Variational Inference of overparameterized Bayesian Neural Networks: a theoretical and empirical study of tempering

Anonymous Author
Anonymous Institution

Abstract

This paper studies the Variational Inference (VI) used for training Bayesian Neural Networks (BNN) in the overparameterized regime, *i.e.*, when the number of neurons tends to infinity. More specifically, we consider overparameterized two-layer BNN trained with VI and point out a critical issue in the mean-field regime. This problem arises from the decomposition of the lower bound on the evidence (ELBO) into two terms: one corresponding to the likelihood function of the model, promoting data-fitting, and the second to the Kullback-Leibler (KL) divergence between the prior distribution and the variational posterior, acting as a regularizer. In particular, we show both theoretically and empirically that there is a reasonable trade-off between these two terms in the overparameterized regime only when the KL term is appropriately re-scaled with respect to the ratio between the the number of observations and neurons. We also illustrate our theoretical results with numerical experiments that highlight the critical choice of this ratio.

1 Introduction

Bayesian neural networks (BNN) have gained popularity in the field of machine learning because they promise to combine the powerful approximation and discrimination properties of (deep) neural networks (NN) with the decision-theoretic approach of Bayesian inference. Among the advantages of BNN is their ability to provide uncertainty quantification, which is a must in many fields - e.g., autonomous driving (Michelmore et al. 2020; McAllister et al. 2017), computer vision (Kendall and Gal 2017), medical diagnosis (Filos et al. 2019). Second, the inclusion of prior information in some cases leads to better generalization error and calibration in classification tasks; see (Jospin et al. 2020; Izmailov et al. 2021) and references therein.

NN can be used to build complex probabilistic models for regression and classification tasks. Given $\bar{w}$ corresponding to the weights and bias of an NN, the network output can be used to define a (conditional) likelihood $\mathrm{L}(\{(x_i, y_i)\}_{i=1}^p|\bar{w})$ of some observed labels $\{y_i\}_{i=1}^p$, $y_i \in \mathsf{Y}$ associated with feature vectors $\{x_i\}_{i=1}^p$, $x_i \in \mathsf{X}$. Specifying a prior distribution for $\bar{w}$ and applying Bayes' rule yields the posterior distribution of weights. In the Bayesian approach, the goal is to find the predictive distribution from new feature vectors defined as an integral with respect to the posterior. One possible approach is to use Markov-Chain Monte Carlo methods - such as Hamiltonian Monte Carlo - for inference in Bayesian neural networks; (Neal et al. 2011; Hoffman, Gelman, et al. 2014; Betancourt 2017). However, the challenge of scaling HMC for applications involving high-dimensional parameter space and large datasets limits its broad application; (Cobb and Jalaian 2021). Computationally cheaper MCMC methods have been proposed, see (Welling and Teh 2011; Chen, Fox, and Guestrin 2014; Brosse, Durmus, and Moulines 2018); but these methods yield biased estimates of posterior expectation, see (Izmailov et al. 2021). A much simpler alternative from a computational standpoint is to use Variational Inference (VI) (Blundell et al. 2015; Gal and Ghahramani 2016; Louizos and Welling 2017; M. Khan et al. 2018), which approximates the posterior with a parametric distribution. Nevertheless, little is known about the validity or limitations of the latter approach, including the choice of prior, variational family, and their interplay.

A number of recent papers have investigated the limiting behavior of gradient descent type algorithms for one or two hidden layers in the overparameterized regime, (Chizat and Bach 2018; Rotskoff et al. 2019; Mei, Montanari, and Nguyen 2018; Tzen and Raginsky 2019; De Bortoli et al. 2020), *i.e.*, the number of hidden neurons goes to infinity. More specifically, it was found that the gradient descent applied to (true) risk minimization can be viewed as a temporal and spatial discretization of the Wasserstein gradient

Preliminary work. Under review by AISTATS 2023. Do not distribute.

flow of a limiting functional, which is defined on the space of probability distributions over the parameters by

$$\mathrm{R}_\mu(\mu) = \int \ell(y, \int s(\bar{w}, x)\mathrm{d}\mu(\bar{w})\mathrm{d}\pi(x, y)) + \mathrm{P}(\mu) \ , \quad (1)$$

where π is the data distribution over $\mathsf{X} \times \mathsf{Y}$, $s(\bar{w}, x)$ is the output prediction of the NN with parameter weights $\bar{w}$ and P plays the role of a penalty function. Roughly speaking, identifying this functional consists in noting that the risk R_w over the weights $\bar{w}$ of a NN coincides with R_μ on the set of empirical measures, *i.e.* for any $\bar{w} = (w_1, \dots, w_N)$ - where N is the number of neurons-, $\mathrm{R}_w(\bar{w}) = \mathrm{R}_\mu(\mu_N)$ with $\mu_N = N^{-1}\sum_{i=1}^N \delta_{w_i}$. This result emphasizes that in the overparameterized regime, the weights of a NN act as particle discretization of probability measures and the final prediction of a NN has a form of continuous mixture.

We are interested here in performing a similar analysis but for Variational Inference (VI) of two-layer Bayesian Neural Networks (BNN). In this setting, the weights of the NN are no longer fixed, but are sampled from a variational posterior, and the prediction of the NN is the empirical average of the prediction of each sample. The variational posterior is obtained by maximizing an objective function, the Evidence Lower Bound (ELBO) over a parameter space Ξ^N. It was empirically found that the maximization of the "vanilla" ELBO function can lead to very poor inference. To address this problem, a modification of this objective function is often considered, resulting from a decomposition into two terms of this function: one corresponding to the Kullback-Leibler (KL) divergence to the prior and the other to a marginal likelihood term. Based on this decomposition, the modified version of ELBO, called partially tempered ELBO, consists in multiplying the KL term by a temperature parameter.

Although this change has been justified intuitively or by purely statistical considerations, to our knowledge no formal results have been derived. Our first contribution is to show that this procedure is indeed mandatory in the overparameterized regime. More precisely, we show that if the temperature is not scaled appropriately with respect to the number of neurons and data points, one of the two terms becomes dominant and therefore the resulting posterior collapses onto the prior. Our second contribution is to propose a partially tempered version of ELBO that mitigates this effect, and that converges to a well-defined limiting functional when the number of neurons and data points approaches infinity, and the temperature parameter is appropriately chosen, similarly to risk minimization (1). Our conclusions are twofold. First, we highlight that under a specific model, the temperature in the ELBO has to scale as p/N where p is the number of observations and N is the number of neurons. Second, we show that performing VI for BNN in the overparameterized regime is equivalent to risk minimization over an extended space of probability measures. In summary, then, the use of VI for BNN enriches traditional NN models by producing predictions from a hierarchical mixture distribution. Ultimately, this shows that using VI to train overparameterized BNN amounts to empirical risk minimization and therefore should be interpreted with cautious for performing Bayesian uncertainty quantification.

This paper is organized as follows. Section 2 introduces the background of VI on BNN. Section 3 characterizes the inadequacy of these models in the limiting case of the mean field, when the data or prior variance do not scale, and identifies the well-posed regime. Section 4 discusses connections to related work and alternative choices of scaling for infinite-witdh NN. In Section 5, some numerical experiments are presented to illustrate our claims.

Notations. For any set $\mathsf{E} \subset \mathbb{R}^d$, we denote by $\mathcal{P}(\mathsf{E})$ the set of probability measures on E. For any $\nu, \mu \in \mathcal{P}(\mathsf{E})$ we denote by $\nu \otimes \mu$ the product measure and define by induction $\nu^{\otimes N} = \nu^{\otimes N-1} \otimes \nu$. For $\sigma \in \mathbb{R}^d$, the diagonal matrix in $\mathbb{R}^{d\times d}$ with diagonal $\sigma^2 = [\sigma_1^2, \dots, \sigma_d^2]$ will be denoted $\mathrm{diag}(\sigma^2)$. For any measurable map $\mathscr{T} : \mathbb{R}^d \to \mathbb{R}^d$ and probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$, we denote by $\mathscr{T}_\# \nu$ the pushforward measure of ν by $\mathscr{T}$. It is characterized by the transfer lemma, i.e. $\int F(y)\mathrm{d}\mathscr{T}_\#\nu(y) = \int F(\mathscr{T}(z))\mathrm{d}\nu(z)$ for any measurable and bounded function F. For $\mu, \pi \in \mathcal{P}(\mathsf{E})$, the Kullback-Leibler divergence between a distribution μ and π is defined by $\mathrm{KL}(\mu|\pi) = \int \log(\mathrm{d}\mu/\mathrm{d}\pi)d\mu$ where $\mathrm{d}\mu/\mathrm{d}\pi$ is the Radon-Nikodym derivative if μ is absolutely continuous with respect to π, and $\mathrm{KL}(\mu|\pi) = +\infty$ otherwise.

2 Variational inference for BNN objective

Consider a supervised setting where we have access to i.i.d. samples $\{(x_i, y_i)\}_{i=1}^p$, from a distribution π on $\mathsf{X} \times \mathsf{Y} \subset \mathbb{R}^{d_\mathsf{X}} \times \mathbb{R}^{d_\mathsf{Y}}$, and aim at predicting y given a new observation x. In this paper, we focus on a fully connected NN with one hidden layer and N neurons, and activation function $h : \mathbb{R}^{d_\mathsf{X}} \times \mathsf{X} \to \mathbb{R}$. A common example is

$$h(b_j, x) = \sigma(\langle b, x\rangle) \ , \quad (2)$$

for $b \in \mathbb{R}^{d_\mathsf{X}}$ and $x \in \mathsf{X}$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is the Rectified Linear Unit $\sigma(t) = \max(0, t)$ or sigmoid function $\sigma(t) = \mathrm{e}^t/(1 + \mathrm{e}^t)$, for $t \in \mathbb{R}$. In addition, for each neuron $j \in \{1, \dots, N\}$, denote by $b_j \in \mathbb{R}^{d_\mathsf{X}}$ and $a_j \in \mathbb{R}^{d_\mathsf{Y}}$ the j-th weights of the hidden and output layers respectively, and set $w_j = (b_j, a_j) \in \mathbb{R}^d$, $d = d_\mathsf{X} + d_\mathsf{Y}$, and $\bar{w} = (w_j)_{j=1}^N$ all the weights of the NN under consideration. With this notation, for each input $x \in \mathsf{X}$, the output prediction $f_{\bar{w}} : \mathsf{X} \to \mathbb{R}^{d_\mathsf{Y}}$ of the neural network can be written as:

$$f_{\bar{w}}(x) = \frac{1}{N}\sum_{j=1}^N s(w_j, x), \quad s(w_j, x) = a_j h(b_j, x). \quad (3)$$

Given a loss function $\ell : \mathsf{Y} \times \mathsf{Y} \to \mathbb{R}_+$, we use the prediction function $f_{\bar{w}}$ to define the conditional likelihood

$$\mathrm{L}(y|x, \bar{w}) \propto \exp(-\ell(f_{\bar{w}}(x), y)) \ , \tag{4}$$

with respect to the Lebesgue measure on $\mathbb{R}^{d\times N}$ denoted by $\mathrm{Leb}_{d\times N}$. Then, choosing a prior pdf P_0 on $\bar{w}$, the posterior pdf P of the weights is proportional to $\bar{w} \mapsto P_0(\bar{w}) \prod_{i=1}^p \mathrm{L}(y_i|x_i, \bar{w})$. We perform Bayesian inference using VI (M. E. Khan and Rue 2021; Blei, Kucukelbir, and McAuliffe 2017; Blundell et al. 2015; Graves 2011; M. Khan et al. 2018). The general procedure is to consider a variational family of pdfs $\mathscr{F}_\Theta = \{q_{\boldsymbol{\theta}} \ : \ \boldsymbol{\theta} \in \Theta\}$, for $\Theta \subset \mathbb{R}^{d_\theta}$ and to maximize the Evidence Lower Bound (ELBO) defined for any $\boldsymbol{\theta} \in \Theta$ by:

$$\begin{aligned}\mathrm{ELBO}^N(\boldsymbol{\theta}) &= -\mathrm{KL}(q_{\boldsymbol{\theta}} \,|\, P_0) \\ &+ \sum_{i=1}^p \int_{\mathbb{R}^{N\times d}} \log \mathrm{L}(y_i|x_i, \bar{w}) q_{\boldsymbol{\theta}}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w}) \ . \end{aligned} \tag{5}$$

It is known that maximizing ELBO^N is equivalent to minimizing $\boldsymbol{\theta} \mapsto \mathrm{KL}(q_{\boldsymbol{\theta}} \,|\, P)$. For this reason, VI consists in approximating the posterior distribution P by $q_{\boldsymbol{\theta}^\star}$ with $\boldsymbol{\theta}^\star \in \operatorname{argmax} \mathrm{ELBO}^N$. The first term in (5) acts as a penalty term to control the deviation of $q_{\boldsymbol{\theta}^*}$ from the prior P_0, while the second term plays the role of empirical risk and promotes data-fitting . In practice, however, it has been shown that the choice of the prior and the variational approximation ELBO^N is crucial for good performance. It was proposed by (G. Zhang et al. 2018; M. Khan et al. 2018; Osawa et al. 2019; Ashukha et al. 2020) to weaken the regularization term KL and consider a partially tempered version of ELBO^N, which for a cooling parameter $\eta > 0$ is given by

$$\begin{aligned}\mathrm{ELBO}^N_\eta(\boldsymbol{\theta}) &= -\eta \mathrm{KL}(q_{\boldsymbol{\theta}} \,|\, P_0) \\ &+ \sum_{i=1}^p \int_{\mathbb{R}^{N\times d}} \log \mathrm{L}(y_i|x_i, \bar{w}) q_{\boldsymbol{\theta}}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w}) \ . \end{aligned} \tag{6}$$

It has been shown in (Wenzel et al. 2020; Wilson and Izmailov 2020) that ELBO^N_η is the same as ELBO^N but considering instead of the true posterior P, a partially tempered posterior $P_T \propto \mathrm{L}^{1/T} P_0$, where the likelihood function is tempered for some temperature $T \geqslant 0$. The parameter η (or equivalently the temperature T) controls the tradeoff of the likelihood term with respect to the prior. Setting $\eta < 1$ corresponds to a *cold posterior*, where the likelihood term is strengthened so that the posterior is concentrated in regions of high likelihood. The case $\eta = 1$ corresponds to "plain" Bayesian inference, while $\eta > 1$ corresponds to *warm posterior* where the prior has a stronger influence on the posterior.

In a series of paper, (Grünwald 2012; Grünwald and Van Ommen 2017; Bhattacharya, Pati, and Yang 2019; Heide et al. 2020; Grunwald, Steinke, and Zakynthinou 2021) have shown, significantly extending earlier results of (Barron and Cover 1991; T. Zhang 2006), that partially tempered posteriors may have better statistical properties under model misspecification than the "plain" posterior as the number of data points goes to infinity (expressed in terms posterior contraction around the best approximation of the truth). These results have been derived for Generalized Linear Models and it is not clear how these results extend to BNN.

(Wilson and Izmailov 2020) more informally argues that tempering is not inconsistent with Bayesian principles and that it may be particularly relevant in a parametric setting (where the model is defined by parameters), as opposed to Bayesian Nonparametric approaches - e.g., Gaussian processes. Namely, while in nonparametric approaches the model capacity is automatically scaled with the available data, this is not the case in parametric approaches, where the model capacity (which is determined by the number of neurons and the neural network architecture) is chosen by the user. Model misspecification is the rule in such case, as we show in Section 3 for neural networks with a hidden layer. Other works have questioned the role of data augmentation to justify tempering. In (Aitchison 2020), the author argues that the choice of likelihood does not reflect properly the data: "curated" datasets such as CIFAR10, where many labelers agree on the label of a data point, are in favor of cold posteriors, while adding noise to the labels reverses this effect. In (Nabarro et al. 2021), the authors investigate the role of data augmentation further and conclude that the right model should include some data augmentation but tempering still demands an explanation in that setting. In the present work, we consider the role of overparametrization of the models. A priori, both effects - dataset curation and overparametrization of BNN - seem to encourage tempered posteriors through different but complementary aspects. Still, to the best of our knowledge, the choice of temperature with respect to the number of data points and network parameters has not been investigated theoretically, in particular in the context of BNN.

Other studies, e.g., (Farquhar, Osborne, and Gal 2019), noted that a potential cause of the predominance of the KL term in (5) stems from the choice of the prior. Indeed, it has been noticed that the role of P_0 is important since it leads to very different inferences, see (Fortuin 2021). In particular, using priors on $\bar{w}$ which factorize over the weights, *i.e.*,

$$P_0(\bar{w}) = \prod_{j=1}^N P_0^1(w_j) \ , \tag{7}$$

do not yield optimal performance and as a result (Tran et al. 2020; Fortuin et al. 2021; Ober and Aitchison 2021; Sun et al. 2019) have proposed the design of new priors which introduce correlation amongst the weights and/or heavier tails than Gaussian ones.

In the present work, we take a novel approach to justify the use of ELBO_η^N based on the so-called overparameterized regime and study the impact of the choice of the cooling parameter η. We assume that the prior and posterior factorizes over the neurons, *i.e.*, the prior takes the form (7) and for each $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N) \in \Xi^N$, $q_{\boldsymbol{\theta}}(\bar{w}) = \prod_{i=1}^N q_{\theta_j}^1(w_j)$, where P_0^1 and $\{q_{\theta_j}^1\}_{j=1}^N$ are distributions over $\Xi \subset \mathbb{R}^d$. In this case, the variational parameter space $\Theta = \Xi^N$ and the prior distribution for each neuron P_0^1 is the same. Further, we assume that for any $\theta \in \Xi$, the variational distribution writes $q_\theta^1 = \mathscr{T}_{\theta\#}\gamma$, i.e. as the pushforward of a reference probability measure with density γ by $\mathscr{T}_\theta$ where $\{\mathscr{T}_\theta : \theta \in \Xi\}$ is a family of C^1-diffeomorphisms on $\mathbb{R}^d$. A common choice for $\mathscr{T}_\theta$ is, setting $\theta = (\mu, \sigma) \in \mathbb{R}^d \times (\mathbb{R}_+^*)^d$,

$$\mathscr{T}_\theta : z \mapsto \mu + \sigma \odot z \;, \tag{8}$$

where $\odot$ is the component wise product; but of course much more sophisticated choices are possible. Then, by (3)-(4) and a change of variable, the ELBO can be expressed as

$$\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) = -\eta \sum_{j=1}^N \mathrm{KL}(q_{\theta_j}^1 | P_0^1) - \sum_{i=1}^p \mathrm{G}^N(\boldsymbol{\theta}; (x_i, y_i)) \tag{9}$$

with denoting the output of a neuron parametrized by $\theta \in \mathbb{R}^d$ for an input x_i by

$$\phi(\theta, z, x_i) = s(\mathscr{T}_\theta(z), x_i) \;, \tag{10}$$

and $\boldsymbol{z} = (z_1, \ldots, z_N) \in \mathbb{R}^{d \times N}$,

$$\mathrm{G}^N(\boldsymbol{\theta}; (x, y)) = \int \ell \left(y, \sum_{j=1}^N \frac{\phi(\theta_j, z_j, x)}{N} \right) \gamma^{\otimes N}(\mathrm{d}\boldsymbol{z}) \;. \tag{11}$$

The decomposition of the KL term in (9) as N terms results from the choice of considering priors and posteriors that factorize over neurons. Although the VI framework we are considering may seem overly simplistic in light of the above, it is the one most commonly used in practice, and therefore it is still very important to obtain useful guidelines for implementation in order to optimize its performance. Moreover, it is a first step before considering other VI methods with more complex priors and/or variational families. The expression of ELBO_η^N shows that the parameter η must be chosen to balance the two terms in (6)-(9) to obtain a well-posed objective functional as $N, p \to +\infty$ and a variational posterior $q_{\boldsymbol{\theta}^\star}$ different from the prior. Without this parameter, optimizing the ELBO^N (5) leads to the collapse of the variational posterior to the prior, as shown in the following proposition.

Proposition 1. Assume that $\mathscr{F}_\Theta$ is a family of Gaussians with diagonal covariance matrices, that $P_0 \in \mathscr{F}_\Theta$ and that X is compact. Let $\boldsymbol{\theta}^{*,N} = \mathrm{argmax}_{\boldsymbol{\theta} \in \Theta} \mathrm{ELBO}^N(\boldsymbol{\theta})$. Assume also that ℓ is the square loss or cross-entropy, and that σ is Lipschitz. Then, $\mathrm{KL}(q_{\boldsymbol{\theta}^{*,N}}, P_0) \to 0$ as $N \to \infty$.

This result and its proof, that can be found in Appendix A, are inspired from (Coker et al. 2022, Theorem 1, 2) who show that the moments of the predictive posterior collapse to the ones of the prior and that the KL converges to 0 as $N \to \infty$, when ℓ is the square loss or logistic loss and σ is odd. Proposition 1 states an analog result, but which holds for additional losses (i.e., also cross-entropy) and more general activation functions (e.g., non odd ones as RelU). This is partly due to our different scaling of the output of the neural network in N^{-1}, see (3), that differs from theirs, in $N^{-1/2}$. To obtain their result, (Coker et al. 2022) fundamentally rely on a kind of central limit theorem, which explains their scaling. This is the main reason why they must assume that the activation function is an odd function. Note that this latter condition is not satisfied for RelU. In contrast, by considering a mean-field regime with a scaling in N^{-1}, we can get rid off this condition by relying on a law of large numbers and encompass a larger set of losses and activation functions, including the ReLU activation function. The result of Proposition 1 highlights that optimizing ELBO^N becomes ill-posed as $N \to \infty$. This suggests that the optimal variational posterior tends to ignore the data fitting term in (9), and that η must be chosen to rebalance ELBO^N. In the next section, we provide a theoretical framework supporting tempering and then present our main results regarding the choice of η.

3 Identifying well-posed regimes for the ELBO with product priors

We follow the approach outlined in (Chizat and Bach 2018; Rotskoff et al. 2019; Mei, Montanari, and Nguyen 2018) for ERM. We first generalize the definition of ELBO_η^N defined in (9) over Ξ^N, to probability measures ν on Ξ. Indeed, the following result states that ELBO_η^N can be expressed as a functional of the empirical measure over the weights $\nu_N^{\boldsymbol{\theta}}$ defined for each variational parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \in \Xi^N$ by

$$\nu_N^{\boldsymbol{\theta}} = N^{-1} \sum_{i=1}^N \delta_{\theta_i} \;, \tag{12}$$

where δ_θ is the Dirac mass at $\theta \in \Xi$. Define $\mathcal{P}_N(\Xi)$ the subset of $\mathcal{P}(\Xi)$ which can be written as (12) for some $\boldsymbol{\theta} \in \Xi^N$ (i.e. discrete measures supported on N parameters).

Proposition 2. For any $N \in \mathbb{N}^*$, there exists a function F_η^N defined over $\mathcal{P}_N(\Xi)$ and valued in $\mathbb{R} \cup \{+\infty\}$ such that $\mathrm{F}_\eta^N(\nu_N^{\boldsymbol{\theta}}) = \mathrm{ELBO}_\eta^N(\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \in \Xi^N$.

Proof. Denote by $\mathscr{S}_N$ the set of permutations over $\{1, \ldots, N\}$ and for any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N) \in \Theta$, $\tau \in \mathscr{S}_N$, $\boldsymbol{\theta}^\tau = (\theta_{\tau(1)}, \ldots, \theta_{\tau(N)})$. Note that for any $\tau \in \mathscr{S}_N$, $\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) = \mathrm{ELBO}_\eta^N(\boldsymbol{\theta}^\tau)$. The proof is then completed upon using that $\boldsymbol{\theta} \mapsto \nu_N^{\boldsymbol{\theta}}$ is a bijection from $\Xi^N / \sim$ to

$\mathcal{P}_N(\Xi)$, where $\sim$ is the equivalence relation defined by $\boldsymbol{\theta} \sim \boldsymbol{\theta}'$ if $\exists \tau \in \mathscr{S}_N$ s.t. $\boldsymbol{\theta}' = \boldsymbol{\theta}^\tau$. □

Proposition 2 is a first step towards identifying an objective functional defined on $\mathcal{P}(\Xi)$, since F_η^N is a reparametrization of the ELBO (i.e., it has the same value) but is defined on empirical measures on supported on N atoms. The main caveat is that F_η^N cannot be non-trivially extended to a functional defined for a general probability measure on Ξ, because it depends on N through the integration of the loss function with respect to the $N \times d$ dimensional Gaussian noise in (11). However, in our next result, we show that, when restricted to empirical probabilities, as $N \to +\infty$, F_η^N is a perturbation of the functional $\tilde{\mathrm{F}}_\eta^N$ defined *over all probabilities in* $\mathcal{P}(\Xi)$ by

$$\tilde{\mathrm{F}}_\eta^N(\nu) = -\sum_{i=1}^p \tilde{\mathrm{G}}(\nu; (x_i, y_i)) - \eta N \int \mathrm{KL}(q_\theta^1 | P_0^1) \mathrm{d}\nu(\theta) \;, \tag{13}$$

where

$$\tilde{\mathrm{G}}(\nu; (x, y)) = \ell\left(y, \iint \phi(\theta, z, x) \mathrm{d}\nu(\theta) \mathrm{d}\gamma(z)\right) \;, \tag{14}$$

and ϕ is given by (10). The main difference between F_η^N and $\tilde{\mathrm{F}}_\eta^N$ is the place where the integration occurs with respect to the Gaussian measure. For F_η^N, the integration is on the loss function in (11), while in $\tilde{\mathrm{F}}_\eta^N$, the integration is only on the second argument of the loss in (14). We now define for any $\theta \in \Xi$ and $x \in \mathsf{X}$, $\tilde{\phi}(\theta, x) = \int \phi(\theta, z, x) \mathrm{d}\gamma(z)$. Consider the following assumption:

A1. (i) There exists $\mathtt{L}_\ell > 0$ such that for any $y \in \mathsf{Y}$, the function $\tilde{y} \mapsto \ell(y, \tilde{y})$ is $\mathtt{L}_\ell$-smooth: for any $\tilde{y}_1, \tilde{y}_2 \in \mathsf{Y}$,

$$\|\nabla_{\tilde{y}} \ell(y, \tilde{y}_1) - \nabla_{\tilde{y}} \ell(y, \tilde{y}_2)\| \leqslant \mathtt{L}_\ell \|\tilde{y}_1 - \tilde{y}_2\| \;.$$

(ii) There exists $C_\phi \geqslant 0$, such that for any $\theta \in \Xi$, $x \in \mathsf{X}$,

$$\int \left\|\phi(\theta, z, x) - \tilde{\phi}(\theta, x)\right\|^2 \mathrm{d}\gamma(z) \leqslant C_\phi \;.$$

Note that **A**1-(i) is satisfied for the quadratic or logistic loss if Y is bounded. We give practical conditions on the activation function σ, the prior P_0^1 and the set Ξ to ensure that **A**1-(ii) holds in the case where $\mathscr{T}_\theta$ is supposed to be of the form (8) for any $\theta \in \Xi$, later in this section after stating our general results.

Theorem 3. *Assume **A**1. Then, there exists $C \geqslant 0$ such that for any $N, p \in \mathbb{N}^*$, $\{(x_i, y_i)\}_{i=1}^p \in (\mathsf{X} \times \mathsf{Y})^p$, $\boldsymbol{\theta} \in \Xi^N$ and $\eta > 0$,*

$$\left|\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) - \tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}})\right| \leqslant Cp/N \;,$$

where $\nu_N^{\boldsymbol{\theta}}$ is defined in (12).

Proof. Using that for any $y \in \mathsf{Y}$, the function $\tilde{y} \mapsto \ell(y, \tilde{y})$ is $\mathtt{L}_\ell$-smooth, we get by (Nesterov 2004, Lemma 1.2.3), Proposition 2 and the definitions (9)-(11)-(13)-(14),

$$\begin{aligned}
&\left|\mathrm{F}_\eta^N(\nu_N^{\boldsymbol{\theta}}) - \tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}})\right| \\
&\leqslant \frac{\mathtt{L}_\ell}{2N^2} \sum_{i=1}^p \int \left\|\sum_{j=1}^N \phi(\theta_j, z_j, x_i) - \tilde{\phi}(\theta_j, x_i)\right\|^2 \mathrm{d}\gamma^{\otimes N}(\boldsymbol{z}) \\
&\leqslant \frac{\mathtt{L}_\ell}{2N^2} \sum_{i=1}^p \sum_{j=1}^N \int \left\|\phi(\theta_j, z_j, x_i) - \tilde{\phi}(\theta_j, x_i)\right\|^2 \mathrm{d}\gamma(z) \;.
\end{aligned}$$

The proof follows from **A**1-(ii). □

We also show in the following theorem that the minimization of F_η^N over $\mathcal{P}_N(\Xi^N)$ provides a good approximation for the minimization problem corresponding to $\tilde{\mathrm{F}}_\eta^N$ for sufficiently large N.

Theorem 4. *Assume **A**1 and that there exists $\nu_\star \in \mathcal{P}(\Xi)$ such that $\nu_\star \in \operatorname{argmax}_{\mathcal{P}(\Xi)} \tilde{\mathrm{F}}_\eta^N$. Suppose in addition that there exists $C_\phi^{\nu_\star} \geqslant 0$ such that for any $x \in \mathsf{X}$,*

$$\int \left\|\tilde{\phi}(\theta, x) - \int \tilde{\phi}(\theta', x) \mathrm{d}\nu_\star(\theta')\right\|^2 \mathrm{d}\nu_\star(\theta) \leqslant C_\phi^{\nu_\star} \;. \tag{15}$$

Then, there exists $C \geqslant 0$ such that for any $N, p \in \mathbb{N}^*$, $\{(x_i, y_i)\}_{i=1}^p \in (\mathsf{X} \times \mathsf{Y})^p$ and $\eta > 0$,

$$\left|\sup_{\boldsymbol{\theta} \in \Xi^N} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) - \sup_{\nu \in \mathcal{P}(\Xi)} \tilde{\mathrm{F}}_\eta^N(\nu)\right| \leqslant Cp/N \;.$$

Proof. Using Theorem 3, we easily get that for any $\boldsymbol{\theta} \in \Xi^N$,

$$\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) \leqslant \tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}}) + Cp/N \leqslant \sup_\nu \tilde{\mathrm{F}}_\eta^N(\nu) + Cp/N \;,$$

for some $C \geqslant 0$ independent of $\{(x_i, y_i)\}_{i=1}^p \in (\mathsf{X} \times \mathsf{Y})^p$ and $\eta > 0$. On the other hand, we have using that $\nu_\star$ is a maximizer of $\tilde{\mathrm{F}}_\eta^N$,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in \Xi^N} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) \geqslant \\
&\sup_\nu \tilde{\mathrm{F}}_\eta^N(\nu) - \int \left|\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) - \tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}})\right| \mathrm{d}\nu_\star^{\otimes N}(\boldsymbol{\theta}) \\
&\qquad - \int \left|\tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}}) - \tilde{\mathrm{F}}_\eta^N(\nu_\star)\right| \mathrm{d}\nu_\star^{\otimes N}(\boldsymbol{\theta}) \;.
\end{aligned} \tag{16}$$

Using **A**1, for any $y \in \mathsf{Y}$, $\tilde{y} \mapsto \ell(y, \tilde{y})$ is $\mathtt{L}_\ell$-smooth, we get by Nesterov 2004, Lemma 1.2.3, setting $\tilde{\phi}_N = (1/N) \sum_{j=1}^N \tilde{\phi}(\theta_j, x_i)$,

$$\begin{aligned}
&\int \left|\tilde{\mathrm{F}}_\eta^N(\nu_N^{\boldsymbol{\theta}}) - \tilde{\mathrm{F}}_\eta^N(\nu_\star)\right| \mathrm{d}\nu_\star^{\otimes N}(\boldsymbol{\theta}) \\
&\leqslant \mathtt{L}_\ell \sum_{i=1}^p \int \left\|\tilde{\phi}_N - \int \tilde{\phi}(\theta', x_i) \mathrm{d}\nu_\star(\theta')\right\|^2 \mathrm{d}\nu_\star^{\otimes N}(\boldsymbol{\theta}) \\
&\leqslant \mathtt{L}_\ell p C_\phi^{\nu_\star}/N \;.
\end{aligned} \tag{17}$$

Combining (16), (17) and Theorem 3 concludes the proof. □

The bounds of Theorems 3 and 4, concentrate for $p/N \to 0$, which corresponds to the current practical overparametrized regimes for NN, where the number of parameters of the network is larger than the number of data points. We now set the cooling parameter as $\eta = \tau p/N$ with $\tau > 0$. As stated in the next proposition, whose proof can be found in Appendix C, this tempering prevents the collapse of the variational posterior onto the prior when optimizing the tempered ELBO, in contrast with Proposition 1.

Proposition 5. Let $\boldsymbol{\theta}^{*,N} = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta})$. Assume $\eta = \tau p/N$ and that ℓ is the square loss. Then, $\limsup_{N\to\infty} \mathrm{KL}(q_{\boldsymbol{\theta}^{*,N}}, P_0) > 0$ as $N \to \infty$.

With the particular choice of tempering $\eta = \tau p/N$, the functional $\tilde{\mathrm{F}}_\eta^N$ depends only on the number of observations p but no longer on the number of neurons N. We denote, for that particular choice of η,

$$\begin{aligned}\mathrm{F}_\tau^p(\nu) &= p^{-1}\tilde{\mathrm{F}}_\eta^N(\nu) \\ &= -\frac{1}{p}\sum_{i=1}^p \tilde{\mathrm{G}}(\nu;(x_i,y_i)) - \tau\int \mathrm{KL}(q_\theta^1|P_0^1)\mathrm{d}\nu(\theta)\ .\end{aligned}$$

In our next result, we show that with high probability, $\mathrm{F}_\tau^p(\nu)$ provides a good approximation as $p \to \infty$ of the function

$$\begin{aligned}\mathrm{R}_\tau(\nu) = &-\int \tilde{\mathrm{G}}(\nu;(x,y))\mathrm{d}\pi(x,y) \\ &- \tau\int \mathrm{KL}(q_\theta^1|P_0^1)\mathrm{d}\nu(\theta)\ , \quad (18)\end{aligned}$$

where $\tilde{\mathrm{G}}$ is defined by (14).

Proposition 6. Assume **A**1 and that there exists $M_G > 0$, s.t. for any $\nu \in \mathcal{P}(\Xi)$, $0 \leqslant \tilde{\mathrm{G}}(\nu;(x,y)) \leqslant M_G$, for π-almost all $(x,y) \in \mathsf{X}\times\mathsf{Y}$. Suppose in addition that $\{(x_i,y_i)\}_{i=1}^p$ are i.i.d. with distribution π. Then, for any $\nu \in \mathcal{P}(\Xi)$ and $\delta > 0$, with probability $1-\delta$ at least, it holds

$$|\mathrm{F}_\tau^p(\nu) - \mathrm{R}_\tau(\nu)| \leqslant M_G\sqrt{\log(\delta/2)/(2p)}\ .$$

The proof follows from applying Hoeffding's inequality on the bounded i.i.d. variables $\tilde{\mathrm{G}}(\nu;(x_i,y_i))$ for $i = 1,\dots,p$.

It is worth noting that the limiting risk R_τ is similar to the one obtained in the analysis of the limiting behavior of gradient descent type algorithms for two-layer NN in the overparameterized regime, by (Chizat and Bach 2018; Rotskoff et al. 2019; Mei, Montanari, and Nguyen 2018; Tzen and Raginsky 2019; De Bortoli et al. 2020) - see (1). Moreover, the maximization of the ELBO using gradient descent can be viewed as a temporal and spatial discretization of the Wasserstein gradient flow of the limiting function (18).

We conclude this section by illustrating when the results of this section hold, i.e. when the assumptions are satisfied for a mean-field variational family associated with the family of C^1-diffeomorphisms $\{\mathscr{T}_\theta : \theta \in \Xi\}$ given in (8). Consider the following assumption.

A2. (i) The subset Ξ is a compact set of $\mathbb{R}^d \times (\mathbb{R}_+^*)^d$, and X, Y are compact sets of $\mathbb{R}^{d_\mathsf{X}}, \mathbb{R}^{d_\mathsf{Y}}$.

(ii) The probability measure γ satisfies $\int \|z\|^4 \mathrm{d}\gamma(z) < +\infty$.

(iii) For any $x \in \mathsf{X}$, there exists $\mathtt{L}_h \geqslant 0$ such that the function $b \mapsto h(b,x)$ is $\mathtt{L}_h$-Lipschitz on $\mathbb{R}^{d_\mathsf{X}}$ and $\sup_{x\in\mathsf{X},\, b\in\mathbb{R}^{d_\mathsf{X}}} |h(b,x)|/(1+\|b\|) < +\infty$.

(iv) The prior density P_0^1 is positive on $\mathbb{R}^d$ and satisfies $\theta \mapsto \mathrm{KL}(q_\theta^1|P_0^1)$ is continuous on Ξ.

Note that the condition that for any $x \in \mathsf{X}$, the condition $b \mapsto h(b,x)$ is $\mathtt{L}_h$-Lipschitz is automatically satisfied for h of the form (2) with σ the RELU or sigmoid function if X is bounded. Also, we verify in the next proposition, whose proof can be found in Appendix B, that $\theta \mapsto \mathrm{KL}(q_\theta^1|P_0^1)$ is continuous if P_0^1 and γ are non-degenerate Gaussian distributions.

Proposition 7. Assume **A**1-(i) and **A**2. Then **A**1-(ii) and the conditions of Theorem 4 hold.

4 Discussion on the Lazy versus Mean Field regime for BNN

In (3), we chose to scale the output of the Bayesian Neural Network by a factor $1/N$, where N is the number of neurons. In the Bayesian Neural Network literature, it is also common to consider scaling the output of the neural network by $1/\sqrt{N}$, i.e.

$$f_{\bar{w}}(x) = N^{-1/2}\sum_{j=1}^N s(w_j,x), \quad (19)$$

Regarding standard (non Bayesian) neural networks, both the scalings $1/\sqrt{N}$ and $1/N$ have been considered to study overparametrized (infinite-width) neural networks, and are referred to respectively as the Neural Tangent Kernel (NTK) regime (Jacot, Gabriel, and Hongler 2018) and the Mean Field (MF) regime (Chizat and Bach 2018; Mei, Montanari, and Nguyen 2018; Sirignano and Spiliopoulos 2020). These choices are briefly discussed and compared in (Chizat, Oyallon, and Bach 2019). The scaling $1/\sqrt{N}$ results in a so-called lazy training regime where the output of the neural network hardly varies with respect to its initialization. In contrast, the scaling $1/N$ allows to converge as $N \to \infty$ to a non degenerate dynamic described by a

partial differential equation. We now discuss these choices of possible scalings for Bayesian neural networks.

When the output of a BNN is scaled as (19), it is well-known that the output prediction function $f_{\bar{w}} : \mathsf{X} \to \mathbb{R}^{d_Y}$ under the prior converges weakly to a (prior) Gaussian Process as $N \to \infty$ (Neal 2012; Lee et al. 2019; Matthews et al. 2018; Garriga-Alonso, Rasmussen, and Aitchison 2019; Novak et al. 2019; Hron, Bahri, Sohl-Dickstein, et al. 2020); and under the posterior to a (posterior) Gaussian process (Hron, Bahri, Novak, et al. 2020) (assuming the likelihood is a bounded continuous function of the NN output).

We establish in the supplement, Appendix D, that even when choosing the scaling $1/\sqrt{N}$, our conclusions and in particular Theorem 4 still hold. Indeed, we show that the problem of optimizing the ELBO using (19) can be reformulated as the one corresponding to a $1/N$ scaling. However the choice of the temperature in the resulting ELBO is more subtle.

Here we consider the Mean Field scaling (3) and we obtain results on tempering that are easy to interpret, since the whole (KL) regularization term is reweighted with respect to the number of observations and neurons. We advocate more generally that the MF regime for Bayesian Neural networks deserves to be analyzed since it provides a simple model for overparametrized NN which comes with many interesting consequences that we derived previously; which are first steps towards the full understanding of NN training and performance. We now turn to a practical evaluation of our study.

Remark 8. One may argue that with the $1/N$ scaling (3), the variance of the prior distribution collapses to zero as $N \to \infty$. However, the distribution which really matters is the (variational) posterior for which the variance is expected not to vanish with the tempering we propose, with non odd activation functions. We confirm this statement using a numerical experiment provided in Appendix E.1.

5 Experiments

In this section we illustrate our findings and their practical implications for image classification on standard datasets (MNIST, CIFAR-10). The reader may refer to Appendix E for additional experiments, including on regression tasks, that highlight the importance of rescaling the ELBO. Our code will be publicly available. In this section, we illustrate the influence of the parameter τ through different metrics.

Evaluation. Let $\mathcal{D} = (x_i, y_i)_{i=1}^p$ be a dataset, where $y_i = c \in \{1, \dots, n_l\}$ is a discrete class label. For an input $x \in \mathsf{X}$, the predictive probability of a class c by a neural network with weights $\bar{w}$ is defined by $\Psi_c(f_{\bar{w}}(x))$, where $\Psi_c(f_{\bar{w}}(x))$ denotes the c-th component of the softmax function applied to the output $f_{\bar{w}}(x) \in \mathbb{R}^{n_l}$ of the neural network. The cross entropy loss writes $\ell_{\mathrm{CE}}(y, f_{\bar{w}}(x)) = -\sum_{c=1}^{n_l} \tilde{y}_c \log(\Psi_c(f_{\bar{w}}(x)))$, where $\tilde{y}_c$ denotes the c-th coordinate of a one-hot representation of the label y and the Negative Log Likelihood (NLL) $\sum_{i=1}^p \int_{\mathbb{R}^{N\times d}} \ell_{\mathrm{CE}}(y_i, f_{\bar{w}}(x_i)) q_{\boldsymbol{\theta}}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w})$. The calibration performance of the model can be estimated by the Expected Calibration Error (ECE) (Naeini, Cooper, and Hauskrecht 2015), see also Appendix E.3. We recall that model is calibrated if the predictive posterior is the true probability for each class $c \in \{1, \dots, n_l\}$. However, since these probabilities are unknown, they have to be estimated, e.g. through ECE. As the NLL, ECE penalizes low probabilities assigned to correct predictions and high probabilities assigned to wrong ones; but these evalutation metrics are not strictly equivalent. To make our prediction, for $x \in \mathsf{X}$, we use the posterior predictive distribution defined for a class c as $\int \Psi_c(f_{\bar{w}}(x)) q_{\boldsymbol{\theta}^\star}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w})$ with $\boldsymbol{\theta}^\star$ obtained by minimization of ELBO_η^N by Bayes by Backprop. This integral is estimated by an empirical version

$$\int \Psi_c(f_{\bar{w}}(x)) q_{\boldsymbol{\theta}^\star}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w}) \approx \frac{1}{m} \sum_{l=1}^m \Psi_c(f_{\bar{w}_l}(x)) \,,$$

where for $l = 1, \dots, m$, $\bar{w}_l$ are i.i.d. samples from $q_{\boldsymbol{\theta}^\star}$. All the evaluation metrics mentioned above (NLL, ECE), as well as the accuracy are estimated using the same procedure. We will present our results on the MNIST dataset (where $p = 6.10^4$) and the CIFAR-10 dataset (where $p = 5.10^4$) (Krizhevsky, Hinton, et al. 2009).

Setup. We use a Linear BNN on MNIST, and ResNet20 architecture (He et al. 2016) on CIFAR-10 (Simonyan and Zisserman 2014). For CIFAR-10, we use the standard data augmentation techniques, see (M. Khan et al. 2018). For each neuron, we use a centered Gaussian prior with variance $1/5$, following (Osawa et al. 2019). We train each BNN by Bayes by Backprop (Blundell et al. 2015) with the reparametrization trick (see Appendix F) and using batch normalization (Sergey and Christian 2021).

Figure 1, 2 illustrate the performance of the different models and data sets for different values of τ. We evaluate the models on the test set in terms of their accuracy, NLL, ECE, and average confidence over the test set. In all experiments, we take $m = 50$ to approximate a BNN prediction and average our results over 5 experiments for each τ. It is worth noting that for a large τ, the accuracy decreases while the NLL increases. This is hardly a surprise, since the KL regularization forces the VI posterior to stay close to the prior distribution, resulting in underfitting. At the same time, the ECE value is low because of the poor confidence in the model, which is reflected in the accuracy. For small values of τ, the data fitting term is privileged, so the accuracy of the model is high, while the NLL is low. At the same time, the confidence in the model is very high, resulting in a low ECE. For intermediate values of τ, the accuracy of the models starts to decrease, but slower than the confidence

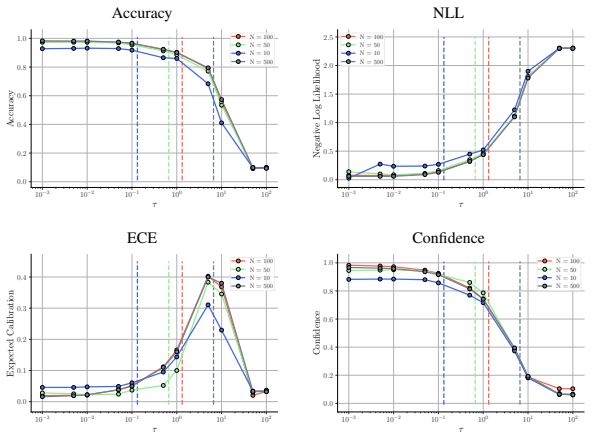

Figure 1: Effect of the temperature for a Linear BNN trained on MNIST. No cooling $\eta = 1$ is indicated by a red line.

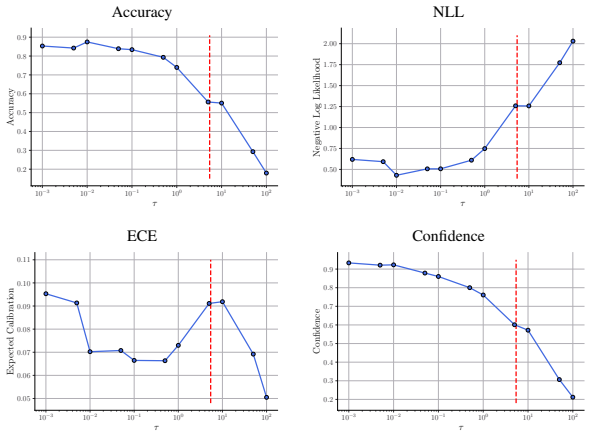

Figure 2: Effect of the temperature for a Resnet20 trained on CIFAR-10. No cooling $\eta = 1$ is indicated by a red line.

in the model, which explains an increase in ECE. We also illustrate the different regimes for the parameter τ with additional experiments in Appendix E, including analysis of the weights distribution and out-of-distribution detection.

Figure 3 displays the NLL and accuracy of 7 networks with an increasing number of neurons, and trained on the same dataset (MNIST) by optimizing the classical ELBO. The ratio p/N evolves here as the dataset is fixed and only the network size changes. Figure 3 highlights that the performance decreases as N increases and suggests the critical role of the ratio p/N.

6 Conclusion

In this work, we studied BNN trained with mean-field VI in the overparameterized regime. We have highlighted both theoretically and numerically that the partially tempered

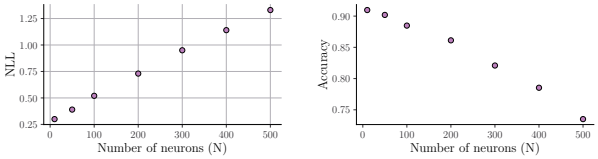

Figure 3: NLL and accuracy after training using the classical ELBO on MNist with an increasing number of neurons

ELBO_η^N advocated for VI for BNN effectively addresses the potential imbalance between the data fitting and KL terms. For mean-field VI and product prior distributions, we found that the cooling parameter must be chosen proportional to the ratio between the number of observations and neurons to achieve a balance between the data fitting and KL regularizer. With this choice, ELBO_η^N converges to a limiting functional that has the same structure as the one given by (Chizat and Bach 2018; Rotskoff et al. 2019; Mei, Montanari, and Nguyen 2018; Tzen and Raginsky 2019; De Bortoli et al. 2020) for empirical risk minimization. We also explained why, in the absence of cooling, the KL term can dominate the data fitting term, typically leading to underfitting of the model, which in practice translates into poor results on all metrics considered. Our work therefore provides a well-grounded theoretical justification for the importance of using a partial tempering in the overparameterized framework, which completes the justifications given by (Wenzel et al. 2020; Izmailov et al. 2021; Nabarro et al. 2021; Noci et al. 2021; Laves et al. 2021). While our theoretical results apply to a neural network with a single hidden layer, we have shown numerically that similar conclusions can be drawn for more general NN architectures. We emphasize that the introduction of a cooling factor into the Mean-Field VI for BNN is not without implications for the validity of Bayesian inference, and that the conclusions that can be drawn in this framework-in particular, Bayesian uncertainty quantification-must therefore be used with care (even though the accuracy, NLL, and ECE metrics obtained with Mean-Field VI compare favorably to their "classical" ERM learning counterparts).

References

Aitchison, L. (2020). "A statistical theory of cold posteriors in deep neural networks". In: *arXiv preprint arXiv:2008.05912*.

Ambrosio, L., N. Gigli, and G. Savaré (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Ashukha, A. et al. (2020). "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning". In: *arXiv preprint arXiv:2002.06470*.

Barron, A. R. and T. M. Cover (1991). "Minimum complexity density estimation". In: *IEEE transactions on information theory* 37.4, pp. 1034–1054.

Betancourt, M. (2017). "A conceptual introduction to Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1701.02434*.

Bhattacharya, A., D. Pati, and Y. Yang (2019). "Bayesian fractional posteriors". In: *The Annals of Statistics* 47.1, pp. 39–66.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Blundell, C. et al. (2015). "Weight uncertainty in neural network". In: *International Conference on Machine Learning*. PMLR, pp. 1613–1622.

Brosse, N., A. Durmus, and E. Moulines (2018). "The promises and pitfalls of Stochastic Gradient Langevin Dynamics". In: *NeurIPS*.

Chen, T., E. Fox, and C. Guestrin (2014). "Stochastic gradient hamiltonian monte carlo". In: *International conference on machine learning*. PMLR, pp. 1683–1691.

Chizat, L. and F. Bach (2018). "On the global convergence of gradient descent for over-parameterized models using optimal transport". In: *arXiv preprint arXiv:1805.09545*.

Chizat, L., E. Oyallon, and F. Bach (2019). "On lazy training in differentiable programming". In: *Advances in Neural Information Processing Systems* 32.

Cobb, A. D. and B. Jalaian (2021). "Scaling Hamiltonian Monte Carlo inference for Bayesian neural networks with symmetric splitting". In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 675–685.

Coker, B. et al. (2022). "Wide Mean-Field Bayesian Neural Networks Ignore the Data". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 5276–5333.

De Bortoli, V. et al. (2020). "Quantitative propagation of chaos for sgd in wide neural networks". In: *arXiv preprint arXiv:2007.06352*.

Farquhar, S., M. Osborne, and Y. Gal (2019). "Radial bayesian neural networks: Robust variational inference in big models". In: *Training* 80, p. 100.

Filos, A. et al. (2019). "A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks". In: *arXiv preprint arXiv:1912.10481*.

Fortuin, V. (2021). "Priors in bayesian deep learning: A review". In: *arXiv preprint arXiv:2105.06868*.

Fortuin, V. et al. (2021). "Bayesian neural network priors revisited". In: *arXiv preprint arXiv:2102.06571*.

Gal, Y. and Z. Ghahramani (2016). "Dropout as a bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. PMLR, pp. 1050–1059.

Garriga-Alonso, A., C. E. Rasmussen, and L. Aitchison (2019). "Deep convolutional networks as shallow gaussian processes". In: *ICLR*.

Graves, A. (2011). "Practical variational inference for neural networks". In: *Advances in neural information processing systems* 24.

Grunwald, P., T. Steinke, and L. Zakynthinou (2021). "PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, pp. 2217–2247.

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

Grünwald, P. and T. Van Ommen (2017). "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it". In: *Bayesian Analysis* 12.4, pp. 1069–1103.

He, K. et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Heide, R. et al. (2020). "Safe-Bayesian generalized linear regression". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2623–2633.

Hernández-Lobato, J. M. and R. Adams (2015). "Probabilistic backpropagation for scalable learning of bayesian neural networks". In: *International conference on machine learning*. PMLR, pp. 1861–1869.

Hoffman, M. D., A. Gelman, et al. (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *J. Mach. Learn. Res.* 15.1, pp. 1593–1623.

Hron, J., Y. Bahri, R. Novak, et al. (2020). "Exact posterior distributions of wide Bayesian neural networks". In: *arXiv preprint arXiv:2006.10541*.

Hron, J., Y. Bahri, J. Sohl-Dickstein, et al. (2020). "Infinite attention: NNGP and NTK for deep attention networks". In: *International Conference on Machine Learning*. PMLR, pp. 4376–4386.

Izmailov, P. et al. (2021). "What Are Bayesian Neural Network Posteriors Really Like?" In: *International Conference on Machine Learning*.

Jacot, A., F. Gabriel, and C. Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31.

Jospin, L. V. et al. (2020). "Hands-on Bayesian Neural Networks–a Tutorial for Deep Learning Users". In: *arXiv preprint arXiv:2007.06823*.

Kendall, A. and Y. Gal (2017). "What uncertainties do we need in bayesian deep learning for computer vision?" In: *arXiv preprint arXiv:1703.04977*.

Khan, M. et al. (2018). "Fast and scalable bayesian deep learning by weight-perturbation in adam". In: *International Conference on Machine Learning*. PMLR, pp. 2611–2620.

Khan, M. E. and H. Rue (2021). "The Bayesian learning rule". In: *arXiv preprint arXiv:2107.04562*.

Krizhevsky, A., G. Hinton, et al. (2009). "Learning multiple layers of features from tiny images". In:

Laves, M.-H. et al. (2021). "Posterior Temperature Optimization in Variational Inference for Inverse Problems". In: *arXiv preprint arXiv:2106.07533*.

Lee, J. et al. (2019). "Wide neural networks of any depth evolve as linear models under gradient descent". In: *Advances in neural information processing systems* 32.

Louizos, C. and M. Welling (2017). "Multiplicative normalizing flows for variational bayesian neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 2218–2227.

Matthews, A. G. d. G. et al. (2018). "Gaussian process behaviour in wide deep neural networks". In: *arXiv preprint arXiv:1804.11271*.

McAllister, R. et al. (2017). "Concrete Problems for Autonomous Vehicle Safety: Advantages of Bayesian Deep Learning". In: *IJCAI*.

Mei, S., A. Montanari, and P.-M. Nguyen (2018). "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33, E7665–E7671.

Michelmore, R. et al. (2020). "Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 7344–7350.

Nabarro, S. et al. (2021). "Data augmentation in Bayesian neural networks and the cold posterior effect". In: *arXiv preprint arXiv:2106.05586*.

Naeini, M. P., G. Cooper, and M. Hauskrecht (2015). "Obtaining well calibrated probabilities using bayesian binning". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Neal, R. M. et al. (2011). "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2.11, p. 2.

Neal, R. M. (2012). *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer.

Netzer, Y. et al. (2011). "Reading digits in natural images with unsupervised feature learning". In:

Noci, L. et al. (2021). "Disentangling the Roles of Curation, Data-Augmentation and the Prior in the Cold Posterior Effect". In: *arXiv preprint arXiv:2106.06596*.

Novak, R. et al. (2019). "Bayesian deep convolutional networks with many channels are gaussian processes". In: *ICLR*.

Ober, S. W. and L. Aitchison (2021). "Global inducing point variational posteriors for bayesian neural networks and deep gaussian processes". In: *International Conference on Machine Learning*. PMLR, pp. 8248–8259.

Osawa, K. et al. (2019). "Practical deep learning with Bayesian principles". In: *arXiv preprint arXiv:1906.02506*.

Rotskoff, G. et al. (2019). "Global convergence of neuron birth-death dynamics". In: *arXiv preprint arXiv:1902.01843*.

Sergey, I. and S. Christian (2021). "Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv 2015". In: *arXiv preprint arXiv:1502.03167*.

Simonyan, K. and A. Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Sirignano, J. and K. Spiliopoulos (2020). "Mean field analysis of neural networks: A law of large numbers". In: *SIAM Journal on Applied Mathematics* 80.2, pp. 725–752.

Sun, S. et al. (2019). "Functional variational Bayesian neural networks". In: *International Conference on Learning Representations*.

Tran, B.-H. et al. (2020). "All you need is a good functional prior for bayesian deep learning". In: *arXiv preprint arXiv:2011.12829*.

Tzen, B. and M. Raginsky (2019). "Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit". In: *arXiv preprint arXiv:1905.09883*.

Welling, M. and Y. W. Teh (2011). "Bayesian learning via stochastic gradient Langevin dynamics". In: *International conference on machine learning*. Citeseer, pp. 681–688.

Wenzel, F. et al. (2020). "How good is the bayes posterior in deep neural networks really?" In: *International conference on machine learning*.

Wilson, A. G. and P. Izmailov (2020). "Bayesian deep learning and a probabilistic perspective of generalization". In: *arXiv preprint arXiv:2002.08791*.

Zhang, G. et al. (2018). "Noisy natural gradient as variational inference". In: *International Conference on Machine Learning*. PMLR, pp. 5852–5861.

Zhang, T. (2006). "Information-theoretic upper and lower bounds for statistical estimation". In: *IEEE Transactions on Information Theory* 52.4, pp. 1307–1321.

A Proof of Proposition 1

We have assumed that $\mathscr{F}_\Theta$ is a family of Gaussians with diagonal covariance matrices, and that $P_0 \in \mathscr{F}_\Theta$ hence there exists $\boldsymbol{\theta}_0 = (\mu, \sigma) \in \mathbb{R}^{N(d_\mathsf{X}+d_\mathsf{Y})} \times \mathbb{R}^{N(d_\mathsf{X}+d_\mathsf{Y})}$ such that $P_0 = q_{\boldsymbol{\theta}_0}$. For ease of notations, we work with P_0 standard Gaussian:

$$P_0(\bar{w}) = \prod_{j=1}^{N}\prod_{l=1}^{d_\mathsf{Y}} \mathcal{N}(a_{i,j}; 0, 1) \times \prod_{j=1}^{N}\prod_{l=1}^{d_\mathsf{X}} \mathcal{N}(b_{j,l}; 0, 1) \tag{20}$$

Our results hold for more general parameters for P_0 but we fix these ones for convenience of notations. The posterior $q_{\boldsymbol{\theta}} \in \mathscr{F}_\Theta$ is:

$$q_{\boldsymbol{\theta}}(\bar{w}) = \prod_{j=1}^{N}\prod_{l=1}^{d_\mathsf{Y}} \mathcal{N}(a_{j,l}; \mu_{a_{j,l}}, \sigma^2_{a_{j,l}}) \times \prod_{j=1}^{N}\prod_{l=1}^{d_\mathsf{X}} \mathcal{N}(b_{j,l}; \mu_{b_{j,l}}, \sigma^2_{b_{j,l}}) \tag{21}$$

We define $a_j = (a_{j,1}, \dots, a_{j,d_\mathsf{Y}}) \in \mathbb{R}^{d_\mathsf{Y}}$ and $b_j = (b_{1,j}, \dots, b_{d_\mathsf{X},j}) \in \mathbb{R}^{d_\mathsf{X}}$ respectively the j^{th} row of the first layer weight matrix and the j^{th} column of the second layer weight matrix. We denote $\mu_{a_j} = (\mu_{a_{j,1}}, \dots, \mu_{a_{j,d_\mathsf{Y}}}) \in \mathbb{R}^{d_\mathsf{Y}}$, $\mu_a = (\mu_{a_1}, \dots, \mu_{a_N}) \in \mathbb{R}^{Nd_\mathsf{Y}}$.

Recall that

$$\mathrm{ELBO}^N(\boldsymbol{\theta}) = -\mathcal{L}(q_{\boldsymbol{\theta}}) - \mathrm{KL}(q_{\boldsymbol{\theta}}|q_{\boldsymbol{\theta}_0}), \quad \text{with } \mathcal{L}(q_{\boldsymbol{\theta}}) = -\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}}\left[\sum_{i=1}^{p} \log(\mathrm{L}(y_i|x_i, \bar{w}))\right], \tag{22}$$

where $\mathrm{L}(y|x, \bar{w}) \propto \exp(-\ell(f_{\bar{w}}(x), y))$ is defined by (4).

By the optimality of $\boldsymbol{\theta}^\star$, we have:

$$\mathrm{ELBO}^N(\boldsymbol{\theta}^\star) \geqslant \mathrm{ELBO}^N(\boldsymbol{\theta}_0),$$

Hence,

$$\mathrm{KL}(q_{\boldsymbol{\theta}^*}|q_{\boldsymbol{\theta}_0}) \leqslant \mathcal{L}(q_{\boldsymbol{\theta}_0}) - \mathcal{L}(q_{\boldsymbol{\theta}^*}). \tag{23}$$

We now deal separately with the square loss (Case 1) and cross-entropy loss (Case 2). Throughout, we will often use the notation $\sigma_j = \sigma(< b_j, x >)$ for any $j = 1, \dots, N$ and a generic point $x \in \mathsf{X}$. Since we have assumed that σ is L-Lipschitz, for any $y \in \mathbb{R}$, $|\sigma(y)| \leq |\sigma(0)| + L|y|$. Also, to explicit the dependence of $\boldsymbol{\theta}^*, \boldsymbol{\theta}_0$ in N we will write their associated distributions $q^N_{\boldsymbol{\theta}^*}$ and $q^N_{\boldsymbol{\theta}_0}$ respectively.

A.1 Case of the square loss

Proof. The idea of the proof is to show that the right hand side term of (23) converges to zero by showing that the two negative log likelihoods converge to the same finite limit, and hence their difference to zero as N goes to infinity. When l is the square loss, for any $q^N_{\boldsymbol{\theta}} \in \mathscr{F}_\Theta$, by (22) we have

$$\mathcal{L}(q^N_{\boldsymbol{\theta}}) = \sum_{i=1}^{p} \mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}\left[\|y_i\|^2 + \|f_{\bar{w}}(x_i)\|^2 - 2\langle y_i, f_{\bar{w}}(x_i)\rangle + \log(Z)\right],$$

where Z is the normalization constant of the model defined by (4). We will show that for both the prior $q^N_{\boldsymbol{\theta}_0}$ and optimal posterior $q^N_{\boldsymbol{\theta}^*}$, the first and second moment of the predictive distribution converge to zero as N goes to infinity.

Under the prior distribution (20), for any $x \in \mathsf{X}$, the first and second moments of the predictive distribution can be written:

$$\begin{aligned}
\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}_0}}[f_{\bar{w}}(x)] &= \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^N \mathsf{w}_j a_j\right] = \frac{1}{N}\sum_{j=1}^N \mathbb{E}[\mathsf{w}_j]\mathbb{E}[a_j] = 0 \\
\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}_0}}[\|f_{\bar{w}}(x)\|^2] &= \mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}_0}}\left[\frac{1}{N^2}\left(\sum_{j=1}^N \mathsf{w}_j^2\|a_j\|^2 + 2\sum_{j=1}^N\sum_{k<j}\mathsf{w}_j\mathsf{w}_k\langle a_j, a_k\rangle\right)\right] \\
&= \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}_0}}[\mathsf{w}_j^2] \leq \frac{1}{N^2}\left(\sum_{j=1}^N |\sigma(0)|^2 + L^2\|x\|^2\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}_0}}[\|b_j\|^2]\right) \\
&\xrightarrow[N\to\infty]{} 0,
\end{aligned}$$

Hence we first obtain:

$$\lim_{N\to\infty}\mathcal{L}(q^N_{\boldsymbol{\theta}_0}) = \sum_{i=1}^p \|y_i\|^2 + \log Z. \tag{24}$$

We now turn to showing that $\mathcal{L}(q^N_{\boldsymbol{\theta}^*})$ has the same limit. First notice that since $\mathcal{L}$ is a positive function, by (23) we have: $\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0}) \leqslant \mathcal{L}(q^N_{\boldsymbol{\theta}_0})$. Since the right-hand term is a converging sequence, it means that $\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0})$ is bounded by a constant C_{KL} independent of N.

By applying Lemma 9 and Lemma 10, we have:

$$\begin{aligned}
\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}^*}}[< y_i, f_{\bar{w}}(x_i) >] &\leqslant \|y_i\|\|\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}^*}}[f_{\bar{w}}(x_i)]\| \leqslant \frac{\phi(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}), \mathsf{X}, d_{\mathsf{Y}})}{\sqrt{N}} \leqslant \frac{\phi(C_{\mathrm{KL}}, \mathsf{X}, d_{\mathsf{Y}})}{\sqrt{N}} \\
\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}^*}}[\|f_{\bar{w}}(x)\|^2] &\leqslant \frac{\psi(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}), \mathsf{X}, d_{\mathsf{Y}})}{\sqrt{N}} \leqslant \frac{\psi(C_{\mathrm{KL}}, \mathsf{X}, d_{\mathsf{Y}})}{\sqrt{N}}
\end{aligned}$$

where the most right hand side inequalities come from the fact that $\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0})$ is bounded by a constant C_{KL} independent of N; and $\phi(C_{\mathrm{KL}}, \mathsf{X}, d_{\mathsf{Y}}), \psi(C_{\mathrm{KL}}, \mathsf{X}, d_{\mathsf{Y}})$ are constants that only depend on the data points $(x_i, y_i)_{i=1}^p$, the spaces X, Y and parameters of the prior distribution (through C_{KL}). Hence, the first and second moments of the predictive under the posterior $q^N_{\boldsymbol{\theta}^*}$ converge to 0. Hence, we obtain:

$$\lim_{N\to\infty}\mathcal{L}(q^N_{\boldsymbol{\theta}^*}) = \sum_{i=1}^p \|y_i\|^2 + \log Z. \tag{25}$$

From (23), (24) and (25) we finally that

$$\lim_{N\to\infty}\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}) = 0.$$

A.2 Case of the cross-entropy

Proof. Similarly to the square loss case, the idea of the proof is to show that $\mathcal{L}(q^N_{\boldsymbol{\theta}_0}), \mathcal{L}(q^N_{\boldsymbol{\theta}^*})$ have the same limit. We will make use of Lemma 11 which specify that limit under a null moment assumption.

Under the prior distribution $q^N_{\boldsymbol{\theta}_0}$,

$$\|\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}_0}}[\frac{1}{N}\sum_{j=1}^N \mathsf{w}_j a_j]\| = \frac{1}{N}\|\sum_{j=1}^N \mathbb{E}[\mathsf{w}_j]\mathbb{E}[a_j]\| = 0,$$

hence by Lemma 11:

$$\lim_{N\to\infty}\mathcal{L}(q^N_{\boldsymbol{\theta}_0}) = p(\log(d_{\mathsf{Y}}) + \log Z).$$

We now turn to the predictive distribution under the posterior $q^N_{\boldsymbol{\theta}^*}$. Recall that since $\mathcal{L}$ is a positive function, using the

optimality of the posterior we have: $\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0}) \leqslant \mathcal{L}(q^N_{\boldsymbol{\theta}_0})$. Since the right-hand term is a converging sequence, it means that $\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0})$ is bounded by a constant C_{KL} independent of N.

By Lemma 9, we can bound the first moment of the predictive distribution as:

$$\|\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}^*}}[f_{\bar{w}}(x)]\| \leqslant \frac{\phi(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}), \mathsf{X}, d_{\mathsf{Y}})}{\sqrt{N}} \leqslant \frac{\phi(C_{\mathrm{KL}}, \mathsf{X}, d_{\mathsf{Y}})}{\sqrt{N}},$$

where the last inequality comes from the fact that the KL term is bounded by a constant C_{KL} independent of N for the optimal variational parameter $\boldsymbol{\theta}^{*,N}$. Moreover, by using similar argument than in the proof of Lemma 9, we can show that each coordinate μ, σ of $\boldsymbol{\theta}^{*,N}$ is bounded as:

- $\mu \leqslant \sqrt{2\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0})} \leqslant \sqrt{2C_{\mathrm{KL}}}$
- $\sigma \leqslant 2\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}) + 1 \leqslant 2C_{\mathrm{KL}} + 1$

It means that each neuron weight has bounded mean and variance. We can thus apply Lemma 11, which yields:

$$\lim_{N\to\infty} \mathcal{L}(q^N_{\boldsymbol{\theta}}) = p(\log(d_{\mathsf{Y}}) - \log Z).$$

As $0 \leqslant \mathrm{KL}(q^N_{\boldsymbol{\theta}^*}|q^N_{\boldsymbol{\theta}_0}) \leqslant \mathcal{L}(q^N_{\boldsymbol{\theta}_0}) - \mathcal{L}(q_{\boldsymbol{\theta}^*})$ we obtain:

$$\lim_{N\to\infty} \mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}) = 0.$$

Lemma 9. Assume the conditions of Proposition 1 hold. Then there exists a function ϕ, increasing in its first variable, such that

$$\|\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}[f_{\bar{w}}(x)]\| \leqslant \frac{\phi(\mathrm{KL}(q^N_{\boldsymbol{\theta}}, q^N_{\boldsymbol{\theta}_0}), \mathsf{X}, d_{\mathsf{Y}})}{\sqrt{N}}.$$

Proof. By Cauchy-Schwartz inequality, the first moment of the predictive distribution under the variational posterior can be upper bounded as:

$$\|\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}[f_{\bar{w}}(x)]\| = \frac{1}{N}\|\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}[\sum_{j=1}^N \sigma(\langle b_j, x\rangle)a_j]\| \leqslant \frac{1}{N}\sum_{j=1}^N |\mathbb{E}[\sigma(\langle b_j, x\rangle)]|\|\mu_{a_j}\|.$$

Since σ is Lipschitz, $|\sigma(x)| \leqslant C_0 + L|x|$ where $C_0 = |\sigma(0)|$. Hence,

$$|\mathbb{E}[\sigma(\langle b_j, x\rangle)]| \leqslant |C_0 + L\mathbb{E}[|\langle b_j, x\rangle|]| \leqslant C_0 + L\sum_{l=1}^{d_{\mathsf{X}}} \mathbb{E}[|b_{j,l}|]|x_l|]$$

Let's start by finding an upper bound for $\mathbb{E}[|b_{j,l}|]$. If $b_{j,l} \sim \mathcal{N}(\mu_{b_{j,l}}, \sigma^2_{b_{j,l}})$, then $|b_{j,l}|$ has an absolute Gaussian distribution and denoting Φ the CDF of a standard Gaussian, we have

$$\mathbb{E}[|b_{j,l}|] = \sigma_{b_{j,l}}\sqrt{\frac{2}{\pi}}\exp\left(\frac{-\mu^2_{b_{j,l}}}{2\sigma^2_{b_{j,l}}}\right) + \mu_{b_{j,l}}\left[1 - 2\Phi\left(-\frac{\mu_{b_{j,l}}}{\sigma_{b_{j,l}}}\right)\right] \leqslant \sigma_{b_{j,l}}\sqrt{\frac{2}{\pi}} + |\mu_{b_{j,l}}|.$$

Recall that the KL between the posterior $q^N_{\boldsymbol{\theta}}$ and prior $q^N_{\boldsymbol{\theta}_0}$ can be written:

$$\mathrm{KL}(q^N_{\boldsymbol{\theta}}|q^N_{\boldsymbol{\theta}_0}) = \frac{1}{2}\sum_{j=1}^N \left[\sum_{l=1}^{d_{\mathsf{X}}}(\mu^2_{b_{j,l}} + \sigma^2_{b_j,l} - \log(\sigma^2_{b_j,l}) - 1) + \sum_{l=1}^{d_{\mathsf{Y}}}(\mu^2_{a_{j,l}} + \sigma^2_{a_j,l} - \log(\sigma^2_{a_j,l}) - 1)\right] \tag{26}$$

Hence, for any $j = 1, \ldots, N$ and $l = 1, \ldots, d_{\mathsf{X}}$:

$$\begin{aligned} |\mu_{b_{j,l}}| \leq \|\mu\|_2 \leq \sqrt{2\mathrm{KL}(q^N_{\boldsymbol{\theta}}|q^N_{\boldsymbol{\theta}_0})}, \\ \sigma_{b_{j,l}} \leq \|\sigma\|_2 \leq |\sigma_{b_{j,l}} + 1 - 1| \leq |\sigma^2_{b_{j,l}} - \log(\sigma^2_{b_{j,l}}) - 1| + 1 \leq 2\mathrm{KL}(q^N_{\boldsymbol{\theta}}|q^N_{\boldsymbol{\theta}_0}) + 1, \end{aligned} \tag{27}$$

and

$$\mathbb{E}[|b_{j,l}|] \leqslant \sqrt{\frac{2}{\pi}}(2\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N)+1)+\sqrt{2\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N)} := D(\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N)) \tag{28}$$

Where D is increasing. Hence, since X is compact, there exists C_{X} such that $\|x\|_1 \leq C_{\mathsf{X}}$ and:

$$|\mathbb{E}[\sigma(\langle b_j, x\rangle)]| \leqslant C_0 + LC_{\mathsf{X}} D(\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N)) := E(\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N), \mathsf{X}), \tag{29}$$

Where E is increasing in its first variable. Finally, since

$$N^{-1}\sum_{j=1}^N \|\mu_{a_j}\|_2 \leq N^{-1}\|\mu_a\|_1 \leq N^{-1}\sqrt{Nd_{\mathsf{Y}}}\|\mu_a\|_2 \leq N^{-\frac{1}{2}}\sqrt{d_{\mathsf{Y}}}\sqrt{2\mathrm{KL}(q_{\boldsymbol{\theta}}^N, q_{\boldsymbol{\theta}_0}^N)},$$

the first moment of the predictive distribution can be upper bounded as:

$$\|\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}}[f_{\bar{w}}(x)\| \leqslant \frac{E(\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N), X)\sqrt{d_{\mathsf{Y}}}\sqrt{2\mathrm{KL}(q_{\boldsymbol{\theta}}^N, q_{\boldsymbol{\theta}_0}^N)}}{\sqrt{N}} := \frac{\phi(\mathrm{KL}(q_{\boldsymbol{\theta}}^N, q_{\boldsymbol{\theta}_0}^N), \mathsf{X}, d_{\mathsf{Y}})}{\sqrt{N}},$$

where ϕ is increasing in its first variable. □

Lemma 10. Assume the conditions of Proposition 1 hold. Then there exists a function ψ depending only on $\mathrm{KL}(q_{\boldsymbol{\theta}}^N, q_{\boldsymbol{\theta}_0}^N)$, X, and d_{Y} such that G, increasing in its first variable, such that:

$$\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}^N}[\|f_{\bar{w}}(x)\|^2] \leqslant \frac{\psi(\mathrm{KL}(q_{\boldsymbol{\theta}}^N, q_{\boldsymbol{\theta}_0}^N), \mathsf{X}, d_{\mathsf{Y}})}{N}.$$

Proof. For a posterior of the form (21), we can write the second moment of the predictive distribution as:

$$\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}^N}[\|f_w(x)\|^2] = \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}[\sigma_j^2]\mathbb{E}[\|a_j\|^2] + \frac{2}{N^2}\sum_{j=1}^N\sum_{k<j}^N \mathbb{E}[\sigma_j]\mathbb{E}[\sigma_k]\mathbb{E}[\langle a_j, a_k\rangle]. \tag{30}$$

We start with the second term on the right hand side of (30). Using $\mathbb{E}[\langle a_j, a_k\rangle] = \langle \mu_{a_j}, \mu_{a_k}\rangle \leq 1/2(\|\mu_{a_j}\|^2 + \|\mu_{a_k}\|^2)$, along with (27) and (29), we have

$$\frac{1}{N^2}\sum_{j=1}^N\sum_{k=1}^N \mathbb{E}[\sigma_j]\mathbb{E}[\sigma_k]\langle \mu_{a_j}, \mu_{a_k}\rangle \leqslant \frac{E^2(\mathrm{KL}(q_{\boldsymbol{\theta}}^N, q_{\boldsymbol{\theta}_0}^N))2\mathrm{KL}(q_{\boldsymbol{\theta}}^N, q_{\boldsymbol{\theta}_0}^N)}{N^2}.$$

We now turn to the first term on the right hand side of (30). We first have for any $j = 1, \ldots, N$, using (26) that:

$$\mathbb{E}[\|a_j\|^2] = \sum_{l=1}^{d_{\mathsf{Y}}} \mathbb{E}[a_{j,l}^2] = \sum_{l=1}^{d_{\mathsf{Y}}}(\sigma_{a_{j,l}}^2 + \mu_{a_{j,l}}^2) \leqslant 2\mathrm{KL}(q_{\boldsymbol{\theta}^*}^N, q_{\boldsymbol{\theta}_0}^N) + d_{\mathsf{Y}}(2\mathrm{KL}(q_{\boldsymbol{\theta}^*}^N, q_{\boldsymbol{\theta}_0}^N)+1)^2 := F(\mathrm{KL}(q_{\boldsymbol{\theta}^*}^N, q_{\boldsymbol{\theta}_0}^N)).$$

Then, using that σ is L-Lipschitz, Cauchy-Schwartz inequality and that since X is compact there exists c_{X} such that $\|x\| \leq c_{\mathsf{X}}$, we have:

$$\mathbb{E}[\sigma_j^2] \leq \mathbb{E}[(C_0 + L|<b_j, x>|)^2] = C_0^2 + 2C_0c_{\mathsf{X}}L\mathbb{E}[\|b_j\|] + L^2c_{\mathsf{X}}^2\mathbb{E}[\|b_j\|^2],$$

where, using (26) and (28),

$$\mathbb{E}[\|b_j\|] \leq \sum_{l=1}^{d_{\mathsf{X}}} \mathbb{E}[|b_{j,l}|] \leq d_{\mathsf{X}} D(\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N)),$$

$$\mathbb{E}[\|b_j\|^2] = \sum_{l=1}^{d_{\mathsf{X}}} \mathbb{E}[b_{j,l}^2] = \sum_{l=1}^{d_{\mathsf{X}}}(\sigma_{b_{j,l}}^2 + \mu_{b_{j,l}}^2) \leq d_{\mathsf{X}}(2\mathrm{KL}(q_{\boldsymbol{\theta}^*}^N, q_{\boldsymbol{\theta}_0}^N)+1)^2 + 2\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N).$$

Hence,

$$\mathbb{E}[\sigma_j^2] \leqslant C_0^2 + 2C_0c_{\mathsf{X}}Ld_{\mathsf{X}}D(\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N)) + L^2c_{\mathsf{X}}^2 2\mathrm{KL}(q_{\boldsymbol{\theta}}^N|q_{\boldsymbol{\theta}_0}^N) := G(\mathrm{KL}(q_{\boldsymbol{\theta}^*}^N, q_{\boldsymbol{\theta}_0}^N)),$$

with R increasing. Hence, the first term on the right hand side of (30) can be bounded as:

$$\frac{1}{N^2}\sum_{j=1}^{N}\mathbb{E}[\sigma_j^2]\mathbb{E}[\|a_j\|^2] \leqslant \frac{G(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}))F(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}))}{N}.$$

Finally, we obtain the desired result with:

$$\psi(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0})) := G(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}))F(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0})) + E^2(\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0}))\sqrt{2\mathrm{KL}(q^N_{\boldsymbol{\theta}^*}, q^N_{\boldsymbol{\theta}_0})}.$$

□

Lemma 11. Let l be the cross-entropy loss, and $q^N_{\boldsymbol{\theta}} \in \mathscr{F}_\Theta$ where $\mathscr{F}_\Theta$ is a family of Gaussians with diagonal covariance matrices, i.e. for any $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\theta} = (\mu, \sigma) \in \mathbb{R}^{Nd_\mathsf{X}} \times R^{Nd_\mathsf{Y}}$. Assume that each coordinate of $\boldsymbol{\theta}$ is bounded by a constant (independent of N) and that $\lim_{N\to\infty} \|\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}}[f_{\bar{w}}(x)]\| = 0$ for any $x \in \mathsf{X}$. Then,

$$\lim_{N\to\infty} \mathcal{L}(q^N_{\boldsymbol{\theta}}) = p(\log(d_\mathsf{Y}) + \log(Z)).$$

Proof. For any $i = 1, \ldots, p$, denote

$$\begin{array}{rccc} l_{y_i}: & \mathbb{R}^{d_\mathsf{Y}} & \longrightarrow & \mathbb{R} \\ & (z_1, \ldots, z_{d_\mathsf{Y}}) & \longmapsto & -\log\left(\frac{e^{z_{y_i}}}{\sum_{j=1}^{d_\mathsf{Y}} e^{z_j}}\right) \end{array},$$

so that $\forall z = (z_1, \ldots z_{d_\mathsf{Y}}) \in \mathbb{R}^{d_\mathsf{Y}}$,

$$|l_{y_i}(z)| = \left|-\log(\exp(z_{y_i}) + \log\left(\sum_{k=1}^{d_\mathsf{Y}}\exp(z_k)\right)\right|. \tag{31}$$

By the definition of $\mathcal{L}$ and plugging $-\log(d_\mathsf{Y}) - \log(Z)$ in (31), we have:

$$|\mathcal{L}(q^N_{\boldsymbol{\theta}}) - p(\log(d_\mathsf{Y}) + \log(Z))| \leqslant \sum_{i=1}^{p}|\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}[f_{\bar{w}}(x, y_i)]| + |\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}[\log\frac{1}{d_\mathsf{Y}}\sum_{k=1}^{d_\mathsf{Y}} e^{f_{\bar{w}}(x,k)}]|,$$

where $f_{\bar{w}}(x, k)$ denotes the k-th coordinate of $f_{\bar{w}}(x) \in \mathbb{R}^{d_\mathsf{Y}}$ for $l = 1, \ldots, d_\mathsf{Y}$. The first term on the right hand side of the previous inequality converges to 0 as N goes to infinity by assumption. Hence, we can focus on the second term. For any $k = 1, \ldots, d_\mathsf{Y}$, since σ is L-Lipschitz,

$$f_{\bar{w}}(x, k) = \frac{1}{N}\sum_{j=1}^{N}\sigma(< b_j, x >)a_{j,k} \leqslant \frac{1}{N}\sum_{j=1}^{N} C_0 a_{j,k} + \frac{L}{N}\sum_{j=1}^{N}\sum_{l=1}^{d_\mathsf{X}}|b_{j,l}||x_l|a_{j,k}.$$

Using the previous inequality along with Jensen's inequality, we have

$$\left|\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}\left[\log\left(\frac{1}{d_\mathsf{Y}}\sum_{k=1}^{d_\mathsf{Y}} e^{\frac{L}{N}\sum_{j=1}^{N}\sum_{l=1}^{d_\mathsf{X}}|b_{j,l}||x_l|a_{j,k}}\right)\right]\right| \leqslant \left|\log\frac{1}{d_\mathsf{Y}}\sum_{k=1}^{d_\mathsf{Y}}\prod_{j=1}^{N}\prod_{l=1}^{d_\mathsf{X}}\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}\left[e^{\frac{L|b_{j,l}||x_l|a_{j,k}}{N}}\right]\right|.$$

Since the posterior is of the form (21) we have for any index (j, k, l):

$$\begin{aligned}\mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}\left[e^{L\frac{|b_{j,l}||x_l|a_{j,k}}{N}}\right] &\leqslant \mathbb{E}_{\bar{w}\sim q^N_{\boldsymbol{\theta}}}\left[e^{L\frac{|b_{j,l}||x_l||a_{j,k}|}{N}}\right] \\ &= \mathbb{E}_{u\sim\mathcal{N}(0,1);v\sim\mathcal{N}(0,1)}\left[e^{\frac{L|\sigma_{b_{j,l}}u+\mu_{b_{j,l}}||x||\sigma_{a_{j,k}}v+\mu_{a_{j,k}}|}}{N}\right] \\ &\leq \mathbb{E}_{u\sim\mathcal{N}(0,1);v\sim\mathcal{N}(0,1)}\left[e^{\frac{|Cb+C||x||Ca+C|}{N}}\right],\end{aligned}$$

for some constant $C > 0$ since by assumption each coordinate of the variational parameter is bounded. By the dominated convergence theorem, when N goes to infinity we have:

$$\mathbb{E}_{u\sim\mathcal{N}(0,1);v\sim\mathcal{N}(0,1)}\left[e^{\frac{L|\sigma_{b_{j,l}}u+\mu_{b_{j,l}}||x|(\sigma_{a_{j,k}}v+\mu_{a_{j,k}})}{N}}\right] = 1 + o\left(\frac{1}{N}\right).$$

Hence,

$$\left|\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}^N}\left[\log\frac{1}{d_{\mathsf{Y}}}\sum_{k=1}^{d_{\mathsf{Y}}} e^{\frac{L}{N}\sum_{j=1}^{N}\sum_{l=1}^{d_{\mathsf{X}}}|b_{j,l}||x_l|a_{j,k}}\right]\right| \leqslant N d_{\mathsf{X}}\log\left(1+o\left(\frac{1}{N}\right)\right)$$

Similarly, we can prove that:

$$\lim_{N\to\infty}\left|\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}^N}\left[\log\frac{1}{d_{\mathsf{Y}}}\sum_{k=1}^{d_{\mathsf{Y}}} e^{\frac{\sigma(0)}{N}\sum_{j=1}^{N}a_{j,k}}\right]\right| = 0$$

Finally, we have:

$$\lim_{N\to\infty}|\mathcal{L}(q_{\boldsymbol{\theta}}^N) - p(\log(d_{\mathsf{Y}}) + \log(Z))| \leqslant \lim_{N\to\infty}\left|N d_{\mathsf{X}}\log\left(1+o\left(\frac{1}{N}\right)\right)\right| = 0.$$

□

B Proof of Proposition 7

We will first need the following technical result.

Lemma 12. Assume **A**1-(i) and **A**2. Then for any $x \in \mathsf{X}$, the function $\theta \mapsto \tilde{\phi}(\theta, x)$ is continuous. In addition, there exists $C \geqslant 0$ such that for any $x \in \mathsf{X}$ and $\theta \in \Xi$, $\|\tilde{\phi}(\theta, x)\| \leqslant C$.

Proof. Since $\phi(\theta, z, x) = ah(b, x)$ and since by **A**2, $b \mapsto h(b, x)$ is continuous for any $x \in \mathsf{X}$, it follows that for any $x \in \mathsf{X}$, $z \in \mathbb{R}^d$, $\theta \mapsto \phi(\theta, z, x)$ is continuous on Ξ. Using (32) and the condition that Ξ is compact, an application of the Lebesgue dominated convergence theorem implies that for any $x \in \mathsf{X}$, the function $\theta \mapsto \tilde{\phi}(\theta, x)$ is continuous. Finally, Eq. (32) and the condition that Ξ is compact shows that there exists $C \geqslant 0$ such that for any $x \in \mathsf{X}$ and $\theta \in \Xi$, $\|\tilde{\phi}(\theta, x)\| \leqslant C$. □

We now prove Proposition 7. We first prove **A**1-(ii). Recall, that for $\theta, z, x \in \Xi \times \mathbb{R}^d \times \mathsf{X}$, $\phi(\theta, z, x) = s(\mathscr{T}_\theta(z), x)$ where $\mathscr{T}_\theta(z) = \mu + \sigma \odot z$. Therefore, by (3), decomposing each weight as $w = (a, b)$ where a is the output weight and b is the hidden weight, $\phi(\theta, z, x) = ah(b, x)$, with $a = \mu_a + \sigma_a \odot z_a$ and $b = \mu_b + \sigma_b \odot z_b$, $\theta = (\theta_a, \theta_b)$, $\theta_a = (\mu_a, \sigma_a) \in \mathbb{R}^{d_{\mathsf{Y}}} \times (\mathbb{R}_+^*)^{d_{\mathsf{Y}}}$ and $\theta_b = (\mu_b, \sigma_b) \in \mathbb{R}^{d_{\mathsf{X}}} \times (\mathbb{R}_+^*)^{d_{\mathsf{X}}}$. Hence,

$$\begin{aligned}\|a\|^2 &\leq 2\|\mu_a\|^2 + 2\|\sigma_a\|^2\|z_a\|^2 \leq 2\|\theta\|^2(1 + \|z_a\|^2)\,,\\ \|b\|^2 &\leqslant 2\|\theta\|^2(1 + \|z_b\|^2)\,.\end{aligned}$$

Also, by **A**2, there exist $C_0, C_1 \geqslant 0$ such that for any x, b, $|h(b, x)| \leq C_0 + C_1\|b\|$. Hence, we have for any $\theta \in \Xi$, $z \in \mathbb{R}^d$ and $x \in \mathsf{X}$,

$$\begin{aligned}&\|\phi(\theta, z, x)\|^2 \leq \|a\|^2(C_0 + C_1\|b\|)^2 \\ &\leq 2\left\|\theta\right\|^2(1 + \left\|z_a\right\|^2)[C_0 + 2C_1\left\|\theta\right\|(1 + \left\|z_b\right\|)^{\frac{1}{2}}]^2 \\ &\leq C_3(1 + \|z\|^4)(1 + \|\theta\|^2)\,,\end{aligned} \tag{32}$$

for some constant $C_3 > 0$. As Ξ is compact and $\int \|z\|^4 \mathrm{d}\gamma(z) < +\infty$, it follows that **A**1-(ii) holds. We now show that $\operatorname{argmax}_{\mathcal{P}(\Xi)} \tilde{\mathrm{F}}_\eta^N \neq \emptyset$. By Lemma 12 in the supplement, $\tilde{\phi}$ is bounded and for any $x \in \mathsf{X}$, $\theta \mapsto \tilde{\phi}(\theta, x)$ is continuous. Using that under **A**2 for any $y \in \mathsf{Y}$, $\tilde{y} \mapsto \ell(y, \tilde{y})$ is continuous, it follows that $\nu \mapsto \tilde{\mathrm{G}}(\nu; (x, y))$ is continuous for the weak topology on $\mathcal{P}(\Xi)$ for any $(x, y) \in \mathsf{X} \times \mathsf{Y}$. In addition, since $\theta \mapsto \mathrm{KL}(q_\theta^1|P_0)$ is continuous, we get since Ξ is compact that $\nu \mapsto \int \mathrm{KL}(q_\theta^1|P_0^1)\mathrm{d}\nu(\theta)$ is continuous for the weak topology. It follows that $\nu \mapsto \tilde{\mathrm{F}}_\eta^N(\nu)$ is continuous for the weak topology. Using Ξ is compact, $\mathcal{P}(\Xi)$ is compact for the weak topology by Ambrosio, Gigli, and Savaré 2008, Theorem 5.1.3, and it follows that $\operatorname{argmax}_{\mathcal{P}(\Xi)} \tilde{\mathrm{F}}_\eta^N \neq \emptyset$. The last condition (15) of Theorem 4 easily follows from Lemma 12.

C Proof of Proposition 5

We will prove the result by contradiction. Let $\boldsymbol{\theta}^{*,N} = \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta} \mathrm{ELBO}_\eta^N(\boldsymbol{\theta})$. Assume that $\mathrm{KL}(q_{\boldsymbol{\theta}^{*,N}}, P_0) \to 0$ as $N \to \infty$.

Recall that the prior writes as (20). Since the posterior $q_{\boldsymbol{\theta}^{*,N}} \in \mathscr{F}_\Theta$ writes:

$$q_{\boldsymbol{\theta}^{*,N}}(\bar{w}) = \prod_{j=1}^{N}\prod_{l=1}^{d_{\mathsf{Y}}} \mathcal{N}(a_{j,l}; \mu^*_{a_{j,l}}, \sigma^{*2}_{a_{j,l}}) \times \prod_{j=1}^{N}\prod_{r=1}^{d_{\mathsf{X}}} \mathcal{N}(b_{j,k}; \mu^*_{b_{j,r}}, \sigma^{*2}_{b_{j,r}}),$$

assuming $\mathrm{KL}(q_{\boldsymbol{\theta}^{*,N}}, P_0) \to 0$ as $N \to \infty$ is equivalent to assuming that for all $j = 1, \dots, N$, $l = 1, \dots, d_{\mathsf{Y}}$, $r = 1, \dots d_{\mathsf{X}}$,

$$\lim_{N\to+\infty} \mu^*_{a_{j,l}} = 0, \ , \lim_{N\to+\infty} \mu_{b^*_{j,r}} = 0, \ \lim_{N\to+\infty} \sigma^*_{a_{j,l}} = 1, \ \lim_{N\to+\infty} \sigma^*_{b_{j,r}} = 1. \tag{33}$$

We consider a small perturbation of the optimal variational distribution. More precisely, we consider the parameter $\tilde{\boldsymbol{\theta}}^{*,N}$ defined by:

- for all $j = 1, \dots, N$, $r = 1, \dots, d_{\mathsf{X}}$, $\tilde{\mu}_{b_{j,r}} = \mu^\star_{b_{j,r}}$
- for all $j = 1, \dots, N$, $r = 1, \dots, d_{\mathsf{X}}$, $\tilde{\sigma}_{b_{j,r}} = \sigma^\star_{b_{j,r}}$
- for all $j = 1, \dots, N$, $l = 1, \dots, d_{\mathsf{Y}}$, if $(j,l) \neq (k,m)$, $\tilde{\mu}_{a_{j,l}} = \mu^\star_{a_{j,l}}$
- for all $j = 1, \dots, N$, $l = 1, \dots, d_{\mathsf{Y}}$, $\sigma_{a_{j,l}} = \sigma^\star_{a_{j,l}}$
- $\tilde{\mu}_{a_{k,m}} = \mu^\star_{a_{k,m}} + \epsilon$

for some $\epsilon > 0$ and $k \in \{1, \dots, N\}, m \in \{1, \dots, d_{\mathsf{Y}}\}$.

Recall that

$$\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}) = -\mathcal{L}(q_{\boldsymbol{\theta}}) - \eta \mathrm{KL}(q_{\boldsymbol{\theta}} | q_{\boldsymbol{\theta}_0}), \quad \text{with } \mathcal{L}(q_{\boldsymbol{\theta}}) = -\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}}}\left[\sum_{i=1}^{p} \log(\mathrm{L}(y_i|x_i,\bar{w}))\right],$$

where $\mathrm{L}(y|x,\bar{w}) \propto \exp(-\ell(f_{\bar{w}}(x), y))$ is defined by (4). By the optimality of $\boldsymbol{\theta}^{*,N}$, we have $\mathrm{ELBO}_\eta^N(\boldsymbol{\theta}^{*,N}) \geqslant \mathrm{ELBO}_\eta^N(\tilde{\boldsymbol{\theta}}^{*,N})$, i.e.

$$\mathcal{L}(q_{\boldsymbol{\theta}^{*,N}}) + \eta \mathrm{KL}(q_{\boldsymbol{\theta}^{*,N}} | P_0) \leqslant \mathcal{L}(q_{\tilde{\boldsymbol{\theta}}^{*,N}}) + \eta \mathrm{KL}(q_{\tilde{\boldsymbol{\theta}}^{*,N}} | P_0)$$

which results in

$$\mathcal{L}(q_{\boldsymbol{\theta}^{*,N}}) - \mathcal{L}(q_{\tilde{\boldsymbol{\theta}}^{*,N}}) \leq \eta \left(\mathrm{KL}(q_{\tilde{\boldsymbol{\theta}}^{*,N}} | P_0) - \mathrm{KL}(q_{\boldsymbol{\theta}^{*,N}} | P_0)\right) = \frac{\eta}{2}\left(\epsilon \mu^*_{a_{k,m}} + \frac{\epsilon^2}{2}\right)$$

where the equality on the right-hand side follows from the construction of $\tilde{\boldsymbol{\theta}}^{*,N}$ w.r.t. $\boldsymbol{\theta}^{*,N}$ and the formula of KL between Gaussians (26).

Now, for the square loss, by (22) we have, denoting $\sigma_j = \sigma(\langle b_j, x_i \rangle)$ (we mask the dependence in the data index i for

lighter notations) for any $j = 1, \dots, N$:

$$\begin{aligned}
\mathcal{L}(q_{\boldsymbol{\theta}^{*,N}}) - \mathcal{L}(q_{\tilde{\boldsymbol{\theta}}^{*,N}}) &= \sum_{i=1}^{p} \mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}^{*,N}}}\left[\|f_{\bar{w}}(x_i)\|^2\right] - \mathbb{E}_{\bar{w}\sim q_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[\|f_{\bar{w}}(x_i)\|^2\right] \\
&\quad - 2\langle y_i, \mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}^{*,N}}}\left[f_{\bar{w}}(x_i)\right] - \mathbb{E}_{\bar{w}\sim q_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[f_{\bar{w}}(x_i)\right]\rangle \\
&= \frac{1}{2N}\sum_{i=1}^{p}\sum_{l=1}^{d_Y}\left(\mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}\left[\sum_{j=1}^{N}\sigma_j^2 a_{j,l}^2\right] - \mathbb{E}_{\theta\sim q_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[\sum_{j=1}^{N}\sigma_j^2 a_{j,l}^2\right]\right) \\
&\quad + \frac{1}{2N^2}\sum_{i=1}^{p}\sum_{l=1}^{d_Y}\sum_{j=1}^{N}\sum_{s=1,s\neq i}^{N}\left(\mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}\left[\sigma_j a_{j,l}\sigma_s a_{s,l}\right] - \mathbb{E}_{\theta\sim q_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[\sigma_j a_{j,l}\sigma_s a_{s,l}\right]\right) \\
&\quad + 2\sum_{i=1}^{p}\left\langle y_i, \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\bar{w}\sim q_{\tilde{\boldsymbol{\theta}}^{*,N}}}[\sigma_j]\mathbb{E}_{\bar{w}\sim q_{\tilde{\boldsymbol{\theta}}^{*,N}}}[a_j] - \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}^{*,N}}}[\sigma_j]\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}^{*,N}}}[a_j]\right\rangle \\
&:= A_N + B_N + C_N,
\end{aligned}$$

First, since the difference is null for $l \neq m$ and $j \neq k$, we have:

$$\begin{aligned}
A_N &= \frac{1}{2N}\sum_{i=1}^{p}\left(\mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}\left[\sigma_k^2 a_{k,m}^2\right] - \mathbb{E}_{\theta\sim q_{\tilde{\boldsymbol{\theta}}^{*,N}}}\left[\sigma_k^2 a_{k,m}^2\right]\right) \\
&= \frac{1}{2N}\sum_{i=1}^{p}\mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}\left[\sigma_k^2\right]\left(\mu^{\star 2}_{a_{k,m}} + \sigma^{\star 2}_{a_{k,m}} - (\mu^{\star}_{a_{k,m}} + \epsilon)^2 - \sigma^{\star 2}_{a_{k,m}}\right) \\
&= \frac{1}{2N}\sum_{i=1}^{p}\mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}\left[\sigma_k^2\right](-2\mu^{\star}_{a_{k,m}}\epsilon - \epsilon^2).
\end{aligned}$$

Then, let's define $\Delta_{j,s,l} = \mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}[\sigma_j a_{j,l}\sigma_s a_{s,l}] - \mathbb{E}_{\theta\sim q_{\tilde{\boldsymbol{\theta}}^{*,N}}}[\sigma_j a_{j,l}\sigma_s a_{s,l}])$. Firstly, if $l \neq m$, then $\Delta_{j,s,l} = 0$ for any $j, s \in \{1, \dots, N\}$. Now, if $l = m$, there are 3 different combinations for the indexes j and s:

- If $j \neq k$ and $s \neq k$, then $\Delta_{j,s,m}$ is also null.
- If $j \neq k$ and $s = k$ then $\Delta_{j,k,m} = \mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}[\sigma_j a_{j,m}\sigma_k](\mu^{\star}_{a_{k,m}} - (\mu^{\star}_{a_{k,m}} + \epsilon))$
- if $j = k$ and $s \neq k$ then $\Delta_{k,s,m} = \mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}[\sigma_k a_{s,m}\sigma_s](\mu^{\star}_{a_{k,m}} - (\mu^{\star}_{a_{k,m}} + \epsilon))$,

Hence

$$B_N = -\frac{1}{N^2}\sum_{i=1}^{p}\sum_{s=1,s\neq k}^{N}\mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^{*,N}}}[\sigma_k a_{s,m}\sigma_s]\epsilon.$$

And for the last term, we have:

$$C_N = \frac{2}{N}\sum_{i=1}^{p} y_{i,m}\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}^{*,N}}}[\sigma_k]\epsilon.$$

Since from the previous computations,

$$\frac{1}{2}\left(\epsilon\mu^{*}_{a_{k,m}} + \frac{\epsilon^2}{2}\right) \geq \frac{1}{\eta}\left(A_N + B_N + C_N\right), \tag{34}$$

we now study the limit of the right-hand side of the previous inequality under the tempering $\eta = \tau p/N$.

Recall that by assuming the collapse of the posterior onto the prior (33), since $\sigma_s = \sigma(\langle b_s, x_i\rangle)$, we have for any $s = 1, \dots, N$

$$m_i^{(1)} := \lim_{N\to\infty}\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}^{*,N}}}[\sigma_s] = \mathbb{E}_{b\sim\mathcal{N}(0,I_{d_X})}[\sigma(\langle b, x_i\rangle)], \quad m_i^{(2)} := \lim_{N\to\infty}\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}^{*,N}}}[\sigma_s^2] = \mathbb{E}_{b\sim\mathcal{N}(0,I_{d_X})}[\sigma^2(\langle b, x_i\rangle)].$$

Hence, we have by assumption (33), successively:

$$\lim_{N\to\infty}\frac{N}{\tau p}A_N=\lim_{N\to\infty}\frac{1}{2\tau p}\sum_{i=1}^{p}\mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^*,N}}\left[\sigma_k^2\right](-2\mu^\star_{a_{k,m}}\epsilon-\epsilon^2)=-\frac{\epsilon^2}{2\tau p}\sum_{i=1}^{p}m_i^{(2)},$$

Then, we obtain, using again assumption (33):

$$\lim_{N\to\infty}\frac{N}{\tau p}B_N=-\lim_{N\to\infty}\frac{1}{N\tau p}\sum_{i=1}^{p}\sum_{s=1,s\neq k}^{N}\mathbb{E}_{\theta\sim q_{\boldsymbol{\theta}^*,N}}[\sigma_k a_{s,m}\sigma_s]=-\lim_{N\to\infty}\frac{1}{N\tau p}\sum_{i=1}^{p}\sum_{s=1,s\neq k}^{N}(m_i^{(1)})^2\mu^*_{a_{s,m}}=0,$$

and finally:

$$\lim_{N\to\infty}\frac{N}{\tau p}C_N=\frac{2\epsilon}{\tau p}\sum_{i=1}^{p}y_{i,m}\lim_{N\to\infty}\mathbb{E}_{\bar{w}\sim q_{\boldsymbol{\theta}^*,N}}[\sigma_k]=\frac{2\epsilon}{\tau p}\sum_{i=1}^{p}y_{i,m}m_i^{(1)}.$$

Hence, considering (34) when $N\to\infty$, we have

$$\lim_{N\to\infty}\left(\epsilon\mu^*_{a_{k,m}}+\frac{\epsilon^2}{2}\right)=\frac{\epsilon^2}{2}\geq-\frac{\epsilon^2}{\tau p}\sum_{i=1}^{p}m_i^{(2)}+\frac{4\epsilon}{\tau p}\sum_{i=1}^{p}y_{i,m}m_i^{(1)},$$

reordering:

$$\epsilon^2\left(1+\frac{1}{\tau p}\sum_{i=1}^{p}m_i^{(2)}\right)\geq\frac{4\epsilon}{\tau p}\sum_{i=1}^{p}y_{i,m}m_i^{(1)}.$$

Now, let $\epsilon=\text{sign}(\sum_{i=1}^{p}y_{i,m}m_i^{(1)})\xi$ for some $\xi\in\mathbb{R}^{+*}$. The previous inequality writes:

$$\xi a:=\xi\left(1+\frac{1}{\tau p}\sum_{i=1}^{p}m_i^{(2)}\right)\geq\frac{4}{\tau p}\left|\sum_{i=1}^{p}y_{i,m}m_i^{(1)}\right|:=b.$$

Choosing $\xi<b/a$, we see that the previous inequality is false, contradicting the assumption (33).

D Tempering the ELBO objective when scaling the output of a neural network by $1/\sqrt{N}$

We now rewrite the ELBO (6) when the scaling in $1/\sqrt{N}$ is adopted; by rewriting both the KL term and likelihood term independently. We denote $\boldsymbol{\theta}^N=(\theta_1^N,\dots,\theta_N^N)$ the variational parameters of a neural network scaled by $1/N$ and $\boldsymbol{\theta}^{\sqrt{N}}=(\theta_1^{\sqrt{N}},\dots,\theta_N^{\sqrt{N}})$ if scaled by $1/\sqrt{N}$. Denoting $a'_j=a_j/\sqrt{N},b'_j=b_j$ for any $j=1,\dots,N$, the output of the neural network under (3) can be written

$$N^{-1}\sum_{j=1}^{N}a_j\sigma(\langle b_j,x\rangle):=N^{-1/2}\sum_{j=1}^{N}a'_j\sigma(\langle b'_j,x\rangle).$$

Hence, the mean and variance of input and output weights can be written respectively

$$(\mu_{b'_j},\sigma_{b'_j})=(\mu_{b_j},\sigma_{b_j}),\ (\mu_{a'_j},\sigma_{a'_j})=(\mu_{a_j}N^{-\frac{1}{2}},\sigma_{a_j}N^{-\frac{1}{2}}).$$

Recall as well that for any $j=1,\dots,N$, $\theta_j^N=(\mu_j,\sigma_j)\in\mathbb{R}^{2d}$, $\mu_j=(\mu_{b_j},\mu_{a_j})\in\mathbb{R}^{d_\mathsf{X}}\times\mathbb{R}^{d_\mathsf{Y}}$, $\sigma_j=(\sigma_{b_j},\sigma_{a_j})\in\mathbb{R}^{d_\mathsf{X}}\times\mathbb{R}^{d_\mathsf{Y}}$. Similarly, $\theta_j^{\sqrt{N}}$ correspond to the parameters b'_j,a'_j for $j=1,\dots,N$. Hence, using the formula for KL divergence between Gaussians, denoting $q_{\boldsymbol{\theta}}^{\sqrt{N}}$ the distribution over weights (a',b') we have

$$\begin{aligned}\text{KL}(q_{\boldsymbol{\theta}^{\sqrt{N}}}|P_0)=\frac{1}{2}\sum_{j=1}^{N}\Bigg[&\sum_{l=1}^{d_\mathsf{X}}(\mu_{b_{j,l}}^2+\sigma_{b_j,l}^2-\log(\sigma_{b_j,l}^2))\\&+\sum_{l=1}^{d_\mathsf{Y}}\left(\frac{\mu_{a_{j,l}}^2}{N}+\frac{\sigma_{a_j,l}^2}{N}-\log\left(\frac{\sigma_{a_j,l}^2}{N}\right)\right)-d_\mathsf{X}-d_\mathsf{Y}\Bigg]:=\alpha_N+\beta_N+C_N,\end{aligned}$$

where C_N is a constant (i.e. does not depend on the variational parameters) depending on N. We can notice than in contrast with the $1/N$ scaling, now the two terms (α_N, β_N) in the KL term, corresponding to the input and output weights, are unbalanced. We propose to rescale the latter as $\eta_N^1 \alpha_N + \beta_N$. We now turn to the data fitting term of the ELBO. Notice first that

$$\phi(\theta_j^{\sqrt{N}}, z, x_i) = a_j' h(b_j', x) = \frac{a_j}{\sqrt{N}} h(b_j, x) = \frac{\phi(\theta_j^N, z, x_i)}{\sqrt{N}}.$$

Hence we have for the likelihood term:

$$\begin{aligned}
\sum_{i=1}^{p} \int_{\mathbb{R}^{N\times d}} \log \mathrm{L}(y_i|x_i,\bar{w}) q_{\boldsymbol{\theta}^{\sqrt{N}}}(\bar{w}) \mathrm{dLeb}_{d\times N}(\bar{w}) &= \sum_{i=1}^{p} \int \ell\left(y, \sum_{j=1}^{N} \frac{\phi(\theta_j^{\sqrt{N}}, z_j, x)}{N}\right) \gamma^{\otimes N}(\mathrm{d}\boldsymbol{z}) \ , \\
&= \sum_{i=1}^{p} \int \ell\left(y, \sum_{j=1}^{N} \frac{\phi(\theta_j^{N}, z_j, x)}{N}\right) \gamma^{\otimes N}(\mathrm{d}\boldsymbol{z}) \\
&= \sum_{i=1}^{p} \mathrm{G}_\Theta^N(\boldsymbol{\theta}^N; (x,y)).
\end{aligned}$$

In the end, we define:

$$\mathrm{ELBO}_{\eta_N^1}^N(\boldsymbol{\theta}^{\sqrt{N}}) = -\eta_N^1 \alpha_N - \beta_N - \sum_{i=1}^{p} \mathrm{G}_\Theta^N(\boldsymbol{\theta}^{\sqrt{N}}; (x_i, y_i)),$$

which is balanced by choosing $\eta_N^1 = \frac{\tau p}{N}$.

E Additional Experiments

E.1 About the posterior variance collapse in the mean field regime

Here we discuss how the variance of the variational posterior behaves when optimizing the balanced ELBO as proposed in Section 3. For the square loss and since the distribution of each neuron is independent, the data-fitting term writes:

$$\begin{aligned}
\mathrm{G}_\Theta^N(\boldsymbol{\theta}; (x,y)) &= C - \frac{2y}{N} \mathbb{E}_{z\sim\gamma^{\otimes N}} \left[\sum_{j=1}^{N} a_j \sigma(\langle b_j, x\rangle)\right] + \frac{1}{N^2} \mathbb{E}_{z\sim\gamma^{\otimes N}} \left[\sum_{j=1}^{N} \|a_j\|^2 \sigma^2(\langle b_j, x\rangle)\right] \\
= C - \frac{2y}{N} \mathbb{E} &\left[\sum_{j=1}^{N} (\mu_{a_j} + \sigma_{a_j} \odot z_{a_j}) \sigma(\langle \mu_{b_j} + \sigma_{b_j} \odot z_{b_j}, x\rangle)\right] + \frac{1}{N^2} \mathbb{E} \left[\sum_{j=1}^{N} \|\mu_{a_j} + \sigma_{a_j} \odot z_{a_j}\|^2 \sigma^2(\langle \mu_{b_j} + \sigma_{b_j} \odot z_{b_j}, x\rangle)\right]
\end{aligned}$$

where the z_{a_j}, z_{b_j} are independent for any $j = 1, \ldots, N$. Consider σ odd, e.g. σ is the identity function. Optimizing over the variance for the variational posterior σ, we have that the σ minimizing $\mathrm{G}_\Theta^N(\boldsymbol{\theta}; (x,y))$ minimizes the last term above, i.e. $N^{-1}\sigma^2$. Hence, adding the KL to a standard Gaussian prior, the variational posterior variance σ minimizing the (rescaled) ELBO minimizes:

$$\frac{\sigma^2}{N} + (\sigma^2 - \log(\sigma^2)).$$

As $N \to \infty$, the first term becomes negligible, and the optimal variance σ collapses to the one of the prior. Moreover, we always have that as $N \to \infty$, the variance of the prior collapses to 0. Indeed, when choosing standard Gaussians as priors, the prior variance is equal to $\frac{1}{N^2}\sum_{j=1}^{N} 1 = 1/N \to 0$. However, when σ is non odd (e.g. RelU), the latter phenomenon - the posterior variance collapse onto zero - does not happen, as shown in the following experiment.

We consider a toy regression $\{(x_i, y_i)\}_{i=1}^p$ where $x_i \sim U[-1,1]$, $y_i = x_i^3 + \epsilon$ and $\epsilon \sim \mathcal{N}(0, 0.001)$. The following dataset is represented in figure 4. The models considered are one hidden layer Bayesian Neural Network with N number of neurons and a non odd activation (ReLU) trained by variational inference with a tempered ELBO ($\eta_N = \tau p/N$). We consider 9 possible values for N ($10^1, 2\times10^1, 10^2, 2\times10^2, 10^3, 2\times10^3, 10^4, 2\times10^4, 10^5$). And we study the dynamic, with respect to N, of the predictive distribution variance, ie, $\mathbb{V}_{w\sim q_\theta^N}[f_w(x)]$. The figure 4 represents the expected predictive distribution variance, ie, $\mathbb{E}_{x\sim U[-1,1]}[\mathbb{V}_{w\sim q_\theta}[f_w(x)]]$ approximated with Monte Carlo after the training. We cannot see any pattern

demonstrating that the variance of the BNN's output tends to zero as N grows. Hence, when the ELBO is well tempered and the activation is non odd then the predictive distribution variance over the prior collapse to zero, but not necessary the one over the variational posterior .

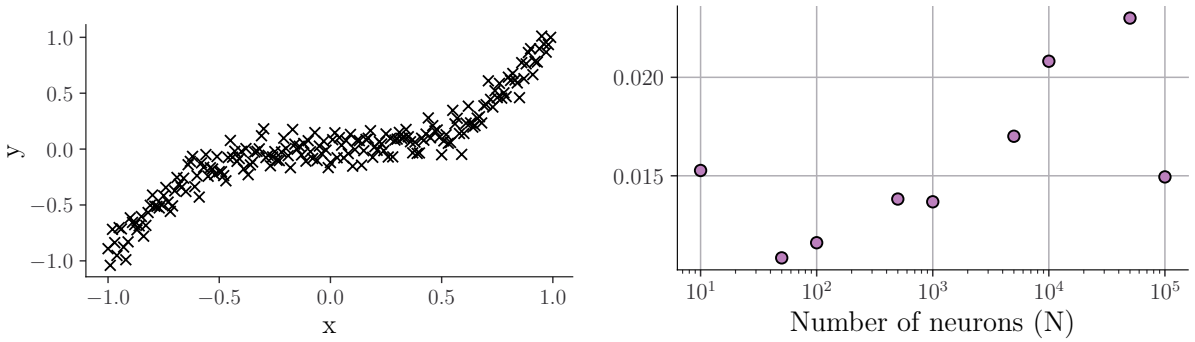

Figure 4: Expected standard deviation of the prediction for different sizes of model

E.2 Balanced ELBO with cooling

We first support with a very simple experiment the theoretical results of Section 3 and the relevance of the form of the parameter $\eta = \tau p/N$ we find. This experiment does not require training, since the goal here is to illustrate how introducing this parameter allows to balance the contributions of the two terms in the decomposition of ELBO_η^N in (6). We choose the architecture of a one hidden layer neural network with RelU activation functions, to which we will refer to as *Linear BNN*. We consider a regression task on the Boston dataset and a classification task on MNIST. We choose a zero-mean Gaussian prior with variance $1/5$ for each neuron. Also, we initialize the variational parameters $\theta = (\mu, \sigma)$ where μ is close to zero and $\sigma = 10^{-3}$. Figure 5 and 6 illustrate the ratio between the likelihood and KL terms in ELBO_η^N when the number of weights grows, for $\eta = 1$ (no cooling), $\eta = \tau p/N$ and different values of the hyperparameter τ, on the MNIST and BOSTON datasets respectively. They confirm that when the number of data points p is fixed and ELBO_η^N is not rescaled, one of the two terms become dominant contrary to the case where we set $\eta = \tau p/N$.

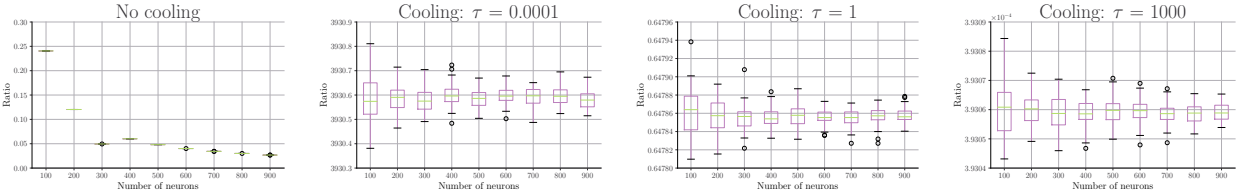

Figure 5: Ratio of the two ELBO^N terms, for a Linear BNN (non trained) on MNIST.

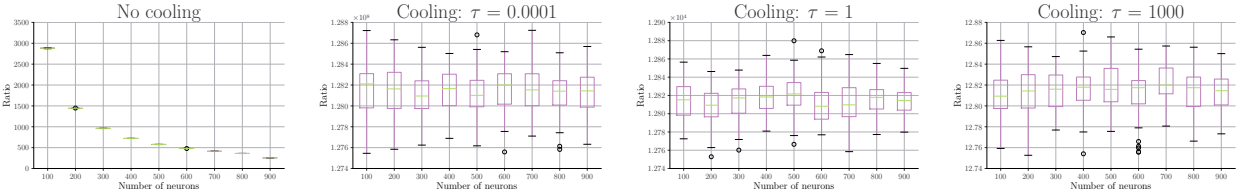

Figure 6: Ratio of the two ELBO^N terms, for a Linear BNN (non trained) on BOSTON.

E.3 ECE definition

For any input x, define $\text{conf}(x) = \max_{c \in \{1,\dots,n_l\}} \Psi_c(f_{\bar{w}}(x))$, i.e., the maximal predicted probability of the network. This quantity can be viewed as a prediction confidence for the input x. ECE discretizes the interval $[0, 1]$ into a given number of

bins B and groups predictions based on the confidence score: $S_b = \{i \in \{1, \dots, p\}, \text{conf}(x_i) \in [b/B, (1+b)/B[\}$. The calibration error is the difference between the fraction of predictions in the bin that are correct (accuracy) and the mean of the probabilities in the bin (confidence).

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{p} |\text{acc}(S_b) - \text{conf}(S_b)| \ ,$$

where p is the total number of data points, and $|S_b|$, $\text{acc}(S_b)$ and $\text{conf}(S_b)$ are the number of predictions, the accuracy and confidence of bin S_b respectively.

E.4 Cooling effect on the distribution of the variational parameters

Figure 7 illustrates the distribution of the variational parameters after training a linear BNN (i.e., single hidden layer with RelU) on MNIST. For a large τ, the distribution of the variational parameters is close to the prior (a centered Gaussian with standard deviation 0.2). For a small τ, we can see that the network has learnt values of σ that are very different from the prior (e.g., close to zero). Intermediate values of τ interpolate between the two previous regimes

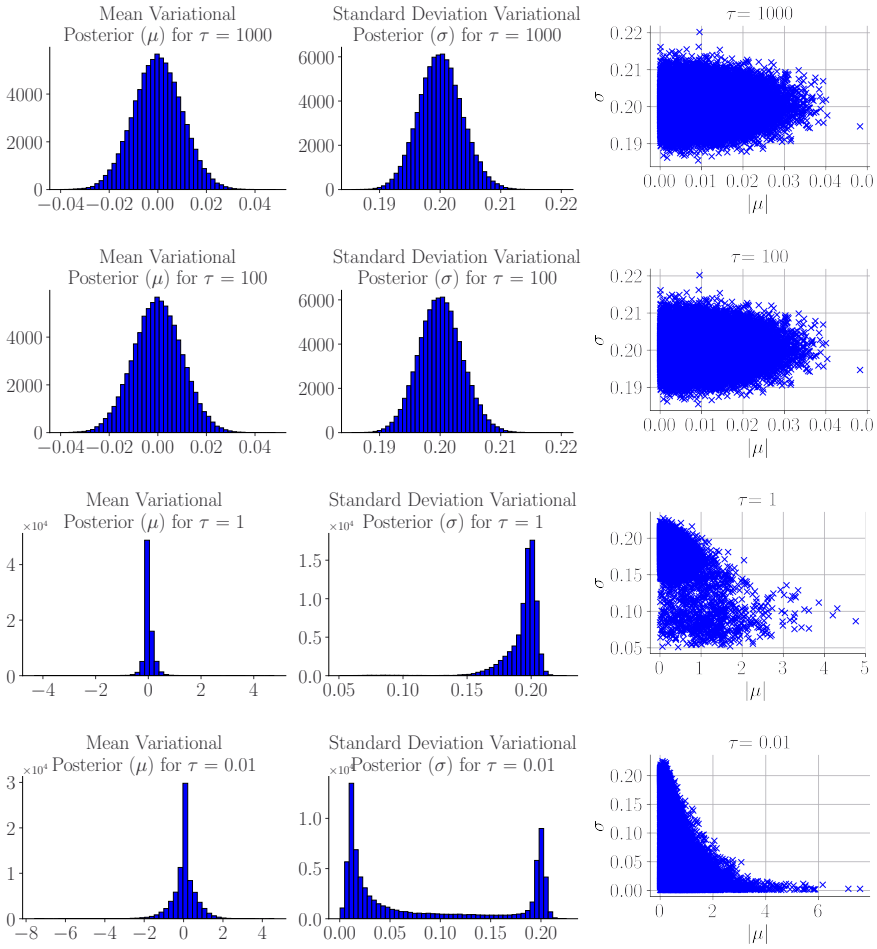

Figure 7: Histograms of the variational parameters $\theta = (\mu, \sigma)$ for a Linear BNN trained on MNIST. From left to right: histogram of variational means, standard deviations, and standard deviation as a function of the norm of the mean.

E.5 OOD detection

We also compare the performance on out-of-distribution of a Resnet20 trained on CIFAR-10 with Bayes by Backprop. We compute the the histogram of predictive entropies for 5000 in-distribution samples and out-of-distribution

samples. Recall that the negative entropy is defined for a vector of class probabilities $[p(y = c|x, \mathcal{D})]_{c\in\{1,\dots,n_l\}}$ as $-\sum_{c=1}^{n_l} p(y = c|x, \mathcal{D}) \log(p(y = c|x, \mathcal{D}))$. The first ones correspond to samples from the test set of CIFAR-10; while the out-of-distribution samples are chosen from another image dataset, namely SVHN Netzer et al. 2011. Our results are to be found in Figure 8 and illustrate again the importance of the parameter τ. When τ is very small, the model is highly confident for in-distribution samples, and has diffuse predictive entropies for out-distribution samples. As τ increases, the model starts to be less confident, resulting in higher entropies on both in-distribution and out-distribution samples, especially for the out-distribution samples. Finally if τ is too large, as the model sticks to the prior distribution, it is not confident neither on the in-distribution nor out-distribution, resulting on a spiky distribution of predictive entropies at high values.

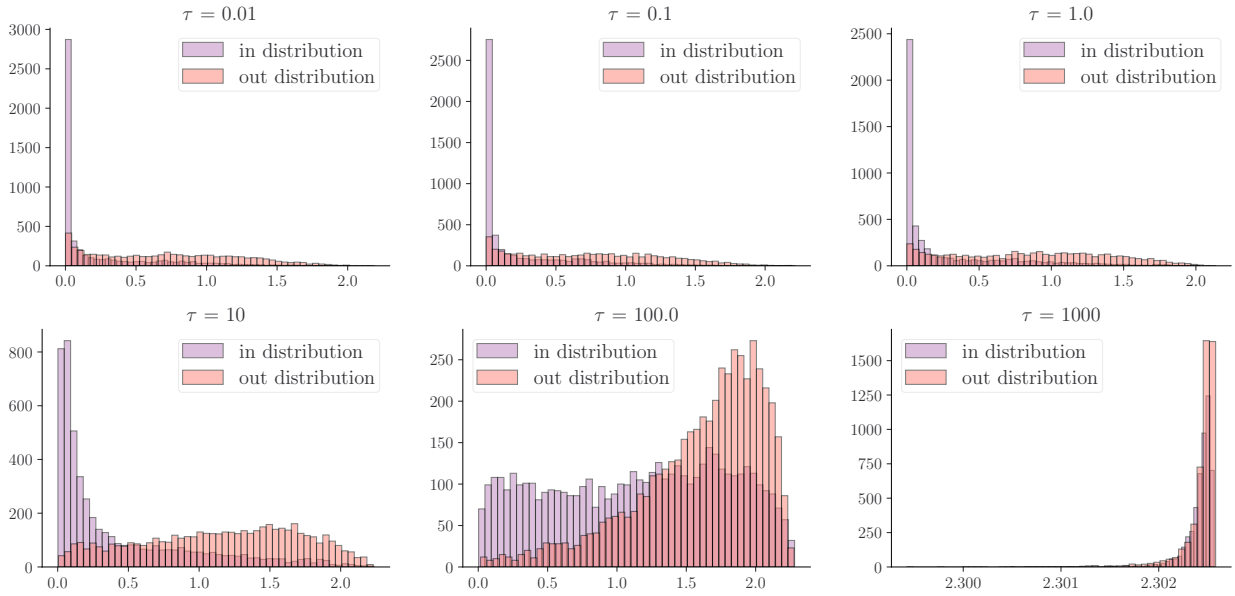

Figure 8: Histogram of the predictive entropies for a Resnet20 trained on CIFAR-10, on 5000 in-distribution (from the test set of CIFAR-10 dataset) and out-of-distribution (from SVHN dataset) samples

F Bayes by Backprop

Several methods have been proposed to optimize ELBO^N. A first and straightforward approach is to apply stochastic gradient descent (SGD), using samples from $q_{\boldsymbol{\theta}_c}$ where $\boldsymbol{\theta}_c$ is the current point, to obtain stochastic estimates for $\nabla_\theta \text{ELBO}^N$. However, the resulting estimation of the gradient suffers from high variance. Alternative algorithms have been proposed to mitigate this effect, such as Probabilistic Backpropagation Hernández-Lobato and Adams 2015 or Bayes by Backprop Blundell et al. 2015. Given a fixed distribution $\bar{\gamma}$ and a parameterized function $g(\boldsymbol{\theta}, \cdot)$, the network parameter $\bar{w}$ is obtained as $\bar{w} = g(\boldsymbol{\theta}, \boldsymbol{z})$, where $\boldsymbol{z}$ is sampled from $\bar{\gamma}$, e.g., from a standard normal distribution. While a new $\boldsymbol{z}$ is sampled at each iteration, its distribution is constant, unlike that of the network parameters $\bar{w}$. As soon as $g(\boldsymbol{\theta}, \cdot)$ is invertible and $\boldsymbol{\gamma}, q(\cdot|\boldsymbol{\theta})$ are non-degenerated probability distributions, we have $q(\bar{w}|\boldsymbol{\theta})d\bar{w} = \bar{\gamma}(\boldsymbol{z})d\boldsymbol{z}$ (see Jospin et al. 2020, Appendix A), and for any differentiable function f:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}_{\bar{w}\sim q(.|\boldsymbol{\theta})}[f(\bar{w}, \boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{z}\sim\bar{\gamma}} \left[\frac{\partial f(\bar{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\partial \bar{w}}{\partial \boldsymbol{\theta}} \frac{\partial f(\bar{w}, \boldsymbol{\theta})}{\partial \bar{w}} \right].$$

Bayes by Backprop uses the previous equality to estimate the gradient of F, because $F = \mathbb{E}_{\bar{w}\sim q(\cdot|\boldsymbol{\theta})}[f(\bar{w}, \boldsymbol{\theta})]$ with $f(\bar{w}, \boldsymbol{\theta}) = \log q(\bar{w}|\boldsymbol{\theta}) - \log P_0(\bar{w}) - \log P(\mathcal{D}|\bar{w})$. More specifically, it performs a stochastic gradient descent for F using a new sample $\boldsymbol{z}$ at each time step to estimate the gradient of F as the parameter $\boldsymbol{\theta}$ is updated. When the step size in this algorithm goes to zero, the Bayes by Backprop dynamics corresponds to a Wasserstein gradient flow of a particular functional defined on the space of probability distributions over $\boldsymbol{\theta}$, which we introduce in the next section.

As in Blundell et al. 2015, we will use a variance reparameterization; $\sigma = \log(1+\exp(\rho)) \in \mathbb{R}^+$ for $\rho \in \mathbb{R}$. Consequently, the variational parameter is given by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{N\times 2d}$ with $\theta_j = (\mu_j, \rho_j) \in \mathbb{R}^{2d}$. We denote by $g : \mathbb{R}^{2d} \times \mathbb{R}^d \to$

Algorithm 1 Bayes by Backprop

Input: step-size $\delta > 0$, number of iterations m_{iter}, number of samples $M_{samples}$.
for each m_{iter} iterations **do**
for each $m = 1, \ldots, M_{samples}$ **do**
1. Sample $\boldsymbol{z} \sim \gamma^{\otimes N}$
2. Let $\bar{w} = \boldsymbol{\mu} + \log(1 + \exp(\boldsymbol{\rho})) \circ \boldsymbol{z}$.

end for
3. Compute

$$g(\bar{w}, \boldsymbol{\theta}) \approx \frac{1}{M_{samples}} \sum_{m=1}^{M_{batch}} \log q(\bar{w}_i|\boldsymbol{\theta}) - \log P_0(\bar{w}_i) P(\mathcal{D}|\bar{w}_i)$$

5. Calculate the gradient with respect to the mean and standard deviation parameter ρ

$$\Delta_\mu = \frac{\partial g(w, \theta)}{\partial w} + \frac{\partial g(w, \theta)}{\partial \mu}$$
$$\Delta_\rho = \frac{\partial g(w, \theta)}{\partial w} \frac{\epsilon}{1 + \exp(\rho)} + \frac{\partial g(w, \theta)}{\partial \rho}$$

6. Update the variational parameters:

$$\mu \longleftarrow \mu - \delta \Delta_\mu$$
$$\rho \longleftarrow \rho - \delta \Delta_\rho$$

end for

$\mathbb{R}, (\theta, z) \mapsto \mu + \log(1 + \exp(\rho)) \odot z$, where $\odot$ denotes the entry-wise multiplication and γ denotes the standard normal distribution over $\mathbb{R}^d$. The Bayes-by-backprop algorithm in this setting is summarized in Algorithm 1.

This algorithm is well suited for minibatch optimisation, when the dataset $\mathcal{D}$ is split into a partition of L subsets (minibatches) $\mathcal{D}_1, \ldots, \mathcal{D}_L$. In this case Graves 2011 proposes to minimise a rescaled NELBO^N for each minibatch $\mathcal{D}_l$, $l = 1, \ldots, L$ as

$$\text{NELBO}_l^N = \frac{1}{L} \text{KL}(q_{\boldsymbol{\theta}}|P_0) - \mathbb{E}_{\bar{w} \sim q_{\boldsymbol{\theta}}}[\log P(\mathcal{D}_l|\bar{w})].$$