol{\theta}) > -1$ (i.e., the lower the better). The derivations of the above equations are given in Appendix 2.8.3.

According to Definition 2.1 and Equation 2.16, DA will induce a stronger CPE if and only if the following condition is satisfied:

$$\text{COV}_{p_{\lambda=1}^{\text{DA}}} \left(n\hat{L}_{\text{DA}}(D, \boldsymbol{\theta}), S_{p_{\lambda=1}^{\text{DA}}}(\boldsymbol{\theta}) \right) > \text{COV}_{p_{\lambda=1}} \left(n\hat{L}(D, \boldsymbol{\theta}), S_{p_{\lambda=1}}(\boldsymbol{\theta}) \right). \quad (2.18)$$

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 2.6.

Finally, comparing Figure 2.6b with Figure 2.7b, we also notice that using a larger neural network enables us to mitigate the CPE because we reduce the underfitting introduced by DA.

2.6.3 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, as precisely defined by Equation 2.15 and Equation 2.18.

Some argue that when using DA, we are not using a proper likelihood function (Izmailov et al., 2021), which could be a problem. Recent works (Nabarro et al., 2022) have developed principle likelihood functions that integrate DA-based approaches, hoping to remove CPE. However, they find that CPE still persists. 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 2.15 and 2.18. 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 2.6c illustrates this situation. Using the analysis given in Section 2.4, 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.

2.7 Conclusions and limitations

Our research contributes to understanding the cold posterior effect (CPE) and its implications for Bayesian deep learning in several ways. Firstly, we theoretically demonstrate that the presence of the CPE implies that the Bayesian posterior is underfitting. Secondly, by building on Zeno et al. (2021), we show that, in general, any tempered posterior can be considered as a proper Bayesian posterior with an alternative likelihood and prior distribution jointly parametrized by T , beyond merely the case of classification. Hence, fine-tuning the temperature parameter T serves as an effective and theoretically sound approach to addressing the underfitting of the Bayesian posterior. Furthermore, we comprehensively discuss the interplay between several factors and CPE, including the use of approximate versus exact inference, model misspecification, and the size of the model and samples. Finally, our analysis in Section 2.6 reveals that data augmentation exacerbates the CPE by intensifying underfitting. This occurs because augmented data provides richer and more reliable information, 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 on the 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.

As a limitation of this work, we want to highlight that the characterization of CPE proposed here is defined only as the local change of Bayes loss at $\lambda = 1$. This approach does not account for scenarios where significant decreases in Bayes loss at other λ values might also indicate the presence of CPE. We believe that our theoretical analysis could be expanded to include these cases as well.

2.8 Appendix

2.8.1 Proofs for Section 2.3

In this section, we provide the proofs for Section 2.3 in the following order. We first prove the derivative of the empirical Gibbs loss in Proposition 2.1.

Then, we show in Proposition 2.6 that for meaningful posteriors (depends on training data), the derivative won't be zero. Before proving Proposition 2.2 and Theorem 2.3, we first provide Proposition 2.7, stating an alternative expression of the derivative of the Bayes loss. The proofs of Proposition 2.2 and Theorem 2.3 then follow from that.

2.8.1.1 Proof of Proposition 2.1

We first show a slightly more general result of $\frac{d}{d\lambda} \mathbb{E}_{p_\lambda}[f(\boldsymbol{\theta})]$ for any function $f(\boldsymbol{\theta})$ that is independent of λ . Recall that the posterior $p_\lambda(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^\lambda p(\boldsymbol{\theta})$. With the fact that $\frac{d}{d\lambda} (p(D|\boldsymbol{\theta})^\lambda p(\boldsymbol{\theta})) = \ln(p(D|\boldsymbol{\theta})) p(D|\boldsymbol{\theta})^\lambda p(\boldsymbol{\theta})$, the derivative

$$\frac{d}{d\lambda} \mathbb{E}_{p_\lambda}[f(\boldsymbol{\theta})] = \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta}) f(\boldsymbol{\theta})] - \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})] \mathbb{E}_{p_\lambda}[f(\boldsymbol{\theta})] = \text{COV}_{p_\lambda}(\ln p(D|\boldsymbol{\theta}), f(\boldsymbol{\theta})), \quad (2.19)$$

where we denote $\text{COV}(X, Y)$ as the covariance of X and Y . Hence, the derivative of the empirical Gibbs loss

$$\begin{aligned} \frac{d}{d\lambda} \hat{G}(p_\lambda, D) &= \frac{d}{d\lambda} \mathbb{E}_{p_\lambda}[-\ln p(D|\boldsymbol{\theta})] \\ &= \text{COV}_{p_\lambda}(\ln p(D|\boldsymbol{\theta}), -\ln p(D|\boldsymbol{\theta})) \\ &= -\mathbb{V}_{p_\lambda}(\ln p(D|\boldsymbol{\theta})). \end{aligned}$$

2.8.1.2 Proposition 2.6

Proposition 2.6. *For any $\lambda > 0$ and $D \neq \emptyset$, if the tempered posterior $p_\lambda(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})^\lambda p(\boldsymbol{\theta})$ satisfies $\mathbb{V}_{p_\lambda}(\ln p(D|\boldsymbol{\theta})) = 0$, then, $p_\lambda(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta})$.*

Proof. First of all, note that the tempered posterior is defined as

$$p_\lambda(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})^\lambda p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})^\lambda p(\boldsymbol{\theta})}.$$

Then,

$$\mathbb{V}_{p_\lambda}(\ln p(D|\boldsymbol{\theta})) = 0 \implies \int_{\boldsymbol{\theta}} p_\lambda(\boldsymbol{\theta}|D) (\ln p(D|\boldsymbol{\theta}) - \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})])^2 = 0$$

Thus, for any $\boldsymbol{\theta} \in \text{supp}(p_\lambda)$, it verifies that

$$\ln p(D|\boldsymbol{\theta}) = \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})].$$

That is, $\ln p(D|\boldsymbol{\theta})$ is constant in the support of p_λ . Let c denote such constant, then

$$p_\lambda(\boldsymbol{\theta}|D) = \frac{e^{c\lambda}p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} e^{c\lambda}p(\boldsymbol{\theta})} = \frac{e^{c\lambda}p(\boldsymbol{\theta})}{e^{c\lambda} \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})} = p(\boldsymbol{\theta}).$$

□

2.8.1.3 Proof of Proposition 2.2 and Theorem 2.3

In order to prove Proposition 2.2 and Theorem 2.3, we first show in Proposition 2.7 that the derivative of the Bayes loss of the tempered posterior p_λ can be expressed by the difference between the empirical Gibbs loss of \bar{p}_λ and the empirical Gibbs loss of p_λ .

Proposition 2.7. *The derivative of the Bayes loss of the tempered posterior p_λ can be expressed by*

$$\frac{d}{d\lambda}B(p_\lambda) = \hat{G}(\bar{p}_\lambda, D) - \hat{G}(p_\lambda, D). \quad (2.20)$$

Proof. By definition,

$$\frac{d}{d\lambda}B(p_\lambda) = \frac{d}{d\lambda}\mathbb{E}_\nu[-\ln \mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]] = -\mathbb{E}_\nu \left[\frac{d}{d\lambda} \ln \mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] \right],$$

where

$$\frac{d}{d\lambda} \ln \mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] = \frac{\frac{d}{d\lambda} \mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]}{\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]} = \frac{\text{COV}_{p_\lambda}(\ln p(D|\boldsymbol{\theta}), p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}))}{\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]}$$

due to Equation 2.19. By expanding the covariance, the above formula further equals to

$$\begin{aligned} & \frac{\mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] - \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})]\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]}{\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]} \\ &= \mathbb{E}_{\bar{p}_\lambda}[\ln p(D|\boldsymbol{\theta})] - \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})], \end{aligned}$$

where the probability distribution $\tilde{p}_\lambda(\boldsymbol{\theta}|D, (\mathbf{y}, \mathbf{x})) \propto p_\lambda(\boldsymbol{\theta}|D)p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. Put everything together, we have

$$\frac{d}{d\lambda}B(p_\lambda) = \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})] - \mathbb{E}_\nu[\mathbb{E}_{\tilde{p}_\lambda}[\ln p(D|\boldsymbol{\theta})]] = \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})] - \mathbb{E}_{\bar{p}_\lambda}[\ln p(D|\boldsymbol{\theta})], \quad (2.21)$$

where

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

The last equality is because

$$\begin{aligned} \mathbb{E}_\nu[\mathbb{E}_{\tilde{p}_\lambda}[\ln p(D|\boldsymbol{\theta})]] &= \int_{(\mathbf{y}, \mathbf{x})} \nu(\mathbf{y}, \mathbf{x}) \int_{\boldsymbol{\theta}} \tilde{p}_\lambda(\boldsymbol{\theta}|D, (\mathbf{y}, \mathbf{x})) \ln p(D|\boldsymbol{\theta}) d\boldsymbol{\theta} d(\mathbf{y}, \mathbf{x}) \\ &= \int_{\boldsymbol{\theta}} \int_{(\mathbf{y}, \mathbf{x})} \nu(\mathbf{y}, \mathbf{x}) \tilde{p}_\lambda(\boldsymbol{\theta}|D, (\mathbf{y}, \mathbf{x})) d(\mathbf{y}, \mathbf{x}) \ln p(D|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \mathbb{E}_\nu[\tilde{p}_\lambda(\boldsymbol{\theta}|D, (\mathbf{y}, \mathbf{x}))] \ln p(D|\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \mathbb{E}_{\bar{p}_\lambda}[\ln p(D|\boldsymbol{\theta})]. \end{aligned}$$

The last expression in Equation 2.21 further equals to $\hat{G}(\bar{p}_\lambda, D) - \hat{G}(p_\lambda, D)$ by definition. \square

2.8.1.3.1 Proof of Proposition 2.2

Note that for any distribution ρ , we have $\hat{G}(\rho, D) := \mathbb{E}_\rho - \ln p(D|\boldsymbol{\theta}) \geq \min_{\boldsymbol{\theta}} - \ln p(D|\boldsymbol{\theta})$. On the other hand, Proposition 2.7 together with Definition 2.1 give that the CPE takes place if and only if

$$\frac{d}{d\lambda} B(p_\lambda)|_{\lambda=1} = \hat{G}(\bar{p}_{\lambda=1}, D) - \hat{G}(p_{\lambda=1}, D) < 0.$$

Therefore, it is not possible to have $\hat{G}(p_{\lambda=1}, D) \not\geq \min_{\boldsymbol{\theta}} - \ln p(D|\boldsymbol{\theta})$ and, at the same time, $\hat{G}(\bar{p}^{\lambda=1}, D) < \hat{G}(p_{\lambda=1}, D)$ because $\hat{G}(\bar{p}^{\lambda=1}, D) \geq \min_{\boldsymbol{\theta}} - \ln p(D|\boldsymbol{\theta})$.

2.8.1.3.2 Proof of Theorem 2.3

It's easy to see from Proposition 2.7 that

$$\frac{d}{d\lambda} B(p_\lambda)|_{\lambda=1} = \hat{G}(\bar{p}_{\lambda=1}, D) - \hat{G}(p_{\lambda=1}, D) = 0$$

if and only if $\hat{G}(\bar{p}_{\lambda=1}, D) = \hat{G}(p_{\lambda=1}, D)$.

2.8.2 Proofs for Section 2.4

2.8.2.1 Proof of Proposition 2.4

First of all, by the definition in Equation 2.2, and assuming a data-independent prior $p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta})$, the tempered posterior is given by

$$p_\lambda(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda p(\boldsymbol{\theta}),$$

where the tempered likelihood fully factorizes as $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda = \prod_{(\mathbf{y}, \mathbf{x}) \in (\mathbf{Y}, \mathbf{X})} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda$. Let a similar but \mathbf{y} -independent function $k(\boldsymbol{\theta}, \mathbf{X}, \lambda) = \prod_{\mathbf{x} \in \mathbf{X}} \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y}$. Therefore, $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda p(\boldsymbol{\theta}) = \frac{p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda}{k(\boldsymbol{\theta}, \mathbf{X}, \lambda)} (k(\boldsymbol{\theta}, \mathbf{X}, \lambda) p(\boldsymbol{\theta}))$, where we can let the new prior

$$q(\boldsymbol{\theta}|\mathbf{X}, \lambda) \propto p(\boldsymbol{\theta}) k(\boldsymbol{\theta}, \mathbf{X}, \lambda) = p(\boldsymbol{\theta}) \prod_{\mathbf{x} \in \mathbf{X}} \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y},$$

and the new posterior

$$q(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \lambda) = \frac{p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda}{k(\boldsymbol{\theta}, \mathbf{X}, \lambda)} = \frac{\prod_{(\mathbf{y}, \mathbf{x}) \in (\mathbf{Y}, \mathbf{X})} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda}{\prod_{\mathbf{x} \in \mathbf{X}} \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y}} = \prod_{(\mathbf{y}, \mathbf{x}) \in (\mathbf{Y}, \mathbf{X})} q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}).$$

2.8.2.2 Proof of Proposition 2.5

The proof is made using differential entropy, i.e. assuming continuous target values \mathbf{y} . The only assumption is that Leibniz integral rule holds for $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$, verifying that

$$\frac{d}{d\lambda} \int (q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) d\mathbf{y} = \int \frac{d}{d\lambda} (q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) d\mathbf{y}.$$

In the case of supervised classification problems, we adopt the Shanon entropy, where equality holds naturally

$$\frac{d}{d\lambda} \sum_{\mathbf{y} \in \mathcal{Y}} (q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) = \sum_{\mathbf{y} \in \mathcal{Y}} \frac{d}{d\lambda} (q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)).$$

From the definition of differential entropy, we got that

$$H(q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) = - \int q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) d\mathbf{y}.$$

Thus, taking derivative w.r.t. λ and exchanging derivative and integral leads to the following expression

$$\begin{aligned}\frac{d}{d\lambda}H(q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) &= - \int \frac{d}{d\lambda}(q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) d\mathbf{y} \\ &= - \int (\ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) + 1) \frac{d}{d\lambda}q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) d\mathbf{y}.\end{aligned}$$

Using that $\int \frac{d}{d\lambda}q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)d\mathbf{y} = \frac{d}{d\lambda} \int q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)d\mathbf{y} = 0$, simplifies the expression as

$$\frac{d}{d\lambda}H(q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)) = - \int \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \frac{d}{d\lambda}q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) d\mathbf{y}.$$

Let us consider now the second term inside the integral. Using the derivative of the quotient rule leads to the following:

$$\begin{aligned}&\frac{d}{d\lambda}q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \\ &= \frac{d}{d\lambda} \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda}{\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y}} \\ &= \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y}} - \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) d\mathbf{y}}{(\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y})^2}.\end{aligned}$$

Where, using the definition of $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$, we got that

$$\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y}} = q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}),$$

and

$$\begin{aligned}&\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda \int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) d\mathbf{y}}{(\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda d\mathbf{y})^2} \\ &= q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \int q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) d\mathbf{y} \\ &= q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})].\end{aligned}$$

As a result, we got that

$$\begin{aligned}&\int \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \frac{d}{d\lambda}q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) d\mathbf{y} \\ &= \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)] - \mathbb{E}_q[\ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)] \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]\end{aligned}$$

Using $q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda)$ definition again:

$$\begin{aligned} & \int \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \frac{d}{d\lambda} q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) d\mathbf{y} \\ &= \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda}{\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda}] - \mathbb{E}_q[\ln \frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda}{\int p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda}] \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] \end{aligned}$$

Where, expanding the logarithms the denominators cancel each other, leading to

$$\begin{aligned} & \int \ln q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) \frac{d}{d\lambda} q(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}, \lambda) d\mathbf{y} \\ &= \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda] - \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})^\lambda] \mathbb{E}_q[\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] \\ &= \lambda \mathbb{V}(\ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})) \geq 0 \end{aligned}$$

As a result, the entropy is negative.

2.8.3 Proofs for Section 2.6

2.8.3.1 Proof of Equation 2.14

Note that

$$\frac{d}{d\lambda} G(p_\lambda) = \frac{d}{d\lambda} \mathbb{E}_{p_\lambda}[L(\boldsymbol{\theta})] = \text{COV}_{p_\lambda}(\ln p(D|\boldsymbol{\theta}), L(\boldsymbol{\theta})) = \text{COV}_{p_\lambda}(-\hat{L}(D, \boldsymbol{\theta}), L(\boldsymbol{\theta})),$$

where the second equality is by applying Equation 2.19. By taking $\lambda = 1$, we obtain the desired derivative.

2.8.3.2 Proof of Equation 2.16

Recall from the proof of Theorem 2.7 that

$$\frac{d}{d\lambda} B(p_\lambda) = \mathbb{E}_{p_\lambda}[\ln p(D|\boldsymbol{\theta})] - \mathbb{E}_{\bar{p}_\lambda}[\ln p(D|\boldsymbol{\theta})] = \mathbb{E}_{\bar{p}_\lambda}[\hat{L}(D, \boldsymbol{\theta})] - \mathbb{E}_{p_\lambda}[\hat{L}(D, \boldsymbol{\theta})],$$

where $\bar{p}_\lambda(\boldsymbol{\theta}|D) = \mathbb{E}_\nu[\tilde{p}_\lambda(\boldsymbol{\theta}|D, (\mathbf{y}, \mathbf{x}))]$ (Equation 2.6), and $\tilde{p}_\lambda(\boldsymbol{\theta}|D, (\mathbf{y}, \mathbf{x})) \propto p_\lambda(\boldsymbol{\theta}|D)p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ is the distribution obtained by updating the posterior p_λ with one new sample (\mathbf{y}, \mathbf{x}) . Therefore,

$$\mathbb{E}_{\bar{p}_\lambda}[\hat{L}(D, \boldsymbol{\theta})] = \mathbb{E}_\nu \mathbb{E}_{\bar{p}_\lambda}[\hat{L}(D, \boldsymbol{\theta})] = \mathbb{E}_\nu \left[\mathbb{E}_{p_\lambda} \left[\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]} \hat{L}(D, \boldsymbol{\theta}) \right] \right].$$

By Fubini's theorem, the above formula further equals to

$$\begin{aligned}\mathbb{E}_{p_\lambda} \left[\mathbb{E}_\nu \left[\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]} \hat{L}(D, \boldsymbol{\theta}) \right] \right] &= \mathbb{E}_{p_\lambda} \left[\mathbb{E}_\nu \left[\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]} \right] \hat{L}(D, \boldsymbol{\theta}) \right] \\ &= \mathbb{E}_{p_\lambda} \left[-S_{p_\lambda}(\boldsymbol{\theta}) \cdot \hat{L}(D, \boldsymbol{\theta}) \right].\end{aligned}$$

On the other hand, since

$$\mathbb{E}_{p_\lambda}[-S_{p_\lambda}(\boldsymbol{\theta})] = \mathbb{E}_{p_\lambda} \left[\mathbb{E}_\nu \left[\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]} \right] \right] = \mathbb{E}_\nu \left[\mathbb{E}_{p_\lambda} \left[\frac{p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{\mathbb{E}_{p_\lambda}[p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})]} \right] \right] = 1,$$

we have

$$\mathbb{E}_{p_\lambda}[\hat{L}(D, \boldsymbol{\theta})] = \mathbb{E}_{p_\lambda}[\hat{L}(D, \boldsymbol{\theta})] \mathbb{E}_{p_\lambda}[-S_{p_\lambda}(\boldsymbol{\theta})].$$

By putting them altogether,

$$\begin{aligned}\frac{d}{d\lambda} B(p_\lambda) &= \mathbb{E}_{p_\lambda} \left[-S_{p_\lambda}(\boldsymbol{\theta}) \cdot \hat{L}(D, \boldsymbol{\theta}) \right] - \mathbb{E}_{p_\lambda}[\hat{L}(D, \boldsymbol{\theta})] \mathbb{E}_{p_\lambda}[-S_{p_\lambda}(\boldsymbol{\theta})] \\ &= -\text{COV} \left(\hat{L}(D, \boldsymbol{\theta}), S_{p_\lambda}(\boldsymbol{\theta}) \right).\end{aligned}$$

2.8.4 Experiment details for Bayesian linear regression on synthetic data with exact inference

In this section we detail the settings of the toy experiment using synthetic data and exact Bayesian linear regression in Figures 2.2 and 2.3. We also show extra results of the derivative of Gibbs loss and Bayes loss w.r.t to λ approximated by samples.

To begin, we will outline the data-generating process for the synthetic data used in the experiment shown in Figures 2.2 and 2.3 and Figures 2.8 and 2.9. We sample x uniformly from the $[-1, 1]$ interval and pass it through a Fourier transformation to construct the input of the data. That is, for a sampled x , the input \mathbf{x} is constructed by a 10-dimensional Fourier basis function $\boldsymbol{\phi}(x) = [g_1(x), \dots, g_K(x)]^T$ for $K = 10$, where the basis functions are defined as follows: $g_1(x) = \frac{1}{\sqrt{2\pi}}$, and for other odd values of k , $g_k(x) = \frac{1}{\sqrt{\pi}} \sin(kx)$, whereas for even values of k , $g_k(x) = \frac{1}{\sqrt{\pi}} \cos(kx)$. The distribution of the output $y \in \mathbb{R}$ given an input \mathbf{x} , denoted as $\nu(y|\mathbf{x})$, follows a Normal

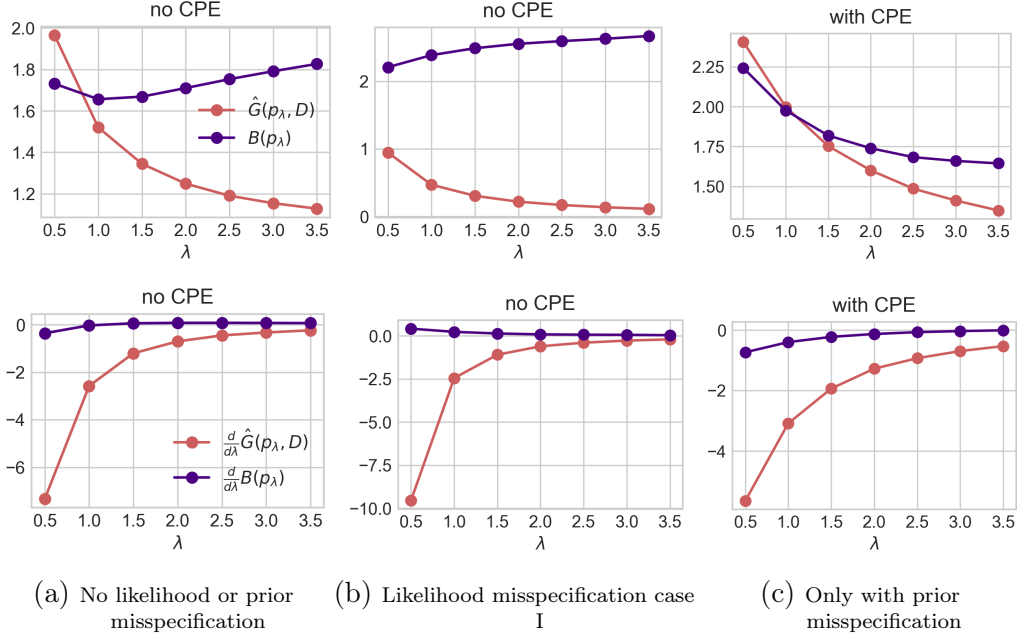


Figure 2.8: **The derivatives $\frac{d}{d\lambda}\hat{G}(p_\lambda, D)$ (Equation 2.5) and $\frac{d}{d\lambda}B(p_\lambda)$ (Equation 2.20) characterize the Gibbs loss $\hat{G}(p_\lambda, D)$ and the Bayes loss $B(p_\lambda)$ perfectly.**

distribution with mean $\mathbf{1}^T \mathbf{x}$ and variance 1.0, where $\mathbf{1}$ is an all-ones vector. That is, $\nu(y|\mathbf{x}) = \mathcal{N}(\mathbf{1}^T \mathbf{x}, 1.0)$.

In our experiment, the likelihood model and the prior model are defined differently for the four settings in Figures 2.2 and 2.3. To enable exact inference, both the likelihood and the prior are Gaussian, which gives a closed-form solution for the posterior predictive. This choice also provides convenience when studying the CPE: different values of λ on the likelihood term can be naturally absorbed into the Gaussian densities by adjusting the variance (dividing by λ) without hindering the exact inference step. We describe them in detail in the following.

1. No likelihood or prior misspecification: likelihood $p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^T \mathbf{x}, 1.0)$, prior $p(\boldsymbol{\theta}) = \mathcal{N}(0, 2)$. This is the baseline for comparison.
2. Likelihood misspecification case I: likelihood $p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^T \mathbf{x}, 0.15)$ (the order of Fourier transformation is $K = 20$, however note that it still contains the $K = 5$ data-generating process in its solution space),

prior $p(\boldsymbol{\theta}) = \mathcal{N}(0, 2)$. In this case, the model is misspecified in a way that it has a smaller variance than the data-generating process.

3. Likelihood misspecification case II: likelihood $p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^T \mathbf{x}, 3.0)$, prior $p(\boldsymbol{\theta}) = \mathcal{N}(0, 2)$. In this case, the model is misspecified in a way that it has a larger variance than the data-generating process. This is similar to one of the scenarios where CPE was found: the curated data has a lower aleatoric uncertainty than the model (Aitchison, 2021).
4. Only with prior misspecification: likelihood $p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^T \mathbf{x}, 1.0)$, prior $p(\boldsymbol{\theta}) = \mathcal{N}(0, 0.5)$. The prior is poorly specified in a way that it is tightly centered at 0 while the best $\boldsymbol{\theta}$ should be 1.

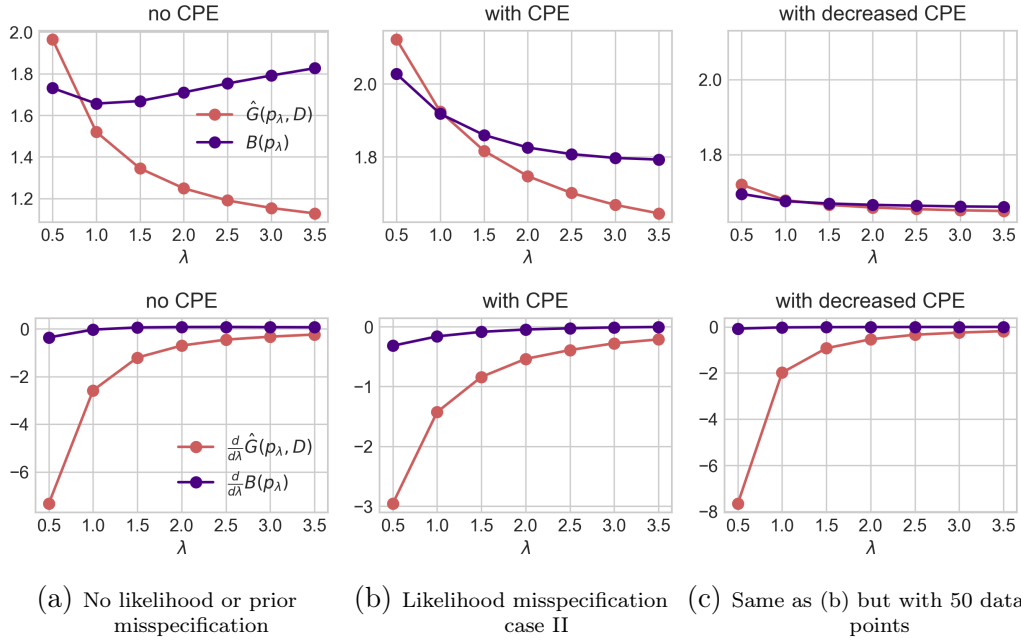


Figure 2.9: **Extended results of Figure 2.8 with more configurations of model misspecification.**

In all the experiments, every training set consists of only 5 samples. Since there are more parameters than the number of training data points, our setting falls within the “overparameterized” regime where CPE has been observed in Bayesian deep learning (Wenzel et al., 2020).

Continuing from Figures 2.2 and 2.3, where we show the Gibbs loss $\hat{G}(p_\lambda, D)$ (training) and the Bayes loss $B(p_\lambda)$ (testing) with respect to λ , we now show their derivatives $\frac{d}{d\lambda}\hat{G}(p_\lambda, D)$ (Equation 2.5) and $\frac{d}{d\lambda}B(p_\lambda)$ (Equation 2.20) respectively in Figures 2.8 and 2.9. Here the losses are included for a clearer depiction of the derivatives. To approximate the Bayes loss for generating the plot, we use 10000 data points sampled from the data-generating distribution. Also, the derivatives are approximated using 10000 samples from the exact posteriors. From Figures 2.8 and 2.9, we could clearly see that the derivatives perfectly characterize the losses in all four settings.

2.8.5 Experiment details for Bayesian neural networks on image data with approximate inference

In this section, we first present in Appendix 2.8.5.1 the architectures of the small and large CNNs used in this paper. As promised in the main text, we then provide results on additional image datasets trained with Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011) in Appendix 2.8.5.2.

2.8.5.1 Architectures of small/large CNN

2.8.5.1.1 Small CNN

The small CNN is similar to LeNet-5, but with 107786 parameters in total:

1. Convolutional layer 1. Input channels: 1 (assuming grayscale images), output channels: 6, kernel size: 5x5, padding: 2, activation: ReLU.
2. Average pooling layer 1. Kernel size: 2x2, stride: 2.
3. Convolutional layer 2. Input channels: 6, output channels: 16, kernel size: 5x5, padding: 2, activation: ReLU.
4. Average pooling layer 2. Kernel size: 2x2, stride: 2.
5. Flattening layer. Flattens the output from the previous layers.
6. Fully connected layer 1. Input features: 784 (16 channels * 7 * 7), output features: 120, activation: ReLU.

7. Fully connected layer 2. Input features: 120, output features: 84, activation: ReLU.
8. Fully connected layer 3 (output layer). Input features: 84, output features: num_classes (specified during instantiation).

2.8.5.1.2 Large CNN

The large CNN is similar to the small CNN, but with 545546 parameters in total:

1. Convolutional layer 1. Input channels: 1 (assuming grayscale images), output channels: 6, kernel size: 5x5, padding: 2, activation: ReLU.
2. Average pooling layer 1. Kernel size: 2x2, stride: 2.
3. Convolutional layer 2. Input channels: 6, output channels: 16, kernel size: 5x5, padding: 2, activation: ReLU.
4. Average pooling layer 2. Kernel size: 2x2, stride: 2.
5. Convolutional layer 3. Input channels: 16, output channels: 120, kernel size: 5x5, padding: 2, activation: ReLU.
6. Flattening layer. Flattens the output from the previous layers.
7. Fully connected layer 1. Input features: 5880 ($120 \text{ channels} \times 7 \times 7$), output features: 84, activation: ReLU.
8. Fully connected layer 2 (output layer). Input features: 84, output features: num_classes (specified during instantiation).

In all the convolutional layers, no stride = 1 and padding is set to *same*.

2.8.5.2 Stochastic gradient Langevin dynamics (SGLD)

Our experiments using SGLD are categorized into 4 groups:

1. Bayesian CNNs (small and large) on MNIST (Figures 2.4 - 2.7 in the main text)
2. Bayesian CNNs (small and large) on Fashion-MNIST (Appendix 2.8.5.2.1)

3. Bayesian ResNets (18 and 50) on CIFAR-10 (Appendix 2.8.5.2.2)
4. Bayesian ResNets (18 and 50) on CIFAR-100 (Appendix 2.8.5.2.3)

where each group evaluates the effect of underfitting on a small model and a large model. Note that as we follow the standard ResNet-18 and ResNet-50, the details of the architectures are omitted. They have around 11 million and 23 million parameters, respectively. We implement with PyTorch (Paszke et al., 2019) and train the model using cyclical learning rate SGLD (cSGLD) (Zhang et al., 2019b) for 1000 epochs. We set the learning rate to $1e-6$ with a momentum term of 0.99. We run cSGLD for 10 trials and collect 10 samples for each trial. Experiments were conducted on NVIDIA A100 GPU, with each trial taking around 30 hours.

2.8.5.2.1 Small and large CNNs via SGLD on Fashion-MNIST

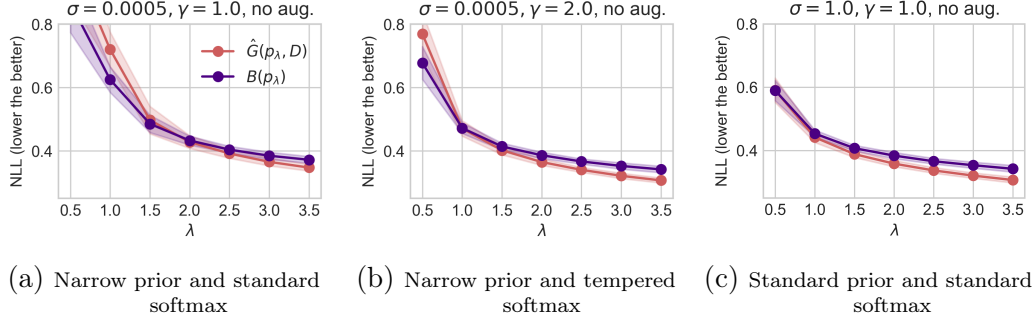


Figure 2.10: Extended results of Figure 2.4 using small CNN via SGLD on Fashion-MNIST.

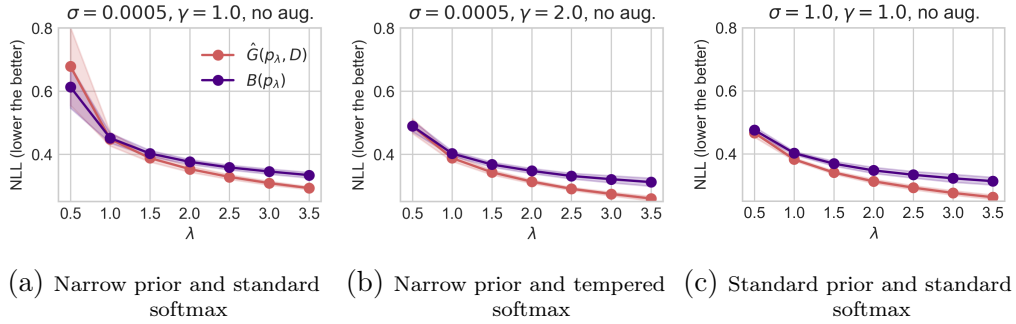


Figure 2.11: Extended results of Figure 2.5 using large CNN via SGLD on Fashion-MNIST.

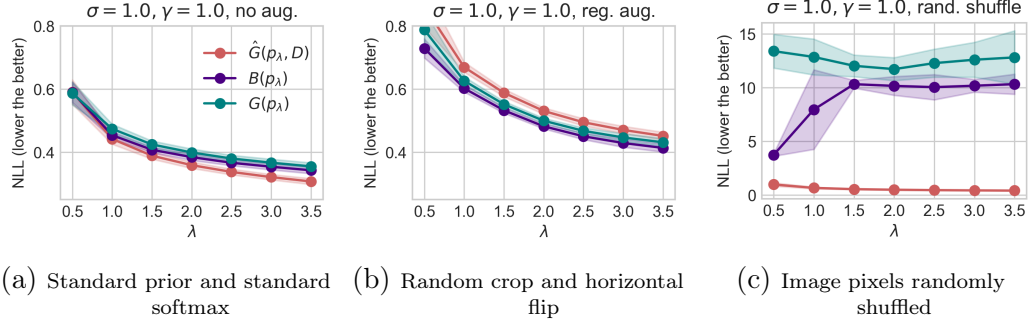


Figure 2.12: Extended results of Figure 2.6 using small CNN via SGLD on Fashion-MNIST.

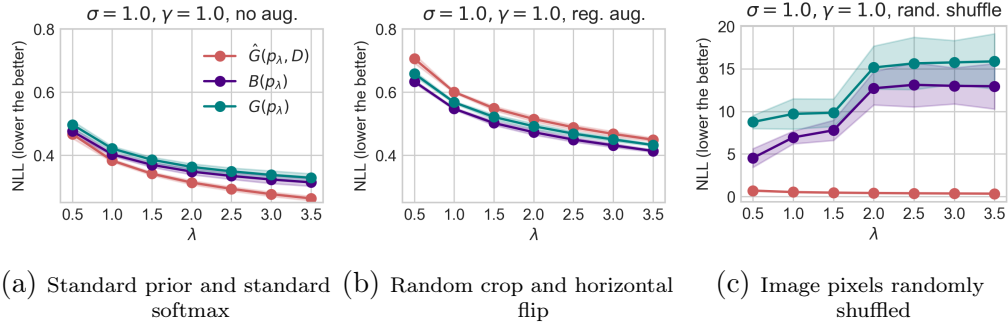


Figure 2.13: Extended results of Figure 2.7 using large CNN via SGLD on Fashion-MNIST.

2.8.5.2.2 ResNet-18 and ResNet-50 via SGLD on CIFAR-10

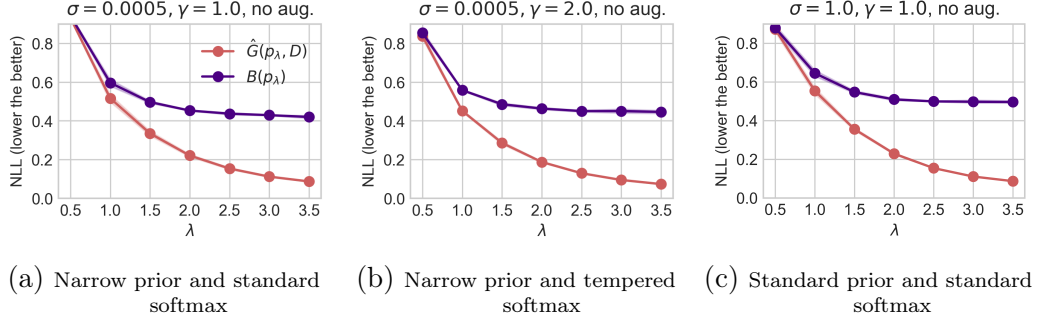


Figure 2.14: Extended results of Figure 2.4 using ResNet-18 via SGLD on CIFAR-10.

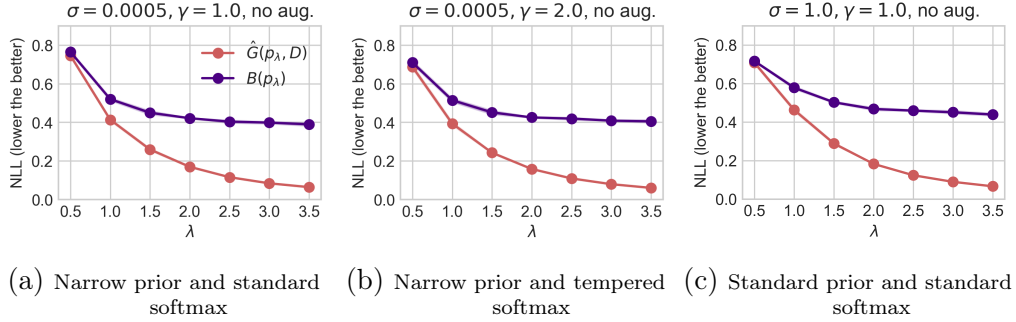


Figure 2.15: Extended results of Figure 2.5 using ResNet-50 via SGLD on CIFAR-10.

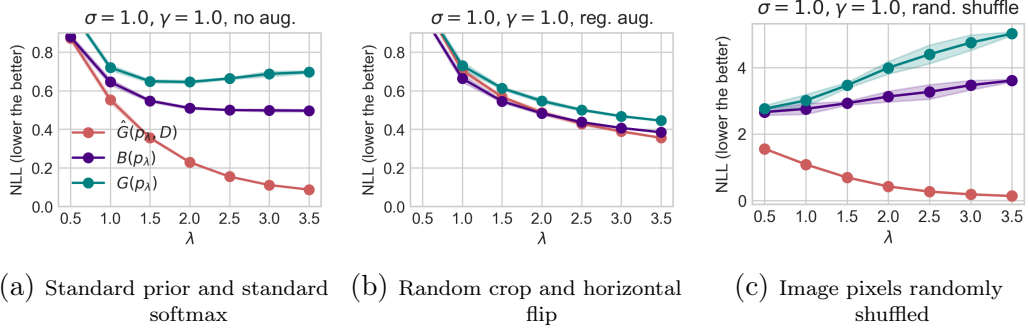


Figure 2.16: Extended results of Figure 2.6 using ResNet-18 via SGLD on CIFAR-10.

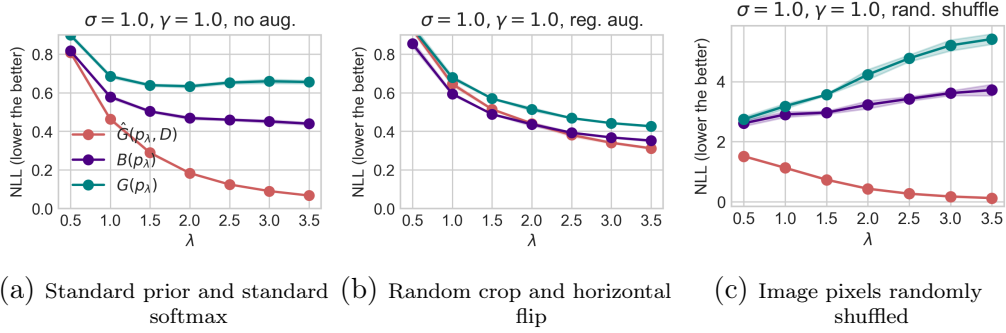


Figure 2.17: Extended results of Figure 2.7 using ResNet-50 via SGLD on CIFAR-10.

2.8.5.2.3 ResNet-18 and ResNet-50 via SGLD on CIFAR-100

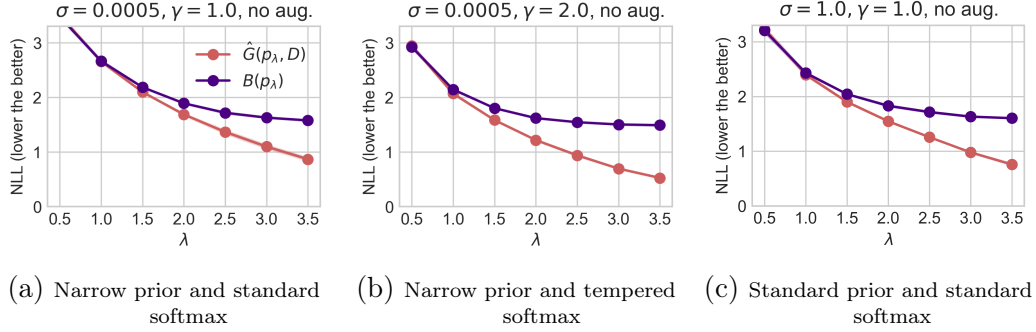


Figure 2.18: Extended results of Figure 2.4 using ResNet-18 via SGLD on CIFAR-100.

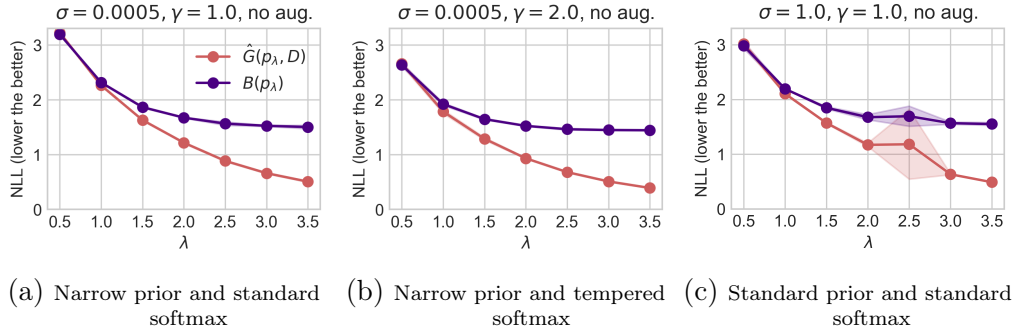


Figure 2.19: Extended results of Figure 2.5 using ResNet-50 via SGLD on CIFAR-100.

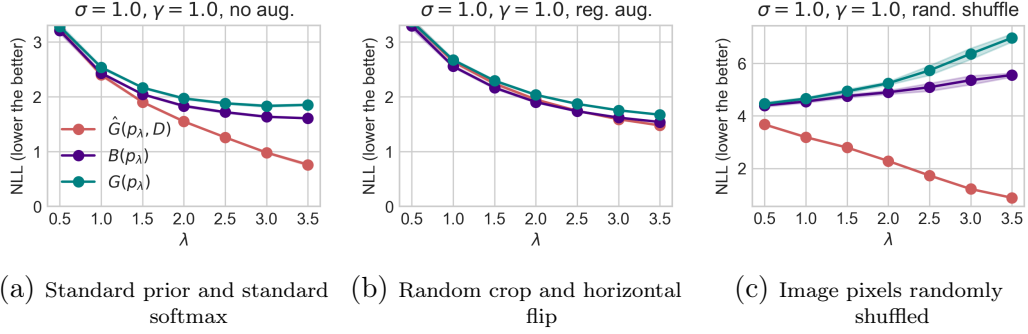


Figure 2.20: Extended results of Figure 2.6 using ResNet-18 via SGLD on CIFAR-100.

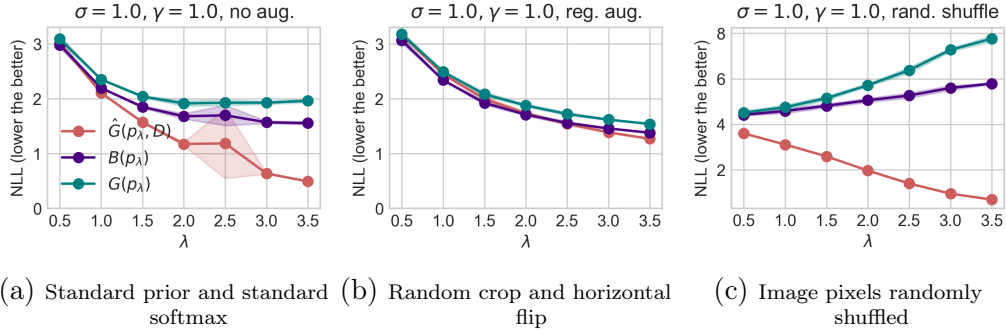


Figure 2.21: Extended results of Figure 2.7 using ResNet-50 via SGLD on CIFAR-100.

Chapter 3

Recursive PAC-Bayes: A Frequentist Approach to Sequential Prior Updates

The work presented in this chapter is based on a paper that is currently under review in *NeurIPS* 2024:

Yi-Shan Wu, Yijie Zhang, Badr-Eddine Chérif-Abdellatif, and Yevgeny Seldin. Recursive PAC-Bayes: A frequentist approach to sequential prior updates with no information loss. *arXiv preprint arXiv:2405.14681*, 2024. URL <https://arxiv.org/abs/2405.14681>.

Abstract

PAC-Bayesian analysis is a frequentist framework for incorporating prior knowledge into learning. It was inspired by Bayesian learning, which allows sequential data processing and naturally turns posteriors from one processing step into priors for the next. However, despite two and a half decades of research, the ability to update priors sequentially without losing confidence information along the way remained elusive for PAC-Bayes. While PAC-Bayes allows construction of data-informed priors, the final confidence intervals depend only on the number of points that were not used for the construction of the prior, whereas confidence information in the prior, which is related to the number of points used to construct the prior, is lost. This limits the possibility and benefit of sequential prior updates, because the final bounds depend only on the size of the final batch.

We present a novel and, in retrospect, surprisingly simple and powerful PAC-Bayesian procedure that allows sequential prior updates with no information loss. The procedure is based on a novel decomposition of the expected loss of randomized classifiers. The decomposition rewrites the loss of the posterior as an excess loss relative to a downscaled loss of the prior plus the downscaled loss of the prior, which is bounded recursively. As a side result, we also present a generalization of the split-kl and PAC-Bayes-split-kl inequalities to discrete random variables, which we use for bounding the excess losses, and which can be of independent interest. In empirical evaluation the new procedure significantly outperforms state-of-the-art.

3.1 Introduction

PAC-Bayesian analysis was born from an attempt to derive frequentist generalization guarantees for Bayesian-style prediction rules (Shawe-Taylor and Williamson, 1997; McAllester, 1998). The motivation was to provide a way to incorporate prior knowledge into the frequentist analysis of generalization. PAC-Bayesian bounds provide high-probability generalization guarantees for randomized classifiers. A randomized classifier is defined by a distribution ρ on a set of prediction rules \mathcal{H} , which is used to sample a prediction rule each time a prediction is to be made. Bayesian posterior is an example of a randomized classifier, whereas PAC-Bayesian bounds hold generally for all randomized classifiers. Prior knowledge is encoded through a prior distribution

π on \mathcal{H} , and the complexity of a posterior distribution ρ is measured by the Kullback-Leibler (KL) divergence from the prior, $\text{KL}(\rho\|\pi)$. PAC-Bayesian generalization guarantees are optimized by posterior distributions ρ that optimize a trade-off between empirical data fit and divergence from the prior in the KL sense.

Selection of a “good” prior plays an important role in the PAC-Bayesian bounds. If one manages to foresee which prediction rules are likely to produce low prediction error and allocate a higher prior mass for them, then the bounds are tighter, because the posterior only needs to make a small deviation from the prior. But if the prior mass on well-performing prediction rules is small, the bounds are loose. A major technique to design good priors is to use part of the data to estimate a good prior and the rest of the data to compute a PAC-Bayes bound. It is known as data-dependent or data-informed priors (Ambroladze et al., 2007). However, all existing approaches to data-informed priors have three major disadvantages. The first is that the bounds are computed on “the rest of the data” that were not used in construction of the prior. Thus, the sample size in the bounds is only a fraction of the total sample size. Therefore, empirically data-informed priors are not always helpful. In many cases starting with an uninformed prior and using all the data to compute the posterior and the bound turns to be superior to sacrificing part of the data for prior construction (Ambroladze et al., 2007; Mhammedi et al., 2019). The second disadvantage is that all the confidence information about the prior is lost in the process. In particular, a prior trained on a few data points is treated in the same way as a prior trained on a lot of data. And a third related disadvantage is that sequential data processing is detrimental, because the bounds only depend on the size of the last chunk and all the confidence information from processing earlier chunks is lost in the process.

Our main contribution is a new (and simple) way of decomposing the loss of a randomized classifier defined by the posterior. We write it as an excess loss relative to a downscaled loss of the randomized classifier defined by the prior plus the downscaled loss of the randomized classifier defined by the prior. The excess loss can be bounded using PAC-Bayes-Empirical-Bernstein-style inequalities (Tolstikhin and Seldin, 2013; Mhammedi et al., 2019; Wu et al., 2021; Wu and Seldin, 2022), whereas the loss of the randomized classifier defined by the prior can be bounded recursively. The recursive bound can both use the data used for construction of the prior and “the rest of the data”, and thereby preserves confidence information on the prior. Our contribution stands out relative to all prior work on PAC-Bayes, and in fact all prior

work on frequentist generalization bounds, because it makes sequential data processing and sequential prior updates meaningful and beneficial.

We note that while several recent papers experimented with sequential posterior updates by using martingale-style analysis, in all these works the prior remained fixed and only the posterior was changing (Chugg et al., 2023; Biggs and Guedj, 2023; Rodríguez-Gálvez et al., 2024). Another line of work used tools from online learning to derive PAC-Bayesian bounds (Jang et al., 2023), and in this context Haddouche and Guedj (2023) have used sequential prior updates, but their bounds hold for a uniform aggregation of sequentially constructed posteriors, which is different from standard posteriors studied in our work. The confidence bounds in their work come primarily from aggregation rather than confidence in individual posteriors in the sequence. Our work is the first one allowing sequential prior updates without loss of confidence information.

An additional side contribution of independent interest is a generalization of the split-kl and PAC-Bayes-split-kl inequalities of Wu and Seldin (2022) from ternary to general discrete random variables. It is based on a novel representation of discrete random variables as a superposition of Bernoulli random variables.

The paper is organized in the following way. In Section 3.2 we briefly survey the evolution of data-informed priors in PAC-Bayes and present our main idea behind Recursive PAC-Bayes; in Section 3.3 we present our generalization of the split-kl and PAC-Bayes-split-kl inequalities, which are later used to bound the excess losses; in Section 3.4 we present the Recursive PAC-Bayes bound; in Section 3.5 we present an empirical evaluation; and in Section 3.6 we conclude with a discussion.

3.2 The evolution of data-informed priors and the idea of recursive PAC-Bayes

In this section we briefly survey the evolution of data-informed priors, and then present our construction of Recursive PAC-Bayes. We consider the standard binary classification setting, with \mathcal{X} being a sample space, \mathcal{Y} a label space, \mathcal{H} a set of prediction rules $h : \mathcal{X} \rightarrow \mathcal{Y}$, and $\ell(h(X), Y) = \mathbf{1}(h(X) \neq Y)$ the zero-one loss function, where $\mathbf{1}(\cdot)$ denotes the indicator function. We let \mathcal{D} denote a distribution on $\mathcal{X} \times \mathcal{Y}$ and $S = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ an

i.i.d. sample from \mathcal{D} . Let $L(h) = \mathbb{E}_{(X,Y) \sim \mathcal{D}}[\ell(h(X), Y)]$ be the expected and $\hat{L}(h, S) = \frac{1}{n} \sum_{i=1}^n \ell(h(X_i), Y_i)$ the empirical loss.

Let ρ be a distribution on \mathcal{H} . A *randomized classifier* associated with ρ samples a prediction rule h according to ρ for each sample $X \in \mathcal{X}$, and applies it to make a prediction $h(X)$. The expected loss of such randomized classifier, which we call ρ , is $\mathbb{E}_{h \sim \rho}[L(h)]$ and the empirical loss is $\mathbb{E}_{h \sim \rho}[\hat{L}(h, S)]$. For brevity we use $\mathbb{E}_\rho[\cdot]$ to denote $\mathbb{E}_{h \sim \rho}[\cdot]$.

We use $\text{KL}(\rho \parallel \pi)$ to denote the Kullback-Leibler divergence between two probability distributions, ρ and π (Cover and Thomas, 2006). For $p, q \in [0, 1]$ we further use $\text{kl}(p \parallel q) = \text{KL}((1-p, p) \parallel (1-q, q))$ to denote the Kullback-Leibler divergence between two Bernoulli distributions with biases p and q .

The goal of PAC-Bayes is to bound $\mathbb{E}_\rho[L(h)]$. Below we present how the bounds on $\mathbb{E}_\rho[L(h)]$ have evolved. In Figure 3.2 we also provide a graphical illustration of the evolution.

3.2.1 Uninformed priors

Early work on PAC-Bayes used *uninformed priors* (McAllester, 1998). An uninformed prior π is a distribution on \mathcal{H} that is independent of the data S . A classical, and still one of the tightest bounds, is the following.

Theorem 3.1 (PAC-Bayes-kl Inequality, Seeger, 2002, Maurer, 2004). *For any probability distribution π on \mathcal{H} that is independent of S and any $\delta \in (0, 1)$:*

$$\mathbb{P}\left(\exists \rho \in \mathcal{P} : \text{kl}\left(\mathbb{E}_\rho[\hat{L}(h, S)] \parallel \mathbb{E}_\rho[L(h)]\right) \geq \frac{\text{KL}(\rho \parallel \pi) + \ln(2\sqrt{n}/\delta)}{n}\right) \leq \delta,$$

where \mathcal{P} is the set of all probability distributions on \mathcal{H} , including those dependent on S .

A posterior ρ that minimizes $\mathbb{E}_\rho[L(h)]$ has to balance between allocating higher mass to prediction rules h with small $\hat{L}(h, S)$ and staying close to π in the $\text{KL}(\rho \parallel \pi)$ sense. Since π has to be independent of S , typical uninformed priors aim “to leave maximal options open” for ρ by staying close to uniform.

3.2.2 Data-informed priors

Ambroladze et al. (2007) proposed to split the data S into two disjoint sets, $S = S_1 \cup S_2$, and use S_1 to construct a *data-informed prior* π and compute a

bound on $\mathbb{E}_\rho[L(h)]$ using π and S_2 . Since in this approach π is independent of S_2 , Theorem 3.1 can be applied. The advantage is that π can use S_1 to give higher mass to promising classifiers, thus relaxing the regularization pressure $\text{KL}(\rho||\pi)$ and making it easier for ρ to allocate even higher mass to well-performing classifiers (those with small $\hat{L}(h, S_2)$). The disadvantage is that the sample size in the bound (the n in the denominator) decreases from the size of S to the size of S_2 . Indeed, Ambroladze et al. observed that the sacrifice of S_1 for prior construction does not always pay off.

3.2.3 Data-informed priors + excess loss

Mhammedi et al. (2019) observed that if we have already sacrificed S_1 for the construction of π , we could also use it to construct a reference prediction rule h^* , typically an Empirical Risk Minimizer (ERM) on S_1 . They then employed the decomposition

$$\mathbb{E}_\rho[L(h)] = \mathbb{E}_\rho[L(h) - L(h^*)] + L(h^*)$$

and used S_2 to give a PAC-Bayesian bound on $\mathbb{E}_\rho[L(h) - L(h^*)]$ and a single-hypothesis bound on $L(h^*)$. The quantity $\mathbb{E}_\rho[L(h) - L(h^*)]$ is known as *excess loss*. The advantage of this approach is that when $L(h^*)$ is a good approximation of $\mathbb{E}_\rho[L(h)]$, the excess loss has lower variance than the plain loss $\mathbb{E}_\rho[L(h)]$ and, therefore, is more efficient to bound, whereas the single-hypothesis bound on $L(h^*)$ does not involve the $\text{KL}(\rho||\pi)$ term. Therefore, it is generally beneficial to use excess losses in combination with data-informed priors. However, as with the previous approach, sacrificing S_1 to learn π and h^* means that the denominator in the bounds (n in Theorem 3.1) reduces to the size of S_2 , and it does not always pay off. (We note that the excess loss is not binary and not in the $[0, 1]$ interval, and in order to exploit small variance it is actually necessary to apply a PAC-Bayes-Empirical-Bernstein-style inequality (Tolstikhin and Seldin, 2013; Mhammedi et al., 2019; Wu et al., 2021) or the PAC-Bayes-split-kl inequality (Wu and Seldin, 2022) rather than Theorem 3.1, but the point about reduced sample size still applies.)

3.2.4 Recursive PAC-Bayes (new)

We introduce the following decomposition of the loss

$$\mathbb{E}_\rho[L(h)] = \mathbb{E}_\rho[L(h) - \gamma\mathbb{E}_\pi[L(h')]] + \gamma\mathbb{E}_\pi[L(h')]. \quad (3.1)$$

As before, we decompose S into two disjoint sets $S = S_1 \cup S_2$. We make the following major observations:

- The quantity $\mathbb{E}_\pi[L(h')]$ on the right is “of the same kind” as $\mathbb{E}_\rho[L(h)]$ on the left.
- We can take an uninformed prior π_0 and apply Theorem 3.1 (or any other suitable PAC-Bayes bound) to bound $\mathbb{E}_\pi[L(h')]$. (The KL term in the bound on $\mathbb{E}_\pi[L(h')]$ will be $\text{KL}(\pi \parallel \pi_0)$.)
- We can restrict π to depend only on S_1 , but still use all the data S in calculation of the PAC-Bayes bound on $\mathbb{E}_\pi[L(h')]$, because π is a posterior relative to π_0 , and a posterior is allowed to depend on all the data, and in particular on any subset of the data. Therefore, the empirical loss $\mathbb{E}_\pi[\hat{L}(h', S)]$ can be computed on all the data S , and the denominator of the bound in Theorem 3.1 can be the size of S , and not the size of S_2 . This is what we call *preservation of confidence information on π* , because all the data S are used to construct a confidence bound on $\mathbb{E}_\pi[L(h')]$, and not just S_2 . This is in contrast to the bound on $L(h^*)$ in the approach of Mhammedi et al. (2019), which only allows to use S_2 for bounding $L(h^*)$. Note that while we use all the data S in calculation of the bound, we only use S_1 and $\mathbb{E}_\pi[\hat{L}(h', S_1)]$ in the construction of π . Nevertheless, we can still use the knowledge that we will have n samples when we reach the estimation phase, i.e., when constructing π we can leave the denominator of the bound at n , allowing more aggressive deviation from π_0 .
- If we restrict π to depend only on S_1 , then it is a valid prior for estimation of any posterior quantity $\mathbb{E}_\rho[\cdot]$ based on S_2 . Thus, if we also restrict γ to depend only on S_1 , we can use any PAC-Bayes-Empirical-Bernstein-style inequality or the PAC-Bayes-split-kl inequality to estimate the excess loss $\mathbb{E}_\rho[L(h) - \gamma \mathbb{E}_\pi[L(h')]]$ based on S_2 , i.e., based on $\mathbb{E}_\rho[\hat{L}(h, S_2) - \gamma \mathbb{E}_\pi[\hat{L}(h', S_2)]]$. If $\gamma \mathbb{E}_\pi[L(h')]$ is a good approximation of $\mathbb{E}_\rho[L(h)]$ and $\mathbb{E}_\rho[L(h)]$ is not close to zero, then the excess loss $\mathbb{E}_\rho[L(h) - \gamma \mathbb{E}_\pi[L(h')]]$ is more efficient to bound than the plain loss $\mathbb{E}_\rho[L(h)]$.
- In general, since $\mathbb{E}_\rho[L(h)]$ is expected to improve on $\mathbb{E}_\pi[L(h')]$, it is natural to set $\gamma < 1$. However, γ is not allowed to depend on S_2 ,

because otherwise $\hat{L}(h, S_2) - \gamma \mathbb{E}_\pi[\hat{L}(h', S_2)]$ becomes a biased estimate of $L(h) - \gamma \mathbb{E}_\pi[L(h')]$. We discuss the choice of γ in more detail when we present the bound and the experiments.

- Biggs and Guedj (2023) have proposed a sequential martingale-style evaluation of a martingale version of $\mathbb{E}_\rho[L(h) - L(h^*)]$ and $L(h^*)$ in the approach of Mhammedi et al., but it has not been shown to yield significant improvements yet. The same “martingalization” can be directly applied to our decomposition, but to keep things simple we stay with the basic decomposition.
- Finally, we note that we can split S_1 further and apply Equation 3.1 recursively to bound $\mathbb{E}_\pi[L(h')]$.

To set notation for recursive decomposition, we use $\pi_0, \pi_1, \dots, \pi_T$ to denote a sequence of distributions on \mathcal{H} , where π_0 is an uninformed prior and $\pi_T = \rho$ is the final posterior. We use $\gamma_2, \dots, \gamma_T$ to denote a sequence of coefficients. For $t \geq 2$ we then have the recursive decomposition

$$\mathbb{E}_{\pi_t}[L(h)] = \mathbb{E}_{\pi_t}[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h)]] + \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h)]. \quad (3.2)$$

To construct π_1, \dots, π_T we split the data S into T non-overlapping subsets, $S = S_1 \cup \dots \cup S_T$. We restrict π_t to depend on $U_t^{\text{train}} = \bigcup_{s=1}^t S_s$ only, and we use $U_t^{\text{val}} = \bigcup_{s=t}^T S_s$ to estimate (recursively) $\mathbb{E}_{\pi_t}[L(h)]$ (see Figure 3.2 in Appendix 3.7.1). Note that S_t is used both for construction of π_t and for estimation of $\mathbb{E}_{\pi_t}[L(h)]$ (it is both in U_t^{train} and U_t^{val}), resulting in efficient use of the data. It is possible to use any standard PAC-Bayes bound, e.g., Theorem 3.1, to bound $\mathbb{E}_{\pi_1}[L(h)]$, and any PAC-Bayes-Empirical-Bernstein-style bound or the PAC-Bayes-split-kl bound to bound the excess losses $\mathbb{E}_{\pi_t}[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h)]]$. The excess losses take more than three values, so in the next section we present a generalization of the PAC-Bayes-split-kl inequality to general discrete random variables, which may be of independent interest. The Recursive PAC-Bayes bound is presented in Section 3.4.

3.3 Split-kl and PAC-Bayes-split-kl inequalities for discrete random variables

The kl inequality is one of the tightest concentration of measure inequalities for binary random variables. Letting $\text{kl}^{-1,+}(\hat{p}, \varepsilon) := \max \{p : p \in [0, 1] \text{ and } \text{kl}(\hat{p}||p) \leq \varepsilon\}$

denote the upper inverse of kl and $\text{kl}^{-1,-}(\hat{p}, \varepsilon) := \min \{p : p \in [0, 1] \text{ and } \text{kl}(\hat{p} \| p) \leq \varepsilon\}$ the lower inverse, it states the following.

Theorem 3.2 (kl Inequality (Langford, 2005; Foong et al., 2021, 2022)). *Let Z_1, \dots, Z_n be independent random variables bounded in the $[0, 1]$ interval and with $\mathbb{E}[Z_i] = p$ for all i . Let $\hat{p} = \frac{1}{n} \sum_{i=1}^n Z_i$ be the empirical mean. Then, for any $\delta \in (0, 1)$:*

$$\mathbb{P}\left(p \geq \text{kl}^{-1,+}\left(\hat{p}, \frac{1}{n} \ln \frac{1}{\delta}\right)\right) \leq \delta \quad ; \quad \mathbb{P}\left(p \leq \text{kl}^{-1,-}\left(\hat{p}, \frac{1}{n} \ln \frac{1}{\delta}\right)\right) \leq \delta.$$

While the kl inequality is tight for binary random variables, it is loose for random variables taking more than two values due to its inability to exploit small variance. To address this shortcoming Wu and Seldin (2022) have presented the split-kl and PAC-Bayes-split-kl inequalities for ternary random variables. Ternary random variables naturally appear in a variety of applications, including analysis of excess losses, certain ways of analysing majority votes, and in learning with abstention. The bound of Wu and Seldin is based on decomposition of a ternary random variable into a pair of binary random variables and application of the kl inequality to each of them. Their decomposition yields a tight bound in the binary and ternary case, but loose otherwise. The same decomposition was used by Biggs and Guedj (2023) to derive a slight variation of the inequality, with the same limitations. We present a novel decomposition of discrete random variables into a superposition of binary random variables. Unlike the decomposition of Wu and Seldin, which only applies in the ternary case, our decomposition applies to general discrete random variables. By combining it with kl bounds for the binary elements we obtain a tight bound. The decomposition is presented formally below and illustrated graphically in Figure 3.1 in Appendix 3.7.1.

3.3.1 Split-kl inequality

Let $Z \in \{b_0, \dots, b_K\}$ be a $(K + 1)$ -valued random variable with $b_0 < b_1 < \dots < b_K$. For $j \in \{1, \dots, K\}$ define $Z_{|j} = \mathbf{1}(Z \geq b_j)$ and $\alpha_j = b_j - b_{j-1}$. Then $Z = b_0 + \sum_{j=1}^K \alpha_j Z_{|j}$. For a sequence Z_1, \dots, Z_n of $(K + 1)$ -valued random variables with the same support, let $Z_{i|j} = \mathbf{1}(Z_i \geq b_j)$ denote the elements of binary decomposition of Z_i .

Theorem 3.3 (Split-kl inequality for discrete random variables). *Let Z_1, \dots, Z_n be i.i.d. random variables taking values in $\{b_0, \dots, b_K\}$ with $\mathbb{E}[Z_i] = p$ for all i . Let $\hat{p}_{|j} = \frac{1}{n} \sum_{i=1}^n Z_{i|j}$. Then for any $\delta \in (0, 1)$:*

$$\mathbb{P}\left(p \geq b_0 + \sum_{j=1}^K \alpha_j \text{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n} \ln \frac{K}{\delta}\right)\right) \leq \delta.$$

Proof. Let $p_{|j} = \mathbb{E}[\hat{p}_{|j}]$, then $p = b_0 + \sum_{j=1}^K \alpha_j p_{|j}$ and

$$\mathbb{P}\left(p \geq b_0 + \sum_{j=1}^K \alpha_j \text{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n} \ln \frac{K}{\delta}\right)\right) \leq \mathbb{P}\left(\exists j : p_{|j} \geq \text{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n} \ln \frac{K}{\delta}\right)\right) \leq \delta,$$

where the first inequality is by the decomposition of p and the second inequality is by the union bound and Theorem 3.2. \square

3.3.2 PAC-Bayes-Split-kl inequality

Let $f : \mathcal{H} \times \mathcal{Z} \rightarrow \{b_0, \dots, b_K\}$ be a $(K+1)$ -valued loss function. (To connect it to the earlier examples, in the binary prediction case we would have $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with elements $Z = (X, Y)$ and $f(h, Z) = \ell(h(X), Y)$, but we will need a more general space \mathcal{Z} later.) For $j \in \{1, \dots, K\}$ let $f_j(\cdot, \cdot) = \mathbb{1}(f(\cdot, \cdot) \geq b_j)$. Let \mathcal{D}_Z be an unknown distribution on \mathcal{Z} . For $h \in \mathcal{H}$ let $F(h) = \mathbb{E}_{\mathcal{D}_Z}[f(h, Z)]$ and $F_j(h) = \mathbb{E}_{\mathcal{D}_Z}[f_j(h, Z)]$. Let $S = \{Z_1, \dots, Z_n\}$ be an i.i.d. sample according to \mathcal{D}_Z and $\hat{F}_j(h, S) = \frac{1}{n} \sum_{i=1}^n f_j(h, Z_i)$.

Theorem 3.4 (PAC-Bayes-Split-kl Inequality). *For any distribution π on \mathcal{H} that is independent of S and any $\delta \in (0, 1)$:*

$$\mathbb{P}\left(\exists \rho \in \mathcal{P} : \mathbb{E}_\rho[F(h)] \geq b_0 + \sum_{j=1}^K \alpha_j \text{kl}^{-1,+}\left(\mathbb{E}_\rho[\hat{F}_j(h, S)], \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{2K\sqrt{n}}{\delta}}{n}\right)\right) \leq \delta,$$

where \mathcal{P} is the set of all possible probability distributions on \mathcal{H} that can depend on S .

Proof. We have $f(\cdot, \cdot) = b_0 + \sum_{j=1}^K \alpha_j f_j(\cdot, \cdot)$ and $F(h) = b_0 + \sum_{j=1}^K \alpha_j F_j(h)$.

Therefore,

$$\begin{aligned} & \mathbb{P} \left(\exists \rho \in \mathcal{P} : \mathbb{E}_\rho[F(h)] \geq b_0 + \sum_{j=1}^K \alpha_j \text{kl}^{-1,+} \left(\mathbb{E}_\rho[\hat{F}_j(h, S)], \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{2K\sqrt{n}}{\delta}}{n} \right) \right) \\ & \leq \mathbb{P} \left(\exists \rho \in \mathcal{P} \text{ and } \exists j : \mathbb{E}_\rho[F_j(h)] \geq \text{kl}^{-1,+} \left(\mathbb{E}_\rho[\hat{F}_j(h, S)], \frac{\text{KL}(\rho \parallel \pi) + \ln \frac{2K\sqrt{n}}{\delta}}{n} \right) \right) \leq \delta, \end{aligned}$$

where the first inequality is by the decomposition of F and the second inequality is by the union bound and application of Theorem 3.1 to F_j (note that f_j is a zero-one loss function). \square

3.4 Recursive PAC-Bayes bound

Now we derive a Recursive PAC-Bayes bound based on the loss decomposition in Equation 3.2. In order to bound $\mathbb{E}_{\pi_t}[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]]$ we need empirical estimates of $L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')]$. We denote $F_{\gamma_t}(h, \pi_{t-1}) = L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}}[L(h')] = \mathbb{E}_{\mathcal{D} \times \pi_{t-1}}[\ell(h(X), Y) - \gamma_t \ell(h'(X), Y)]$, where $\mathcal{D} \times \pi_{t-1}$ is a product distribution on $\mathcal{X} \times \mathcal{Y} \times \mathcal{H}$ and $h' \in \mathcal{H}$ is sampled according to π_{t-1} . We define $f_{\gamma_t}(h, (X, Y, h')) = \ell(h(X), Y) - \gamma_t \ell(h'(X), Y) \in \{-\gamma_t, 0, 1 - \gamma_t, 1\}$, then $F_{\gamma_t}(h, \pi_{t-1}) = \mathbb{E}[f_{\gamma_t}(h, (X, Y, h'))]$. We let $\{b_{t|0}, b_{t|1}, b_{t|2}, b_{t|3}\} = \{-\gamma_t, 0, 1 - \gamma_t, 1\}$ and $f_{\gamma_t|j}(h, (X, Y, h')) = \mathbf{1}(f_{\gamma_t}(h, (X, Y, h')) \geq b_{t|j})$. We let $F_{\gamma_t|j}(h, \pi_{t-1}) = \mathbb{E}[f_{\gamma_t|j}(h, (X, Y, h'))]$, then $F_{\gamma_t}(h, \pi_{t-1}) = -\gamma_t + \sum_{j=1}^3 (b_{t|j} - b_{t|j-1}) F_{\gamma_t|j}(h, \pi_{t-1})$.

For each sample (X_i, Y_i) we sample a prediction rule h_i to serve for h' . We let $\hat{\pi}_{t-1} = \{h_{t-1,1}, h_{t-1,2}, \dots\}$ to be the resulting sequence of prediction rules sampled independently according to π_{t-1} . We only use the first elements of the sequence in accordance with the size of the sample used in the corresponding estimation. We define $\hat{F}_{\gamma_t|j}(h, U_t^{\text{val}}, \hat{\pi}_{t-1}) = \frac{1}{n_t^{\text{val}}} \sum_{(X,Y) \in U_t^{\text{val}}, h' \in \hat{\pi}_{t-1}} f_{\gamma_t|j}(h, (X, Y, h'))$, where $n_t^{\text{val}} = |U_t^{\text{val}}|$ and we take the first n_t^{val} elements from $\hat{\pi}_{t-1}$, one for each sample. Note that $\mathbb{E}[\hat{F}_{\gamma_t|j}(h, S, \hat{\pi}_{t-1})] = F_{\gamma_t|j}(h, \pi_{t-1})$. Now we are ready to state the bound.

Theorem 3.5 (Recursive PAC-Bayes Bound). *Let $S = S_1 \cup \dots \cup S_T$ be an i.i.d. sample split in an arbitrary way into T non-overlapping subsamples, and let $U_t^{\text{train}} = \bigcup_{s=1}^t S_s$ and $U_t^{\text{val}} = \bigcup_{s=t}^T S_s$. Let $n_t^{\text{val}} = |U_t^{\text{val}}|$. Let $\pi_0^*, \pi_1^*, \dots, \pi_T^*$ be a sequence of distributions on \mathcal{H} , where π_t^* is allowed to depend on U_t^{train} , but not the rest of the data. Let $\gamma_2, \dots, \gamma_T$ be a sequence of coefficients, where γ_t*

is allowed to depend on U_{t-1}^{train} , but not the rest of the data. For $t \in \{1, \dots, T\}$ let \mathcal{P}_t be a set of distributions on \mathcal{H} , which are allowed to depend on U_t^{train} . Then for any $\delta \in (0, 1)$:

$$\mathbb{P}(\exists t \in \{1, \dots, T\} \text{ and } \pi_t \in \mathcal{P}_t : \mathbb{E}_{\pi_t}[L(h)] \geq B_t(\pi_t)) \leq \delta,$$

where $B_t(\pi_t)$ is a PAC-Bayes bound on $\mathbb{E}_{\pi_t}[L(h)]$ defined recursively as follows. For $t = 1$

$$B_1(\pi_1) = \text{kl}^{-1,+} \left(\mathbb{E}_{\pi_1}[\hat{L}(h, S)], \frac{\text{KL}(\pi_1 \| \pi_0^*) + \ln \frac{2T\sqrt{n}}{\delta}}{n} \right).$$

For $t \geq 2$ we let $\mathcal{E}_t(\pi_t, \gamma_t)$ denote a PAC-Bayes bound on $\mathbb{E}_{\pi_t}[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}^*}[L(h')]]$ given by

$$\mathcal{E}_t(\pi_t, \gamma_t) = -\gamma_t + \sum_{j=1}^3 (b_{t|j} - b_{t|j-1}) \text{kl}^{-1,+} \left(\mathbb{E}_{\pi_t} \left[\hat{F}_{\gamma_t|j}(h, U_t^{\text{val}}, \hat{\pi}_{t-1}^*) \right], \frac{\text{KL}(\pi_t \| \pi_{t-1}^*) + \ln \frac{6T\sqrt{n_t^{\text{val}}}}{\delta}}{n_t^{\text{val}}} \right)$$

and then

$$B_t(\pi_t) = \mathcal{E}_t(\pi_t) + \gamma_t B_{t-1}(\pi_{t-1}^*). \quad (3.3)$$

Proof. By Theorem 3.1 we have $\mathbb{P}(\exists \pi_1 \in \mathcal{P}_1 : \mathbb{E}_{\pi_1}[L(h)] \geq B_1(\pi_1)) \leq \frac{\delta}{T}$. Further, by Theorem 3.4 for $t \in \{2, \dots, T\}$ we have

$\mathbb{P}(\exists \pi_t \in \mathcal{P}_t : \mathbb{E}_{\pi_t}[L(h) - \gamma_t \mathbb{E}_{\pi_{t-1}^*}[L(h')]] \geq \mathcal{E}_t(\pi_t, \gamma_t)) \leq \frac{\delta}{T}$. The theorem follows by a union bound and the recursive decomposition of the loss (Equation 3.2). \square

Discussion

- Note that π_1^*, \dots, π_T^* can be constructed sequentially, but π_t^* can only be constructed based on the data in U_t^{train} , meaning that in the construction of π_t^* we can only rely on $\mathbb{E}_{\pi_t} \left[\hat{F}_{\gamma_t|j}(h, S_t, \hat{\pi}_{t-1}^*) \right]$, but not on $\mathbb{E}_{\pi_t} \left[\hat{F}_{\gamma_t|j}(h, U_t^{\text{val}}, \hat{\pi}_{t-1}^*) \right]$. Also note that S_t is part of both U_t^{train} and U_t^{val} (see Figure 3.2 in Appendix 3.7.1 for a graphical illustration). In other words, when we evaluate the bounds we can use additional data. And even though the additional data can only be used in the evaluation

stage, we can still use the knowledge that we will get more data for evaluation when we construct π_t^* . For example, we can take

$$\pi_t^* = \arg \min_{\pi} \sum_{j=1}^3 (b_{t|j} - b_{t|j-1}) \text{kl}^{-1,+} \left(\mathbb{E}_{\pi} \left[\hat{F}_{\gamma_t|j}(h, S_t, \hat{\pi}_{t-1}^*) \right], \frac{\text{KL}(\pi \| \pi_{t-1}^*) + \ln \frac{6T\sqrt{n_t^{\text{val}}}}{\delta}}{n_t^{\text{val}}} \right). \quad (3.4)$$

The empirical losses above are calculated on S_t corresponding to π_t^* , but the sample sizes n_t^{val} correspond to the size of the validation set U_t^{val} rather than the size of S_t . This allows to be more aggressive in deviating with π_t^* from π_{t-1}^* by sustaining larger $\text{KL}(\pi_t^* \| \pi_{t-1}^*)$ terms.

- Similarly, $\gamma_2, \dots, \gamma_T$ can also be constructed sequentially, as long as γ_t only depends on U_{t-1}^{train} (otherwise $\hat{F}_{\gamma_t|j}(h, S_t, \hat{\pi}_{t-1}^*)$ becomes a biased estimate of $F_{\gamma_t|j}(h, \pi_{t-1}^*)$).
- We naturally want to have improvement over recursion rounds, meaning $B_t(\pi_t^*) < B_{t-1}(\pi_{t-1}^*)$. Plugging this into Equation 3.3, we obtain $\mathcal{E}(\pi_t^*, \gamma_t) + \gamma_t B_{t-1}(\pi_{t-1}^*) < B_{t-1}(\pi_{t-1}^*)$, which implies that we want γ_t to be sufficiently small to satisfy $\gamma_t < 1 - \frac{\mathcal{E}_t(\pi_t^*, \gamma_t)}{B_{t-1}(\pi_{t-1}^*)}$. At the same time, γ_t should be non-negative. Therefore, improvement over recursion steps can only be maintained as long as $\mathcal{E}_t(\pi_t^*, \gamma_t) < B_{t-1}(\pi_{t-1}^*)$. We also note that if $\mathbb{E}_{\pi_t^*}[L(h)] > 0$, then overly small γ_t can increase the excess loss. Thus, unless the loss is close to zero, γ_t should not be too small either.

3.5 Experiments

In this section, we provide an empirical comparison of our Recursive PAC-Bayes (RPB) procedure to the following prior work: i) Uninformed priors (Uninformed), (Dziugaite and Roy, 2017b); ii) Data-informed priors (Informed) (Ambroladze et al., 2007; Pérez-Ortiz et al., 2021); iii) Data-informed prior + excess loss (Informed + Excess) (Mhammedi et al., 2019; Wu and Seldin, 2022). All the experiments were run on a laptop.

We start with describing the details of the optimization procedure, and then present the results.

3.5.1 Details of the optimization and evaluation procedure

We constructed π_1^*, \dots, π_T^* sequentially using the optimization objective 3.4, and computed the bound using the recursive procedure in Theorem 3.5. There are a few technical details concerning convexity of the optimization procedure and infinite size of the set of prediction rules \mathcal{H} that we address next.

3.5.1.1 Convexification of the loss functions

The functions $f_{\gamma_t|j}(h, (X, Y, h'))$ defined in Section 3.4 are non-convex and non-differentiable: $f_{\gamma_t|j}(h, (X, Y, h')) = \mathbb{1}(f_{\gamma_t}(h, (X, Y, h')) \geq b_{t|j}) = \mathbb{1}(\ell(h(X), Y) - \gamma_t \ell(h'(X), Y) \geq b_{t|j})$. In order to facilitate optimization, we approximate the external indicator function $\mathbb{1}(z \geq z_0)$ by a sigmoid function $\omega(z; c_1, z_0) = (1 + \exp(c_1(z - z_0)))^{-1}$ with a fixed parameter $c_1 > 0$ specified in Appendix 3.7.2.3.

Furthermore, since the zero-one loss $\ell(h(X), Y)$ is also non-differentiable, we adopt the cross-entropy loss, as in most modern training procedures (Pérez-Ortiz et al., 2021). Specifically, for a k -class classification problem, let $h : \mathcal{X} \rightarrow \mathbb{R}^k$ represent the function implemented by the neural network, assigning each class a real value. Let $u = h(X)$ be the assignment, with u_i being the i -th value of the vector. To convert this real-valued vector into a probability distribution over classes, we apply the softmax function $\sigma : \mathbb{R}^k \rightarrow \Delta^{k-1}$, where $\sigma(u)_i = \exp(c_2 u_i) / \sum_j \exp(c_2 u_j)$ for some $c_2 > 0$ for each entry. The cross-entropy loss $\ell^{\text{ce}} : \mathbb{R}^k \times [k] \rightarrow \mathbb{R}$ is defined by $\ell^{\text{ce}}(u, Y) = -\log(\sigma(u)_Y)$. However, since this loss is unbounded, whereas the PAC-Bayes-kl bound requires losses within $[0, 1]$, we enforce a $[0, 1]$ -valued cross-entropy loss by first lower-bounding the probability assigned to Y by taking $\tilde{\sigma}(u)_Y = \max(\sigma(u)_Y, p_{\min})$ for $p_{\min} > 0$, and then rescaling it to $[0, 1]$ by taking $\ell^{\text{ce}}(u, Y) = -\log(\tilde{\sigma}(u)_Y) / \log(1/p_{\min})$ (Dziugaite and Roy, 2017b).

We emphasize that in the evaluation of the bound (using Theorem 3.5), we directly compute the zero-one loss and the $f_{\gamma_t|j}$ functions without employing the approximations.

3.5.1.2 Relaxation of the PAC-Bayes-kl bound

The PAC-Bayes-kl bound is often criticized for being unfriendly to optimization (Rodríguez-Gálvez et al., 2024). Therefore, several relaxations have been

proposed, including the PAC-Bayes-classic bound (McAllester, 1999), the PAC-Bayes- λ bound (Thiemann et al., 2017), and the PAC-Bayes-quadratic bound (Rivasplata et al., 2019b; Pérez-Ortiz et al., 2021), among others. In our optimization we have adopted the bound of McAllester (1999) instead of the kl-based bounds in Equation 3.4.

We again emphasize that in the evaluation of the bound we used the kl-based bounds in Theorem 3.5.

3.5.1.3 Estimation of $\mathbb{E}_\pi[\cdot]$

Due to the infinite size of \mathcal{H} and lack of a closed-form expression for $\mathbb{E}_{\pi_1}[\hat{L}(h, S)]$ and $\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t|j}(h, U_t^{\text{val}}, \hat{\pi}_{t-1}^*)]$ appearing in Theorem 3.5, we approximate them by sampling (Pérez-Ortiz et al., 2021). For optimization, we sample one classifier for each mini-batch during stochastic gradient descent. For evaluation, we sample one classifier for each data in the corresponding evaluation dataset. Due to approximation of the empirical quantities the final bound in Theorem 3.5 requires an additional concentration bound. (We note that the extra bound is only required for computation of the final bound, but not for optimization of $\hat{\pi}_t^*$.) Specifically, let $\hat{\pi}_t^* = \{h_{t,1}, h_{t,2}, \dots, h_{t,m}\}$ be m samples drawn i.i.d. from π_t^* . Then for any function $f(h)$ taking values in $[0, 1]$ (which is the case for $\hat{L}(h, S)$ and $\hat{F}_{\gamma_t|j}(h, U_t^{\text{val}}, \hat{\pi}_{t-1}^*)$) and $\delta' \in (0, 1)$ we have

$$\mathbb{P}\left(\mathbb{E}_{\pi_t^*}[f(h)] \geq \text{kl}^{-1,+}\left(\frac{1}{m} \sum_{i=1}^m f(h_{t,i}), \frac{1}{m} \log \frac{1}{\delta'}\right)\right) \leq \delta'.$$

It is worth noting that $\mathbb{E}_{\pi_t^*}[f(h)]$ is evaluated for a fixed π_t^* , meaning that there is no selection involved, and therefore no KL term appears in the bound above. We, of course, take a union bound over all the quantities being estimated.

3.5.2 Experimental results

We evaluated our approach and compared it to prior work using multi-class classification tasks on MNIST (LeCun and Cortes, 2010) and Fashion MNIST (Xiao et al., 2017) datasets, both with 60000 training data. The experimental setup was based on the work of Dziugaite and Roy (2017b) and Pérez-Ortiz et al. (2021). Similar to them we used Gaussian distributions for all the priors

and posteriors, modeled by probabilistic neural networks. Technical details are provided in Appendix 3.7.2.

The empirical evaluations on MNIST and Fashion MNIST are presented in Table 3.1 and Table 3.2 respectively. For the Uninformed approach, we trained and evaluated the bound using the entire training dataset directly. For the other two baseline methods, Informed and Informed + Excess Loss, we used half of the training data to train the informed prior and an ERM h^* for the excess loss, and the other half to learn the posterior. For our Recursive PAC-Bayes (RPB), we chose $\gamma_t = 1/2$ for all t , and conducted experiments with $T = 2$, $T = 4$, $T = 6$, and $T = 8$ to study the impact of recursion depth. (Each value of T corresponded to a separate run of the algorithm and a separate evaluation of the bound, i.e., they should not be seen as successive refinements.) We applied geometric split of the data, where at each recursion step the data were split in two equal halves. Specifically, for $T = 2$ the split was (30000,30000) points, for $T = 4$ it was (7500, 7500, 15000, 30000) points, for $T = 6$ it was (1875, 1875, 3750, 7500, 15000, 30000) points, and for $T = 8$ it was (468, 469, 938, 1875, 3750, 7500, 15000, 30000) points. The motivation was to let early recursion steps with few data points efficiently learn the prior, while keeping enough data for fine-tuning in the later steps. Note that with this approach the value of $n_t^{\text{val}} = |U_t^{\text{val}}| = \sum_{s=t}^T |S_s|$, which is in the denominator of the bounds in Theorem 3.5, is at least $\frac{n}{2}$.

Table 3.1: Comparison of the classification loss of the final posterior ρ on the entire training data, $\mathbb{E}_\rho[\hat{L}(h, S)]$ (Train 0-1), and on the testing data, $\mathbb{E}_\rho[\hat{L}(h, S^{\text{test}})]$ (Test 0-1), and the corresponding bounds for each method on MNIST. We report the mean and one standard deviation over 5 repetitions. “Unif.” abbreviates the Uniform approach, “Inf.” the Informed, “Inf. + Ex.” the Informed + Excess Loss, and “RPB” the Recursive PAC-Bayes.

	MNIST		
	Train 0-1	Test 0-1	Bound
Uninf.	.343 (2e-3)	.335 (3e-3)	.457 (2e-3)
Inf.	.377 (8e-4)	.371 (6e-3)	.408 (9e-4)
Inf. + Ex.	.157 (2e-3)	.151 (3e-3)	.192 (2e-3)
RPB $T = 2$.143 (2e-3)	.139 (3e-3)	.321 (3e-3)
RPB $T = 4$.112 (1e-3)	.109 (1e-3)	.203 (8e-4)
RPB $T = 6$.103 (1e-3)	.101 (1e-3)	.166 (1e-3)
RPB $T = 8$.101 (1e-3)	.097 (2e-3)	.158 (2e-3)

Table 3.2: Comparison of the classification loss of the final posterior ρ on the entire training data, $\mathbb{E}_\rho[\hat{L}(h, S)]$ (Train 0-1), and on the testing data, $\mathbb{E}_\rho[\hat{L}(h, S^{\text{test}})]$ (Test 0-1), and the corresponding bounds for each method on Fashion MNIST. We report the mean and one standard deviation over 5 repetitions. “Unif.” abbreviates the Uniform approach, “Inf.” the Informed, “Inf. + Ex.” the Informed + Excess Loss, and “RPB” the Recursive PAC-Bayes.

	Fashion MNIST		
	Train 0-1	Test 0-1	Bound
Uninf.	.382 (2e-3)	.384 (2e-3)	.464 (2e-3)
Inf.	.412 (1e-3)	.413 (6e-3)	.440 (1e-3)
Inf. + Ex.	.280 (4e-3)	.285 (5e-3)	.342 (6e-3)
RPB $T = 2$.257 (3e-3)	.266 (5e-3)	.404 (3e-3)
RPB $T = 4$.203 (2e-3)	.213 (3e-3)	.293 (1e-3)
RPB $T = 6$.186 (4e-4)	.198 (1e-3)	.255 (1e-3)
RPB $T = 8$.181 (1e-3)	.192 (3e-3)	.242 (1e-3)

Tables 3.1 and 3.2 show that while $T = 2$, which corresponds to the data split in the Informed and Informed + Excess Loss approaches, outperforms the Uninformed and Informed approaches on both datasets, it still underperforms compared to the Informed + Excess Loss approach. However, with deeper recursions, RPB yields dramatic improvements, offering significant advantages over prior work.

Tables 3.3 and 3.4 provide a glimpse into the training progress of RPB with $T = 8$ by showing the evolution of the key quantities along the recursive process. Similar tables for other values of T are provided in Appendix 3.7.2.4, along with training details for other methods. The tables show an impressive reduction of the KL term and significant improvement of the bound as the recursion proceeds, demonstrating the effectiveness of the approach.

Table 3.3: Insight into the training process of the Recursive PAC-Bayes for $T = 8$ on MNIST. The table shows the evolution of $\mathcal{E}_t(\pi_t^*, \gamma_t)$, $B_t(\pi_t^*)$, and other quantities as the training progresses with t . We define $\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1}) = -\gamma_t + \sum_{j=1}^3 (b_{t|j} - b_{t|j-1}) \hat{F}_{\gamma_t|j}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})$.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \parallel \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.009 (3e-4)		.612 (9e-3)	.532 (.011)
2	59532	-0.046 (4e-3)	.031 (1e-3)	.114 (2e-3)	.421 (5e-3)	.215 (7e-3)
3	59063	.040 (3e-3)	.013 (9e-4)	.125 (3e-3)	.336 (2e-3)	.146 (3e-3)
4	58125	.049 (1e-3)	.005 (3e-4)	.099 (1e-3)	.267 (7e-4)	.120 (2e-3)
5	56250	.052 (4e-4)	.002 (1e-4)	.083 (1e-3)	.217 (1e-3)	.111 (2e-3)
6	52500	.051 (1e-3)	.001 (4e-5)	.076 (1e-3)	.185 (1e-3)	.104 (2e-3)
7	45000	.050 (1e-3)	8e-4 (6e-5)	.073 (1e-3)	.166 (1e-3)	.099 (1e-3)
8	30000	.050 (1e-3)	6e-4 (4e-5)	.074 (1e-3)	.158 (2e-3)	.097 (2e-3)

Table 3.4: Insight into the training process of the Recursive PAC-Bayes for $T = 8$ on Fashion MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \parallel \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.003 (7e-5)		.733 (7e-3)	.686 (8e-3)
2	59532	-0.043 (8e-3)	.023 (6e-4)	.104 (9e-3)	.470 (8e-3)	.309 (7e-3)
3	59063	.083 (4e-3)	.008 (3e-4)	.161 (3e-3)	.396 (4e-3)	.242 (1e-3)
4	58125	.090 (3e-3)	.004 (5e-4)	.142 (4e-3)	.341 (5e-4)	.216 (5e-3)
5	56250	.093 (3e-3)	.001 (2e-4)	.126 (3e-3)	.297 (4e-3)	.204 (4e-3)
6	52500	.090 (1e-3)	6e-4 (6e-5)	.117 (1e-3)	.265 (1e-3)	.195 (3e-3)
7	45000	.090 (1e-3)	4e-4 (2e-5)	.115 (1e-3)	.248 (1e-3)	.195 (5e-4)
8	30000	.090 (1e-3)	4e-4 (1e-5)	.117 (1e-3)	.242 (1e-3)	.192 (3e-3)

3.6 Discussion

We have presented the first PAC-Bayesian bound that supports sequential prior updates and preserves confidence information on the prior. The work closes a long-standing gap between Bayesian and Frequentist learning by making sequential data processing and sequential updates of prior knowledge meaningful and beneficial in the frequentist framework, as it has always been

in the Bayesian framework. We have shown that apart from theoretical beauty the approach is beneficial in practice.

3.7 Appendix

3.7.1 Illustrations

In this appendix we provide graphical illustrations of the basic concepts presented in the paper.

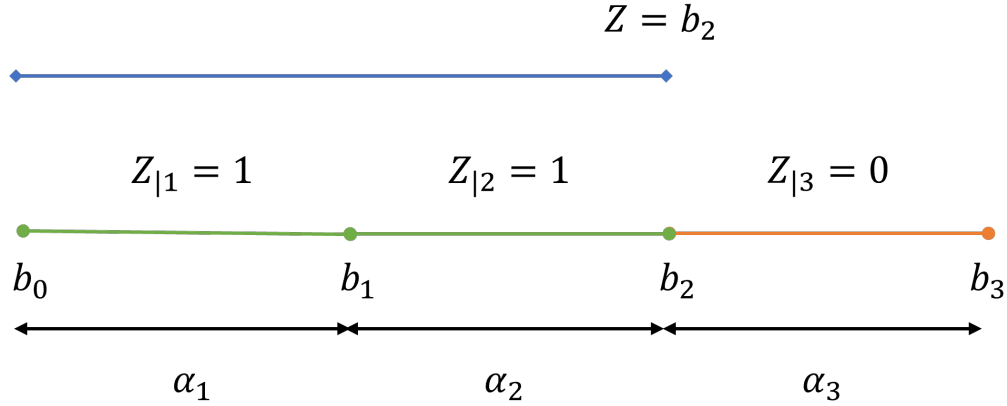


Figure 3.1: **Decomposition of a discrete random variable into a superposition of binary random variables.** The figure illustrates a decomposition of a discrete random variable Z with domain of four values $b_0 < b_1 < b_2 < b_3$ into a superposition of three binary random variables, $Z = b_0 + \sum_{j=1}^3 \alpha_j Z_{|j}$. A way to think about the decomposition is to compare it to a progress bar. In the illustration Z takes value b_2 , and so the random variables $Z_{|1}$ and $Z_{|2}$ corresponding to the first two segments “light up” (take value 1), whereas the random variable $Z_{|3}$ corresponding to the last segment remains “turned off” (takes value 0). The value of Z equals the sum of the lengths α_j of the “lighted up” segments.

Uninformed prior (McAllester, 1998)	<table><tr><td colspan="2">ρ</td></tr><tr><td colspan="2">$\mathbb{E}_{\rho}[L(h)]$</td></tr></table>	ρ		$\mathbb{E}_{\rho}[L(h)]$		S																	
ρ																							
$\mathbb{E}_{\rho}[L(h)]$																							
Informed Prior (Ambroladze et al., 2007)	<table><tr><td>S_1</td><td>S_2</td></tr><tr><td>π</td><td></td></tr><tr><td></td><td>ρ</td></tr><tr><td></td><td>$\mathbb{E}_{\rho}[L(h)]$</td></tr></table>	S_1	S_2	π			ρ		$\mathbb{E}_{\rho}[L(h)]$	S													
S_1	S_2																						
π																							
	ρ																						
	$\mathbb{E}_{\rho}[L(h)]$																						
Informed Prior + Excess Loss (Mhammedi et al., 2019)	<table><tr><td>S_1</td><td>S_2</td></tr><tr><td>$\pi ; h^*$</td><td></td></tr><tr><td></td><td>ρ</td></tr><tr><td></td><td>$\mathbb{E}_{\rho}[L(h) - L(h^*)] ; L(h^*)$</td></tr></table>	S_1	S_2	$\pi ; h^*$			ρ		$\mathbb{E}_{\rho}[L(h) - L(h^*)] ; L(h^*)$	S													
S_1	S_2																						
$\pi ; h^*$																							
	ρ																						
	$\mathbb{E}_{\rho}[L(h) - L(h^*)] ; L(h^*)$																						
Recursive PAC-Bayes $T = 2$	<table><tr><td>S_1</td><td>S_2</td></tr><tr><td>$\pi_1 ; \gamma_2$</td><td></td></tr><tr><td>$\mathbb{E}_{\pi_1}[L(h)]$</td><td></td></tr><tr><td></td><td>π_2</td></tr><tr><td></td><td>$\mathbb{E}_{\pi_2}[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h)]]$</td></tr></table>	S_1	S_2	$\pi_1 ; \gamma_2$		$\mathbb{E}_{\pi_1}[L(h)]$			π_2		$\mathbb{E}_{\pi_2}[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h)]]$	S											
S_1	S_2																						
$\pi_1 ; \gamma_2$																							
$\mathbb{E}_{\pi_1}[L(h)]$																							
	π_2																						
	$\mathbb{E}_{\pi_2}[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h)]]$																						
Recursive PAC-Bayes $T = 3$	<table><tr><td>S_1</td><td>S_2</td><td>S_3</td></tr><tr><td>$\pi_1 ; \gamma_2$</td><td></td><td></td></tr><tr><td>$\mathbb{E}_{\pi_1}[L(h)]$</td><td></td><td></td></tr><tr><td></td><td>$\pi_2 ; \gamma_3$</td><td></td></tr><tr><td></td><td>$\mathbb{E}_{\pi_2}[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h)]]$</td><td></td></tr><tr><td></td><td></td><td>π_3</td></tr><tr><td></td><td></td><td>$\mathbb{E}_{\pi_3}[L(h) - \gamma_3 \mathbb{E}_{\pi_2}[L(h)]]$</td></tr></table>	S_1	S_2	S_3	$\pi_1 ; \gamma_2$			$\mathbb{E}_{\pi_1}[L(h)]$				$\pi_2 ; \gamma_3$			$\mathbb{E}_{\pi_2}[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h)]]$				π_3			$\mathbb{E}_{\pi_3}[L(h) - \gamma_3 \mathbb{E}_{\pi_2}[L(h)]]$	S
S_1	S_2	S_3																					
$\pi_1 ; \gamma_2$																							
$\mathbb{E}_{\pi_1}[L(h)]$																							
	$\pi_2 ; \gamma_3$																						
	$\mathbb{E}_{\pi_2}[L(h) - \gamma_2 \mathbb{E}_{\pi_1}[L(h)]]$																						
		π_3																					
		$\mathbb{E}_{\pi_3}[L(h) - \gamma_3 \mathbb{E}_{\pi_2}[L(h)]]$																					

Figure 3.2: **Evolution of PAC-Bayes.** The figure shows how data are used by different methods. Dark yellow shows data used directly for optimization of the indicated quantities. Light yellow shows data involved indirectly through dependence on the prior. Light green shows data used for estimation of the indicated quantities. In Recursive PAC-Bayes data are released and used sequentially chunk-by-chunk, as indicated by the dashed lines. For example, in the $T = 3$ case $\mathbb{E}_{\pi_1}[L(h)]$ is first evaluated on S_1 to construct π_1 and γ_2 , then in the first recursion step on $S_1 \cup S_2$, and in the last step on all S .

3.7.2 Experimental details

In this section, we provide the details of the datasets in Appendix 3.7.2.1, our neural network architectures in Appendix 3.7.2.2, and other details in Appendix 3.7.2.3. We provide further statistics for all the methods on both datasets in Appendix 3.7.2.4.

3.7.2.1 Datasets

We perform our evaluation on two datasets, MNIST (LeCun and Cortes, 2010) and Fashion MNIST (Xiao et al., 2017). We will introduce these two datasets in the following.

3.7.2.1.1 MNIST

The MNIST (Modified National Institute of Standards and Technology) dataset is one of the most renowned and widely used datasets in the field of machine learning, particularly for training and testing in the domain of image processing and computer vision. It consists of a large collection of handwritten digit images, spanning the numbers 0 through 9.

The MNIST dataset comprises a total of 70,000 grayscale images of handwritten digits, where the training set has 60,000 images and the test set has 10,000 images. Each image in the dataset is 28x28 pixels, resulting in a total of 784 pixels per image. The images are in grayscale, with pixel values ranging from 0 (black) to 255 (white). Each image is associated with a label from 0 to 9, indicating the digit that the image represents. The images are typically stored in a single flattened array of 784 elements, although they can also be represented in a 28x28 matrix format.

3.7.2.1.2 Fashion MNIST

The Fashion MNIST dataset is a contemporary alternative to the traditional MNIST dataset, created to provide a more challenging benchmark for machine learning algorithms. It consists of images of various clothing items and accessories, offering a more complex and varied dataset for image classification tasks.

The Fashion MNIST dataset contains a total of 70,000 grayscale images, where the training set has 60,000 images and the test set has 10,000 images. Each image in the dataset is 28x28 pixels, resulting in a total of 784 pixels per

image. The images are in grayscale, with pixel values ranging from 0 (black) to 255 (white). Each image is associated with one of 10 categories, representing different types of fashion items. The categories are: 1. T-shirt/top 2. Trouser 3. Pullover 4. Dress 5. Coat 6. Sandal 7. Shirt 8. Sneaker 9. Bag 10. Ankle boot. Similar to MNIST, the images are stored in a single flattened array of 784 elements but can also be represented in a 28x28 matrix format.

3.7.2.2 Neural network architectures

For all methods, we adopt a family of factorized Gaussian distributions to model both priors and posteriors, characterized by the form $\pi = \mathcal{N}(w, \sigma \mathbf{I})$ where $w \in \mathbb{R}^d$ denotes the mean vector, and σ represents the scalar variance. We use feedforward neural networks for the MNIST dataset (LeCun and Cortes, 2010), while using convolutional neural networks for the Fashion MNIST dataset (Xiao et al., 2017).

Both our feedforward neural network and convolutional neural network are probabilistic, and each layer has a factorized (i.e. mean-field) Gaussian distribution.

Our feedforward neural network has the following architecture:

1. Input layer. Input size: 28×28 (flattened to 784 features).
2. Probabilistic linear layer 1. Input features: 784, output features: 600, activation: ReLU.
3. Probabilistic linear layer 2. Input features: 600, output features: 600, activation: ReLU.
4. Probabilistic linear layer 3. Input features: 600, output features: 600, activation: ReLU.
5. Probabilistic linear layer 4. Input features: 600, output features: 10, activation: Softmax.

Our convolutional neural network has the following architecture:

1. Input layer. Input size: $1 \times 28 \times 28$.
2. Probabilistic convolutional layer 1. Input channels: 1, output channels: 32, kernel size: 3x3, activation: ReLU.

3. Probabilistic convolutional layer 2. Input channels: 32, output channels: 64, kernel size: 3x3, activation: ReLU.
4. Max pooling layer. Pooling size: 2x2.
5. Flattening layer. Flattens the output from the previous layers into a single vector.
6. Probabilistic linear layer 1. Input features: 9216, output features: 128, activation: ReLU.
7. Probabilistic linear layer 2 (output layer). Input features: 128, output features: 10, activation: Softmax.

3.7.2.3 Other details in the experiments

3.7.2.3.1 General for all methods

The methods in comparisons are trained and evaluated using the procedure described in Section 3.2 and visually illustrated in Figure 3.2. We will provide some further details for each method later in the following. For all methods in comparison, we apply the optimization and evaluation method described in Section 3.5.1. For the approximation described in Section 3.5.1.1, we set the parameters $c_1 = c_2 = 5$. The lower bound for the prediction $p_{\min} = 1e - 5$. The δ in our bound and all the other methods is selected to be $\delta = 0.025$. As mentioned in Section 3.5.1.2, we use the PAC-Bayes-classic bound by McAllester in replacement of PAC-Bayes-kl when doing optimization. Note that for all methods, we also have to estimate the empirical loss of the posterior $\mathbb{E}_{\pi}[\cdot]$ described in Section 3.5.1.3. We also allocate the budget for the union bound for the estimation such that these estimations in the bound are controlled with probability at least $1 - \delta'$, where we chose $\delta' = 0.01$. Therefore, the ultimate bounds for all methods hold with probability at least $1 - \delta - \delta'$. Note that we do not consider such bounds during optimization but only when estimating the bounds.

For all methods, we adopt a family of factorized Gaussian distributions to model both priors and posteriors of all the learnable parameters of the classifiers, characterized by the form $\pi = \mathcal{N}(w, \sigma \mathbf{I})$ where $w \in \mathbb{R}^d$ denotes the mean vector, and σ represents the scalar variance. For all methods, we initialize an uninformed prior $\pi_0 = \mathcal{N}(w_0, \sigma_0 \mathbf{I})$ that is independent of data,

where the mean is randomly initialized, and the variance σ_0 is initialized to 0.03 (Pérez-Ortiz et al., 2021).

In the training process of all methods in our experiments, we set the batch size to 250, the number of training epochs to 200, and use stochastic gradient descent with a learning rate of 0.001 and a momentum of 0.95.

3.7.2.3.2 Uninformed priors

We take π_0 defined above as the uninformed prior. We then learn the posterior ρ from the prior using the entire training dataset S , applying a PAC-Bayes bound. We evaluate the bound using, again, the entire training dataset S .

3.7.2.3.3 Data-informed priors

We start with the same π_0 as the uninformed prior. We train the informed prior π_1 using S_1 with $|S_1| = |S|/2$ by minimizing a PAC-Bayes bound. The posterior ρ is then learned using the informed prior π_1 and the subset S_2 with $|S_2| = |S|/2$, again by minimizing a PAC-Bayes bound. The bound is evaluated using S_2 .

3.7.2.3.4 Data-informed priors + excess loss

We train the informed prior π_1 and the reference classifier h^* using S_1 that contains half of the training dataset. π_1 is obtained by minimizing a PAC-Bayes bound with the uninformed prior π_0 , while the reference classifier h^* is obtained by an empirical risk minimizer (ERM). The posterior ρ is obtained by minimizing a PAC-Bayes bound on the excess loss between ρ and h^* . The prior used in the bound for both training and evaluation is the data-informed prior π_1 . Therefore, the data for both training and evaluation of ρ must be the other half of data S_2 .

3.7.2.4 Further results for the experiments

In this section, we report some more statistics for all methods.

For all methods, to calculate the classification loss of ρ on the testing data, $\mathbb{E}_\rho[\hat{L}(h, S^{\text{test}})]$ (Test 0-1), we sample one classifier for each data. The train 0-1 loss for all methods is computed on the entire training dataset S , while the test 0-1 loss for all methods is computed on the test dataset S_{test} .

3.7.2.4.1 Recursive PAC-Bayes

We report the additional results of Recursive PAC-Bayes on MNIST with $T = 2$ in Table 3.5, $T = 4$ in Table 3.6, and $T = 6$ in Table 3.7. We report Recursive PAC-Bayes on Fashion MNIST with $T = 2$ in Table 3.8, $T = 4$ in Table 3.9, and $T = 6$ in Table 3.10.

Table 3.5: Insight into the training process of the Recursive PAC-Bayes for $T = 2$ on MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \parallel \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.024 (3e-5)		.370 (1e-3)	.254 (2e-3)
2	30000	.013 (3e-3)	.024 (1e-4)	.136 (3e-3)	.321 (3e-3)	.139 (3e-3)

Table 3.6: Insight into the training process of the Recursive PAC-Bayes for $T = 4$ on MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \parallel \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.023 (8e-5)		.374 (1e-3)	.258 (1e-3)
2	52500	-4e-4 (1e-3)	.025 (3e-4)	.118 (1e-3)	.305 (1e-3)	.126 (2e-3)
3	45000	.053 (1e-3)	.002 (9e-5)	.087 (2e-3)	.240 (2e-3)	.114 (1e-3)
4	30000	.054 (1e-3)	.001 (2e-5)	.083 (1e-3)	.203 (8e-4)	.109 (1e-3)

Table 3.7: Insight into the training process of the Recursive PAC-Bayes for $T = 6$ on MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \parallel \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.019 (7e-5)		.425 (1e-3)	.311 (3e-3)
2	58125	-0.013 (1e-3)	.032 (6e-4)	.128 (2e-3)	.341 (3e-3)	.139 (1e-3)
3	56250	.050 (1e-3)	.003 (1e-4)	.093 (6e-4)	.264 (1e-3)	.117 (2e-3)
4	52500	.051 (1e-3)	.001 (6e-5)	.080 (9e-4)	.212 (5e-4)	.108 (2e-3)
5	45000	.051 (1e-3)	9e-4 (3e-5)	.076 (2e-3)	.182 (1e-3)	.104 (6e-4)
6	30000	.049 (1e-3)	7e-4 (3e-5)	.074 (1e-3)	.166 (1e-3)	.101 (1e-3)

Table 3.8: Insight into the training process of the Recursive PAC-Bayes for $T = 2$ on Fashion MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.011 (3e-5)		.466 (1e-3)	.389 (5e-3)
2	30000	.064 (3e-3)	.013 (2e-4)	.171 (4e-3)	.404 (3e-3)	.266 (5e-3)

Table 3.9: Insight into the training process of the Recursive PAC-Bayes for $T = 4$ on Fashion MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		.011 (1e-4)		.476 (4e-3)	.397 (6e-3)
2	52500	.032 (6e-4)	.017 (9e-4)	.147 (2e-3)	.386 (3e-3)	.240 (4e-3)
3	45000	.100 (3e-3)	3e-3 (1e-4)	.138 (5e-3)	.331 (4e-3)	.222 (5e-3)
4	30000	.095 (2e-3)	7e-4 (5e-5)	.128 (2e-3)	.293 (1e-3)	.213 (3e-3)

Table 3.10: Insight into the training process of the Recursive PAC-Bayes for $T = 6$ on Fashion MNIST.

t	n_t^{val}	$\mathbb{E}_{\pi_t}[\hat{F}_{\gamma_t}(h, U_t^{\text{val}}, \hat{\pi}_{t-1})]$	$\frac{\text{KL}(\pi_t^* \pi_{t-1}^*)}{n_t^{\text{val}}}$	$\mathcal{E}_t(\pi_t^*, \gamma_t)$	$B_t(\pi_t^*)$	Test 0-1
1	60000		9e-3 (8e-5)		.534 (4e-3)	.462 (5e-3)
2	58125	.013 (6e-3)	.023 (1e-3)	.151 (4e-3)	.418 (5e-3)	.254 (6e-3)
3	56250	.091 (3e-3)	3e-3 (4e-4)	.141 (2e-3)	.350 (3e-3)	.223 (1e-3)
4	52500	.090 (2e-3)	9e-4 (9e-5)	.121 (1e-3)	.296 (1e-3)	.207 (1e-3)
5	45000	.090 (1e-3)	6e-4 (3e-5)	.117 (2e-3)	.265 (2e-3)	.199 (2e-3)
6	30000	.093 (1e-3)	5e-4 (2e-5)	.122 (1e-3)	.255 (1e-3)	.198 (1e-3)

3.7.2.4.2 Uninformed priors

We report the additional results of uninformed priors (McAllester, 1998) on MNIST and Fashion MNIST in Table 3.11. As described earlier in Section 3.2, Section 3.5, and Section 3.7.2.3, we evaluate the bound using the entire training set.

Table 3.11: Further details to compute the bound for the uninformed prior approach on MNIST and Fashion MNIST.

	$\mathbb{E}_\rho[\hat{L}(h, S)]$	$\frac{\text{KL}(\rho \parallel \pi_0)}{n}$	Bound	Test 0-1
MNIST	.343 (2e-3)	.023 (4e-5)	.457 (2e-3)	.335 (3e-3)
F-MNIST	.382 (2e-3)	.011 (8e-6)	.464 (2e-3)	.384 (5e-3)

3.7.2.4.3 Data-informed priors

We report the additional results of data-informed priors (Ambroladze et al., 2007) on MNIST and Fashion MNIST in Table 3.12. As described earlier in Section 3.2, Section 3.5, and Section 3.7.2.3, we evaluate the bound using S_2 that is independent of the data-informed prior π_1 .

Table 3.12: Further details to compute the bound for the data-informed prior on MNIST and Fashion MNIST.

	$\mathbb{E}_\rho[\hat{L}(h, S_2)]$	$\frac{\text{KL}(\rho \parallel \pi_0)}{ S_2 }$	Bound	Test 0-1
MNIST	.376 (8e-4)	8e-4 (9e-6)	.408 (9e-4)	.371 (6e-3)
F-MNIST	.412 (1e-3)	4e-4 (7e-6)	.440 (1e-3)	.413 (6e-3)

3.7.2.4.4 Data-informed priors + excess loss

We report the additional results of data-informed priors + excess loss (Mhammedi et al., 2019) on MNIST and Fashion MNIST in Table 3.13 and 3.14. As described earlier in Section 3.2, Section 3.5, and Section 3.7.2.3, we evaluate the bound using S_2 that is independent of the data-informed prior π_1 and the reference prediction rule h^* . The bound is composed of two parts: a bound on the excess loss of ρ with respect to h^* (Excess bound) and a single hypothesis bound on h^* (h^* bound). We report the two components of the bound in Table 3.13. We provide further details to compute these bounds from the losses of their corresponding quantities in Table 3.14.

Table 3.13: Details to compute the bound for the data-informed prior and excess loss on MNIST and Fashion MNIST. The table shows the bound on the excess loss of ρ with respect to h^* (Excess bound) and a single hypothesis bound on h^* (h^* bound).

	Ex. Bound	h^* Bound	Bound	Test 0-1
MNIST	.162 (1e-3)	.029 (4e-4)	.192 (2e-3)	.151 (3e-3)
F-MNIST	.196 (5e-3)	.145 (1e-3)	.342 (6e-3)	.285 (5e-3)

Table 3.14: Further details to compute the bound for the data-informed prior and excess loss on MNIST and Fashion MNIST. The table shows the empirical excess loss $\mathbb{E}_\rho[\hat{\Delta}(h, h^*, S_2)]$, where we define $\Delta(h, h^*, S_2) = \hat{L}(h, S_2) - \hat{L}(h^*, S_2)$, and its bound (Excess Bound). It also shows the empirical loss of the reference prediction rule $\hat{L}(h^*, S_2)$ and its bound. The computation of such bound does not involve the KL term.

	$\mathbb{E}_\rho[\hat{\Delta}(h, h^*, S_2)]$	$\frac{\text{KL}(\rho \parallel \pi_1)}{ S_2 }$	Ex. Bound	$\hat{L}(h^*, S_2)$	h^* Bound	Bound
MNIST	-0.011 (3e-3)	.035 (5e-4)	.162 (1e-3)	.026 (4e-4)	.029 (4e-4)	.192 (2e-3)
F-MNIST	.104 (6e-3)	.018 (5e-4)	.196 (5e-3)	.112 (1e-3)	.145 (1e-3)	.342 (6e-3)

Chapter 4

Another Perspective on Recursive PAC-Bayes: Connections to Cold Posteriors and KL-Annealing

4.1 Introduction

The field of machine learning has experienced tremendous advancements in recent years, with Bayesian inference playing a pivotal role in the development of robust and reliable models. In particular, the PAC-Bayesian framework has been instrumental in providing generalization guarantees for learning algorithms by incorporating prior knowledge into the learning process (Shawe-Taylor and Williamson, 1997; McAllester, 1998). Despite its theoretical strengths, traditional PAC-Bayesian approaches face significant challenges when it comes to sequential prior updates, often losing valuable confidence information in the process (McAllester, 1998; Ambroladze et al., 2007; Mhammedi et al., 2019). This limitation has spurred the development of novel methodologies aimed at preserving the information of priors as they are updated with new data (Chugg et al., 2023; Biggs and Guedj, 2023; Rodríguez-Gálvez et al., 2024).

In this context, Recursive PAC-Bayes (RPB), introduced in Chapter 3, emerges as a powerful technique that addresses the shortcomings of traditional PAC-Bayesian methods. By employing a novel decomposition of expected loss and recursive bounding techniques, Recursive PAC-Bayes enables sequential prior updates without losing confidence information. This innovative approach not only preserves the integrity of prior knowledge but also continually enhances the learning process.

Parallel to these advancements, cold posteriors (Wenzel et al., 2020) and KL-annealing (Bowman et al., 2016; Bingham et al., 2019; Zhang et al., 2019a; Sønderby et al., 2016), two commonly used practical techniques, have shown empirical success in the application of Bayesian methods. Cold posteriors (Wenzel et al., 2020), discussed thoroughly in Chapter 2 and characterized by the use of a temperature scaling factor of less than one ($T < 1$), have demonstrated significant improvements in the performance of Bayesian neural networks. This counterintuitive adjustment challenges traditional Bayesian paradigms, offering new insights into optimizing posterior distributions.

KL-annealing (Bowman et al., 2016; Bingham et al., 2019; Zhang et al., 2019a; Sønderby et al., 2016), another practical technique, gradually increases the weight of the Kullback-Leibler (KL) divergence term in the loss function during training. This method strikes a crucial balance between fitting the empirical data and maintaining a close distance between the prior and posterior distributions, thus enhancing the stability and performance of variational

inference.

In this chapter, we aim to explore the connections between these methodologies and Recursive PAC-Bayes. By delving into the updating rules of Recursive PAC-Bayes, we will uncover how its recursive updates can be viewed through the lens of cold posteriors and KL-annealing. Through this exploration, we seek to provide another perspective on the success of Recursive PAC-Bayes.

- In Section 4.2, we elucidate the connection between cold posteriors and Recursive PAC-Bayes.
- In Section 4.3, we provide a brief introduction to KL-annealing and demonstrate its connection to Recursive PAC-Bayes.

4.2 Recursive PAC-Bayes and cold posteriors

To see the connection between Recursive PAC-Bayes and cold posteriors (Wenzel et al., 2020) (if needed, please refer to Chapter 2 for introduction or background knowledge), we first present the form of cold posteriors (Equation 2.2):

$$p_\lambda(\boldsymbol{\theta}|D) \propto p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})^\lambda p(\boldsymbol{\theta}), \quad \lambda > 1 \quad \text{where } \lambda = \frac{1}{T}.$$

Such cold posteriors can be obtained by optimizing the (generalized) ELBO objective (Equation 2.3) (Alquier et al., 2016; Higgins et al., 2017). To make it clearer, we rewrite Equation 2.3 in the form of average loss:

$$\begin{aligned} p_\lambda(\boldsymbol{\theta}|D) &= \arg \min_{\rho} \mathbb{E}_{\rho}[-\ln p(D|\boldsymbol{\theta})] + \frac{1}{\lambda} \text{KL}(\rho(\boldsymbol{\theta}|D), p(\boldsymbol{\theta})) \\ &= \arg \min_{\rho} \mathbb{E}_{\rho}[-\ln \prod_{(\mathbf{y}, \mathbf{x}) \in D} p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] + \frac{1}{\lambda} \text{KL}(\rho(\boldsymbol{\theta}|D), p(\boldsymbol{\theta})) \\ &= \arg \min_{\rho} \mathbb{E}_{\rho}[-\sum_{(\mathbf{y}, \mathbf{x}) \in D} \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] + \frac{1}{\lambda} \text{KL}(\rho(\boldsymbol{\theta}|D), p(\boldsymbol{\theta})) \\ &= \arg \min_{\rho} \mathbb{E}_{\rho}[-\frac{1}{n} \sum_{(\mathbf{y}, \mathbf{x}) \in D} \ln p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] + \underbrace{\frac{1}{\lambda}}_{\text{temperature } T} \underbrace{\frac{\text{KL}(\rho(\boldsymbol{\theta}|D), p(\boldsymbol{\theta}))}{n}}_{\text{average KL}}, \end{aligned}$$

where n is the size of the dataset D used for constructing the posterior. It is now clear that obtaining cold posteriors ($T < 1$, or $\lambda > 1$ equivalently) corresponds to searching a posterior that optimizes a (generalized) ELBO objective where the coefficient of the average KL term is lesser than 1.

On the other hand, in Recursive PAC-Bayes, we obtain the sequential posteriors using Equation 3.4:

$$\begin{aligned}
 \pi_t^* &= \arg \min_{\pi} \sum_{j=1}^3 (b_{t|j} - b_{t|j-1}) \\
 &\quad \cdot \text{kl}^{-1,+} \left(\mathbb{E}_{\pi} \left[\hat{F}_{\gamma_t|j}(h, S_t, \hat{\pi}_{t-1}^*) \right], \frac{\text{KL}(\pi \| \pi_{t-1}^*) + \ln \frac{6T\sqrt{n_t^{\text{val}}}}{\delta}}{n_t^{\text{val}}} \right) \\
 &= \arg \min_{\pi} \sum_{j=1}^3 (b_{t|j} - b_{t|j-1}) \\
 &\quad \cdot \text{kl}^{-1,+} \left(\mathbb{E}_{\pi} \left[\hat{F}_{\gamma_t|j}(h, S_t, \hat{\pi}_{t-1}^*) \right], \underbrace{\frac{|S_t|}{n_t^{\text{val}}}}_{\text{temperature } T} \underbrace{\frac{\text{KL}(\pi \| \pi_{t-1}^*) + \ln \frac{6T\sqrt{n_t^{\text{val}}}}{\delta}}{|S_t|}}_{\text{average KL}} \right).
 \end{aligned}$$

As shown in Figure 3.2, for every step t until the final step T , the denominator of the coefficient of the average KL term, $n_t^{\text{val}} = |U_t^{\text{val}}| = \sum_{s=t}^T |S_s|$, is always larger than the size of the current dataset $|S_t|$ used for the construction of the posteriors, thus resulting in a cold temperature $T = \frac{|S_t|}{n_t^{\text{val}}} < 1$.

In this sense, Recursive PAC-Bayes can be seen as implicitly utilizing cold posteriors for all sequential steps t except for the last step T . As Bayesian neural networks ($\lambda = 1$) are shown to have an underfitting phenomenon with curated image datasets (please refer to Chapter 2), such a procedure allows Recursive PAC-Bayes to make more aggressive updates for fitting the data instead of staying close to the uninformed priors, especially in the early stages. Thus, it mitigates the underfitting problem and achieves better performance.

However, while this connection helps explain the success of Recursive PAC-Bayes, it is important to note the differences. Recursive PAC-Bayes is only “partially cold” compared to cold posteriors, and in each update step on the corresponding chunk of data S_t , the implicit temperature is different, increasing gradually from the lowest temperature ($T \ll 1$, or $\lambda \gg 1$ equivalently) to normal ($T = 1$, or $\lambda = 1$). More importantly, the implicit

temperature in Recursive PAC-Bayes, $\frac{|S_t|}{n_t^{\text{val}}}$, is determined by the data splitting strategy rather than as a pre-defined hyperparameter in cold posteriors. Interestingly, obtained from rigorous theoretical bounds, such a method might provide insights for choosing appropriate temperatures for Bayesian predictive models.

4.3 Recursive PAC-Bayes and KL-annealing

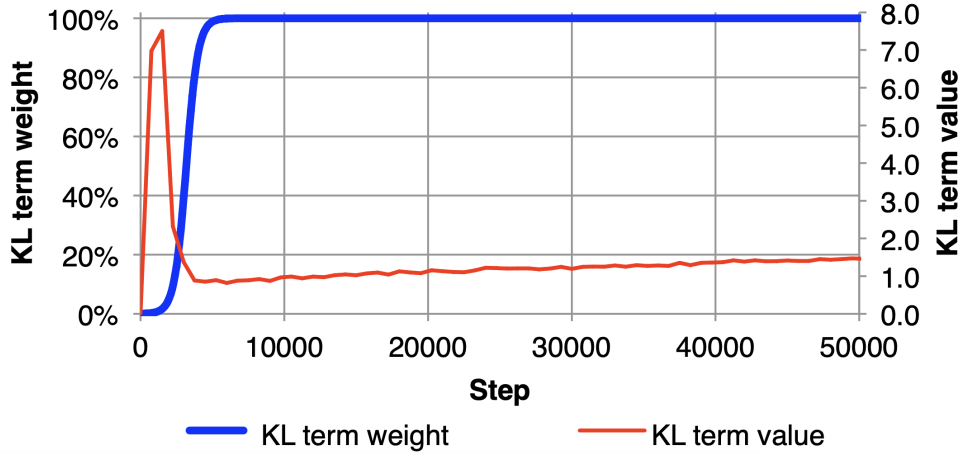


Figure 4.1: **Figure 2 in Bowman et al. (2016): a graphical illustration of KL-Annealing.**

In the realm of Bayesian inference and variational methods, KL-annealing (Bowman et al., 2016; Bingham et al., 2019; Zhang et al., 2019a; Sønderby et al., 2016) has emerged as a powerful technique to improve the training of probabilistic models, particularly variational autoencoders (VAEs) and Bayesian neural networks. KL-annealing addresses the challenge of balancing the trade-off between data fidelity and regularization during the optimization process, a crucial aspect in ensuring that the model learns a meaningful representation of the data while avoiding overfitting.

KL-annealing addresses this challenge by gradually increasing the weight of the KL divergence term in the ELBO during training, as demonstrated graphically in Figure 4.1. This approach allows the model to initially focus on fitting the data, gradually introducing the regularization effect of the KL term

to ensure the optimization of the posterior distribution remains well-behaved. The modified ELBO with annealed KL divergence is given by:

$$\text{ELBO}_\beta = \mathbb{E}_{q(\boldsymbol{\theta})} [\log p(D \mid \boldsymbol{\theta})] - \beta(t) \cdot \text{KL}(q(\boldsymbol{\theta}), p(\boldsymbol{\theta})),$$

where $\beta(t)$ is an annealing parameter that starts from a small value (often close to zero) and is gradually increased to one during the training process.

KL-annealing offers several benefits that enhance the training and performance of probabilistic models. By gradually increasing the weight of the KL divergence term in the ELBO during training, it allows the model to initially focus on fitting the data accurately, leading to improved convergence rates and more precise representations. This technique helps prevent mode collapse, ensuring that the model captures diverse features of the data. Additionally, KL-annealing stabilizes the training dynamics by avoiding abrupt changes in the optimization landscape, resulting in a balanced trade-off between data fidelity and regularization. This balance improves the model’s ability to generalize well to new, unseen data, making it a valuable approach in variational inference and Bayesian neural networks.

As shown in Section 4.2, the connection between Recursive PAC-Bayes and KL-annealing becomes evident: both methods gradually increase the KL coefficient, which converges to one at the final stages of optimization. Although Recursive PAC-Bayes may benefit from KL-annealing, there is a significant difference. Recursive PAC-Bayes employs different KL coefficients (i.e., temperatures) for each chunk of data, with optimization performed until convergence. In contrast, KL-annealing trains on all data at once, using smaller-than-one KL coefficients (cold temperatures) for more efficient optimization in the early stages, but ultimately converges with a KL coefficient equal to one.

4.4 Conclusion

In this chapter, we explored the Recursive PAC-Bayes (RPB) framework and its connections to cold posteriors and KL-annealing. Recursive PAC-Bayes addresses traditional PAC-Bayesian challenges by enabling sequential prior updates without losing confidence information, thereby enhancing the learning process with each iteration.

The concept of cold posteriors, which involves a temperature scaling factor less than one, aligns with Recursive PAC-Bayes’s approach of using “cold”

coefficients in the KL term during optimization (except for the final update). This adjustment allows Recursive PAC-Bayes to make more aggressive updates early in the learning process, helping to mitigate underfitting and improve performance.

Similarly, KL-annealing gradually increases the weight of the KL divergence term during training, balancing data fidelity and regularization dynamically for better optimization. Recursive PAC-Bayes's recursive updates inherently incorporate this principle, leading to stable training dynamics and better generalization.

By considering the perspectives of cold posteriors and KL-annealing, we showed that Recursive PAC-Bayes implicitly leverages their strengths. This connection provides a deeper understanding of the success of Recursive PAC-Bayes and highlights its potential for advancing machine learning models.

Chapter 5

(Generalized) ELBO Revisited: Another Decomposition for Mean-Field Variational Global Latent Variable Models

5.1 Introduction

Variational inference (Blei et al., 2017; Jordan et al., 1999; Wainwright and Jordan, 2008) has become a cornerstone technique for approximate Bayesian inference in complex models. The Evidence Lower Bound (ELBO) is a crucial objective in this framework, facilitating the optimization of variational distributions to approximate true posteriors. While Hoffman and Johnson (2016) have provided insightful decompositions of the ELBO for local latent variable models like Variational Autoencoders (VAEs) (Kingma and Welling, 2014), there remains a need for a deeper understanding of ELBO dynamics in global latent variable models.

In this chapter, we introduce a novel decomposition for variational global latent variable models, where the latent variables (usually the model or network weights) are shared across all data points, contrasting with the local latent variables (usually latent representation of each data point) in models like VAEs. More specifically, our decomposition holds for any mean-field variational distribution, which is the common design choice in practice.

Our proposed decomposition aims to shed light on several aspects of the training and evaluation of variational global latent variable models:

- Understanding training dynamics: by dissecting the ELBO into interpretable components, we can gain insights into how different parts of the objective function influence the training process, potentially leading to more effective optimization strategies.
- Evaluating the variational posteriors: the decomposition provides a finer-grained analysis of the variational posterior quality, enabling practitioners to diagnose and address issues in the approximation process more effectively.
- Finer control of the optimization process: with a clearer understanding of the contributions of different terms in the ELBO, we can develop more nuanced training controls, e.g., inserting separate coefficients for each term, potentially improving convergence and performance.

5.2 Rewriting the (generalized) ELBO for mean-field variational global latent variable models

In (generalized) variational inference, we consider approximating the (tempered) Bayesian posterior $p(\boldsymbol{\theta}|D)^\lambda$ with a variational distribution $q(\boldsymbol{\theta})$. Such variational posteriors are typically obtained by optimizing the (generalized) ELBO objective (essentially Equation 2.3) within the selected variational family:

$$q(\boldsymbol{\theta}) = \arg \min_q \mathbb{E}_q[-\ln p(D|\boldsymbol{\theta})] + \frac{1}{\lambda} \text{KL}(q(\boldsymbol{\theta}), p(\boldsymbol{\theta})),$$

where $p(\boldsymbol{\theta})$ is the prior and typically chosen to be the standard normal distribution for tractability. In practice, people often consider a factorized (i.e., mean-field) variational distribution, $q(\boldsymbol{\theta}) = \prod_{d=1}^D q_d(\theta_d)$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the parameters of the predictive model, is a D -dimension vector and d indexes the dimension.

Inspired by Hoffman and Johnson (2016), we present a similar but novel decomposition of the KL term in ELBO. However, while Hoffman and Johnson (2016)’s decomposition is for variational local latent variables (e.g. VAEs), our decomposition applies to variational global latent variable models (e.g. Bayesian neural networks). To begin our decomposition, we first define the followings:

$$\begin{aligned} q(d, \theta) &= q(d)q(\theta|d), & q(\theta|d) &= q_d(\theta_d), & q(d) &= \frac{1}{D}, & q^{\text{avg}}(\theta) &= \frac{1}{D} \sum_{d=1}^D q_d(\theta_d) \\ p(d, \theta) &= p(d)p(\theta|d), & p(\theta|d) &= p(\theta), & p(d) &= \frac{1}{D}, \end{aligned}$$

where the index d is treated as a random variable and $q^{\text{avg}}(\theta) = \sum_{d=1}^D q(d, \theta) = \frac{1}{D} \sum_{d=1}^D q_d(\theta_d)$ is the average (or “mixture”) variational posterior over the dimensions. Then, the (generalized) ELBO can be rewritten as:

$$\text{ELBO}_\lambda = \mathbb{E}_q[-\ln p(D|\boldsymbol{\theta})] + \frac{1}{\lambda} \left(D * (\text{KL}(q^{\text{avg}}(\theta), p(\theta)) + \mathbb{I}_{q(d, \theta)}[d, \theta]) \right), \quad (5.1)$$

where $\mathbb{I}_{q(d,\theta)}[d, \theta]$ is the mutual information between the index d and the parameter θ of the d -th dimension.

Next, we demonstrate the decomposition of the KL term in Equation 5.1:

$$\begin{aligned}
& \text{KL}(q(\boldsymbol{\theta}), p(\boldsymbol{\theta})) \\
&= \sum_{d=1}^D \text{KL}(q_d(\theta_d), p(\theta)) \\
&= \sum_{d=1}^D \int q_d(\theta_d) \log \frac{q_d(\theta_d)}{p(\theta)} d\theta \quad \text{definition of KL} \\
&= \sum_{d=1}^D \int q(\theta|d) \log \frac{q(\theta|d)}{p(\theta|d)} d\theta \quad \text{substitute the definitions accordingly} \\
&= D \sum_{d=1}^D \int \frac{1}{D} q(\theta|d) \log \frac{q(\theta|d) \frac{1}{D}}{p(\theta|d) \frac{1}{D}} d\theta \quad \text{add } D \text{ and } \frac{1}{D} \\
&= D \sum_{d=1}^D \int q(d) q(\theta|d) \log \frac{q(\theta|d) q(d)}{p(\theta|d) p(d)} d\theta \quad \text{substitute the definitions accordingly} \\
&= D \sum_{d=1}^D \int q(d, \theta) \log \frac{q(d, \theta)}{p(d, \theta)} d\theta \quad \text{product rule} \\
&= D \sum_{d=1}^D \int q(d, \theta) \log \frac{q^{\text{avg}}(\theta) q(d|\theta)}{p(\theta) p(d|\theta)} d\theta \quad \text{product rule} \\
&= D \sum_{d=1}^D \int q(d, \theta) \left(\log \frac{q^{\text{avg}}(\theta)}{p(\theta)} + \log \frac{q(d|\theta)}{p(d)} \right) d\theta \quad \text{since } p(d|\theta) = p(d) \\
&= D \left(\sum_{d=1}^D \int q(d, \theta) \log \frac{q^{\text{avg}}(\theta)}{p(\theta)} d\theta + \sum_{d=1}^D \int q(d, \theta) \log \frac{q(d|\theta)}{p(d)} d\theta \right) \quad \text{combine the log} \\
&= D \left(\int \sum_{d=1}^D q(d, \theta) \log \frac{q^{\text{avg}}(\theta)}{p(\theta)} d\theta + \int \sum_{d=1}^D q(d, \theta) \log \frac{q(d|\theta)}{p(d)} d\theta \right) \quad \text{Fubini's theorem} \\
&= D \left(\int q^{\text{avg}}(\theta) \log \frac{q^{\text{avg}}(\theta)}{p(\theta)} d\theta + \int \sum_{d=1}^D q(d, \theta) \log \frac{q(d, \theta)}{q(d) q^{\text{avg}}(\theta)} d\theta \right) \quad d \text{ is integrated out} \\
&= D (\text{KL}(q^{\text{avg}}(\theta), p(\theta)) + \mathbb{I}_{q(d,\theta)}[d, \theta]) \quad \text{substitute the definitions accordingly}
\end{aligned}$$

Intuitively, the mutual information term $\mathbb{I}_{q(d,\theta)}[d, \theta]$ could be able to describe the learning process. In the beginning, $\mathbb{I}_{q(d,\theta)}[d, \theta]$ should be 0 as we start from a factorized prior, where the dimension d is independent of the parameter θ . As learning progresses, $\mathbb{I}_{q(d,\theta)}[d, \theta]$ should increase gradually and stay at a stable level. On the other hand, the average variational posterior, $q^{\text{avg}}(\theta) = \frac{1}{D} \sum_{d=1}^D q_d(\theta_d)$, could be arbitrarily close to the prior $p(\theta)$, enabling the term $\text{KL}(q^{\text{avg}}(\theta), p(\theta))$ to be arbitrarily small. Think of each dimension $q_d(\theta_d)$, even though being different and diverse in itself, could still constitute a mixture centered at 0.

Furthermore, one might consider adding extra parameters to control the optimization objective as needed (Esmaili et al., 2019), ending in the following form for (generalized) ELBO:

$$\text{ELBO}_\lambda = \mathbb{E}_q[-\ln p(D|\boldsymbol{\theta})] + D * \left(\frac{1}{\lambda_1} \text{KL}(q^{\text{avg}}(\theta), p(\theta)) + \frac{1}{\lambda_2} \mathbb{I}_{q(d,\theta)}[d, \theta] \right). \quad (5.2)$$

5.3 Conclusion

In this chapter, we have introduced a novel decomposition for the Evidence Lower Bound (ELBO) tailored to mean-field variational global latent variable models. This decomposition could provide a deeper understanding of the training dynamics and optimization challenges inherent in these models, which involve latent variables shared across all data points. By dissecting the ELBO into more interpretable components, this decomposition facilitates finer control over the optimization process and allows for a more detailed evaluation of the quality of variational posteriors. These insights might lead to the development of more effective training strategies, ultimately enhancing the performance and reliability of variational global latent variable models.

Chapter 6

Summary and Discussion

In Chapter 2, we provided a detailed analysis of the cold posterior effect (CPE) in Bayesian deep learning. Theoretically, the Bayesian posterior is considered optimal when the model is perfectly specified. However, empirical evidence shows that posteriors with a temperature $T < 1$ can outperform the standard Bayesian posterior, a phenomenon known as CPE. We investigated the underlying causes of CPE, establishing that it occurs primarily due to underfitting resulting from model misspecification. Moreover, we demonstrated that these tempered posteriors are valid Bayesian posteriors corresponding to different combinations of likelihoods and priors, thus justifying temperature adjustment as an effective strategy for addressing underfitting. This finding is significant as it not only deepens our theoretical understanding of CPE but also provides a practical method for improving Bayesian model performance through temperature fine-tuning.

In Chapter 3, we presented the development of Recursive PAC-Bayes (RPB), a novel framework for sequentially updating posteriors in a manner that preserves confidence information. Traditional PAC-Bayesian analysis faces challenges in maintaining confidence when posteriors are updated sequentially, limiting its utility in real-world applications. The introduction of RPB addresses this issue by decomposing the expected loss of randomized classifiers into an excess loss relative to a downscaled prior loss. This approach enables recursive bounding, allowing for meaningful and beneficial sequential data processing without losing crucial confidence information. We further enhanced the practical applicability of PAC-Bayesian analysis by generalizing key inequalities and developing a robust optimization procedure based on RPB. The empirical results indicated that RPB significantly outperforms existing

methods, making it a valuable tool for improving the generalization capabilities of neural networks.

In Chapter 4, we explored the Recursive PAC-Bayes (RPB) framework and its relationship with cold posteriors and KL-annealing, two empirically effective techniques in Bayesian modeling. We demonstrated that the intermediate posteriors generated by RPB’s updating rules are inherently “cold”, aligning with the principles underlying cold posteriors. Furthermore, RPB’s recursive updates is similar to the process of KL-annealing, with the potential of bringing in more stable training dynamics. These connections not only shed light on the success of RPB but also suggest intriguing avenues for future research.

In Chapter 5, we introduced a novel decomposition of the evidence lower bound (ELBO) for variational global latent variable models. The proposed decomposition breaks down the ELBO into more interpretable components, with the potential of offering a finer-grained understanding of the training dynamics and posterior quality in these models. This decomposition might improve optimization strategies by allowing for more nuanced control of the temperature parameter T , leading to better convergence and model performance.

List of Publications

The work presented in this thesis has led to this publication (Zhang et al., 2024) and this preprint (Wu et al., 2024). Additionally, the work during my Ph.D. has resulted in another publication (Lyu et al., 2023). However, for the coherence of the contributions, Lyu et al. (2023) is not included in the thesis.

1. Yijie Zhang, Yi-Shan Wu, Luis A. Ortega, and Andres R Masegosa. The cold posterior effect indicates underfitting, and cold posteriors represent a fully bayesian method to mitigate it. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=GZORXGxHHT>.
2. Yi-Shan Wu, Yijie Zhang, Badr-Eddine Chérif-Abdellatif, and Yevgeny Seldin. Recursive PAC-Bayes: A frequentist approach to sequential prior updates with no information loss. *arXiv preprint arXiv:2405.14681*, 2024. URL <https://arxiv.org/abs/2405.14681>.
3. Yunlian Lyu, Aymeric Côme, Yijie Zhang, and Mohammad Sadegh Talebi. Scaling up q-learning via exploiting state-action equivalence. *Entropy*, 25(4):584, 2023.

Bibliography

- Abdullah A. Abdullah, Masoud Muhammed Hassan, and Yaseen T. Mustafa. A review on bayesian deep learning in healthcare: Applications and challenges. *IEEE Access*, 10:36538–36562, 2022.
- Ben Adlam, Jasper Snoek, and Samuel L. Smith. Cold posteriors and aleatoric uncertainty, 2020.
- 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.
- Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.
- 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.
- Thomas Bayes. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. Communicated by R. Price, in a letter to J. Canton.

- Felix Biggs and Benjamin Guedj. Tighter PAC-Bayes generalisation bounds by leveraging example difficulty. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.
- Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.
- Christopher M. Bishop. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer, 2007.
- P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016a. doi: 10.1111/rssb.12158. URL <https://academic.oup.com/jrsssb/article/78/5/1103/7040623>.
- 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, 2016b.
- 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.
- Natalia Bochkina and Peter J. Green. Bernstein - von mises theorem and misspecified models: a review. *arXiv preprint arXiv:2204.13614*, 2022. URL <https://arxiv.org/abs/2204.13614>.
- Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *CoNLL*, pages 10–21. ACL, 2016.
- Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.
- Badr-Eddine Chérif-Abdellatif and Pierre Alquier. Mmd-bayes: Robust bayesian estimation via maximum mean discrepancy. In *AABI*, volume 118 of *Proceedings of Machine Learning Research*, pages 1–21. PMLR, 2019.

- Ben Chugg, Hongjian Wang, and Aaditya Ramdas. A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *Journal of Machine Learning Research*, 24(372), 2023.
- Shay B. Cohen. *Bayesian Analysis in Natural Language Processing, Second Edition*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2019.
- Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2nd edition, 2006.
- Pedro Domingos. Bayesian averaging of classifiers and the overfitting problem. In *ICML*, 2000.
- Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*. AUAI Press, 2017a.
- Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*, 2017b.
- Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N. Siddharth, Brooks Paige, Dana H. Brooks, Jennifer G. Dy, and Jan-Willem van de Meent. Structured disentangled representations. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 2525–2534. PMLR, 2019.
- Andrew Foong, Wessel Bruinsma, David Burt, and Richard Turner. How tight can pac-bayes be in the small data regime? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Andrew Y. K. Foong, Wessel P. Bruinsma, and David R. Burt. A note on the chernoff bound for random variables in the unit interval. *arXiv preprint arXiv.2205.07880*, 2022.
- 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.

- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *arXiv preprint arXiv:1506.02142*, 2016. URL <https://arxiv.org/abs/1506.02142>.
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013a.
- Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2013b. ISBN 9781439840955.
- Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pages 1884–1892, 2016.
- John Geweke, Gary Koop, and Herman van Dijk, editors. *The Oxford Handbook of Bayesian Econometrics*. Oxford University Press, Oxford, 2011. ISBN 9780199559084.
- Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nat.*, 521(7553):452–459, 2015.
- Abhik Ghosh and Ayanendranath Basu. Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016. doi: 10.1007/s10463-014-0499-0. URL <https://econpapers.repec.org/RePEc:spr:aistmt:v:68:y:2016:i:2:p:413-437>.
- 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.
- 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. doi: 10.1214/17-BA1085. URL <https://doi.org/10.1214/17-BA1085>.
- 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*, pages 1321–1330. PMLR, 2017.

- Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schön. Evaluating scalable bayesian deep learning methods for robust computer vision. In *CVPR Workshops*, pages 1289–1298. Computer Vision Foundation / IEEE, 2020.
- Maxime Haddouche and Benjamin Guedj. PAC-Bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research Journal*, 2023.
- 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.
- Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016.
- C. C. Holmes and S. G. Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017. doi: 10.1093/biomet/asx010. URL <https://academic.oup.com/biomet/article/104/2/497/3074978>.
- Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International conference on artificial intelligence and statistics*, pages 703–711. PMLR, 2021.
- 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.
- Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter PAC-Bayes bounds through coin-betting. In *Proceedings of the Conference on Learning Theory (COLT)*, 2023.
- Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999.

- 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.
- Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.
- Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes’ rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23:1–109, 2022. URL <https://jmlr.org/papers/v23/19-1047.html>.
- John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6, 2005.
- Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL <http://yann.lecun.com/exdb/mnist/>.
- Sanae Lotfi, Pavel Izmailov, Gregory W. Benton, Micah Goldblum, and Andrew Gordon Wilson. Bayesian model selection, the marginal likelihood, and generalization. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 14223–14247. PMLR, 2022.
- Yunlian Lyu, Aymeric Côme, Yijie Zhang, and Mohammad Sadegh Talebi. Scaling up q-learning via exploiting state-action equivalence. *Entropy*, 25(4):584, 2023.
- 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.
- Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.
- David McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Conference on Learning Theory (COLT)*, 1998.
- David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37, 1999.

- Rowan McAllister, Yarin Gal, Alex Kendall, Mark van der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In *IJCAI*, pages 4745–4753. ijcai.org, 2017.
- Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected Bernstein inequality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- Jeffrey W. Miller and David B. Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527): 1113–1125, 2019. doi: 10.1080/01621459.2018.1469995. URL <http://hdl.handle.net/10.1080/01621459.2018.1469995>.
- 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*, pages 8270–8298. PMLR, 2022.
- Seth Nabarro, Stoil Kanev, 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.
- Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *NeurIPS*, pages 11013–11023, 2019.
- 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.
- Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Approximate thompson sampling via epistemic neural networks. In *UAI*, volume 216 of *Proceedings of Machine Learning Research*, pages 1586–1595. PMLR, 2023.
- 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*, pages 8024–8035, 2019.
- María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *J. Mach. Learn. Res.*, 22:227:1–227:40, 2021.
- María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 2021.
- Omar Rivasplata, Vikram M. Tankasali, and Csaba Szepesvári. Pac-bayes with backprop. *CoRR*, abs/1908.07380, 2019a.
- Omar Rivasplata, Vikram M Tankasali, and Csaba Szepesvari. Pac-bayes with backprop. *arXiv preprint arXiv:1908.07380*, 2019b.
- Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110), 2024.
- Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3, 2002.
- John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the Conference on Learning Theory (COLT)*, 1997.
- Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.
- Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. How to train deep variational autoencoders and probabilistic ladder networks. *CoRR*, abs/1602.02282, 2016.
- Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2017.

- Ilya Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.
- Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2): 1–305, 2008.
- Hao Wang and Dit-Yan Yeung. A survey on bayesian deep learning. *arXiv preprint arXiv:1604.01662*, 2016. URL <https://arxiv.org/abs/1604.01662>.
- Gregory Way and Casey Greene. Bayesian deep learning for single-cell analysis. *Nature Methods*, 15, 11 2018. doi: 10.1038/s41592-018-0230-9.
- Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, pages 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.
- Andrew Gordon Wilson. The case for bayesian deep learning. *CoRR*, abs/2001.10995, 2020.
- Yi-Shan Wu and Yevgeny Seldin. Split-kl and pac-bayes-split-kl inequalities for ternary random variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Yi-Shan Wu, Andres Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Chebyshev-cantelli pac-bayes-bennett inequality for the weighted majority vote. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Yi-Shan Wu, Yijie Zhang, Badr-Eddine Chérif-Abdellatif, and Yevgeny Seldin. Recursive PAC-Bayes: A frequentist approach to sequential prior updates with no information loss. *arXiv preprint arXiv:2405.14681*, 2024. URL <https://arxiv.org/abs/2405.14681>.

- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- Chen Zeno, Itay Golan, Ari Pakman, and Daniel Soudry. Why cold posteriors? on the suboptimal generalization of optimal bayes estimates. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021. URL <https://openreview.net/forum?id=cu6zDHCfhZx>.
- Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):2008–2026, 2019a.
- 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*, 2019b.
- Tong Zhang. From ε -entropy to kl-entropy: Analysis of minimum information complexity density estimation. *Annals of Statistics*, 2006.
- Yijie Zhang, Yi-Shan Wu, Luis A. Ortega, and Andres R Masegosa. The cold posterior effect indicates underfitting, and cold posteriors represent a fully bayesian method to mitigate it. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=GZORXGxHHT>.
- Zihao Zhang, Stefan Zohren, and Stephen Roberts. Bdlob: Bayesian deep convolutional neural networks for limit order books. *arXiv preprint arXiv:1811.10041*, 2019c. URL <https://arxiv.org/abs/1811.10041>.
- Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the imagenet scale: a pac-bayesian compression approach. In *ICLR (Poster)*. OpenReview.net, 2019.