3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–8

Bootstrap Ensembles as Variational Inference

Anonymous Authors
Anonymous Institution

1. Introduction

Ensemble methods have a long history of successful use in machine learning to improve performance over individual models Dietterich (2000). Recently, there has been a surge of interest in ensemble methods for Deep Neural Networks (DNNs) (Lakshminarayanan et al., 2017; Osband et al., 2018). Contrary to early works on the topic, currents attempts use ensembles to characterize the uncertainty in predictions, rather than to improve performance (Hansen and Salamon, 1990). Although the result of this practice is not conceptually too different from having a distribution of predictive models as in Bayesian inference, and there are some notable theoretical contributions characterizing its properties (Newton and Raftery, 1994; Heskes, 1997; Efron, 2012), to the best of our knowledge, ensemble-based methods lack the principled mathematical framework of Bayesian statistics.

The aim of this paper is to gain some insights on ensemble methods by viewing these through the lenses of variational inference. Ensemble-based methods rely on a wide range of practices, which are difficult to place under a single unified framework, including repeated training via random initialization (Szegedy et al., 2015; Sutskever et al., 2014; Krizhevsky et al., 2012), random perturbation of the data (Lakshminarayanan et al., 2017), or more recently, random perturbation of the prior (Osband et al., 2018; Pearce et al., 2018). In this work, we focus on parametric bootstrap (Efron and Tibshirani, 1993), whereby many perturbed replicates of the original data are generated by introducing noise from a parametric distribution, and a new model is optimized for each perturbed version of the original loss function. Then, the ensemble of models, each member of which is represented by a "particle" in the parameter space, makes it possible to obtain a family of predictions on unseen data, which can be used to quantify uncertainty in a frequentist sense. We seek to investigate whether this ensemble has any connection with the ensemble of models obtained in Bayesian statistics when sampling parameters from their posterior distribution.

We interpret the particles as samples from an unknown distribution approximating the posterior over model parameters, and we derive conditions under which any optimization steps of the particles improves the quality of the approximation to the posterior. Remarkably, the conditions that we derive do not require assumptions on the distributional form assumed by the particles and they are purely *geometrical*, involving first and second derivative of the log-likelihood w.r.t. model parameters. We make use of variational arguments to show that, in the linear regression case with a Gaussian likelihood, any optimization steps of the particles associated with a perturbed replicate of the data yields an improvement of the KL divergence between the distribution of the particles and the posterior. Interestingly, *the conditions that we derive suggest that this is also the case when the Hessian trace of the model function w.r.t. the parameters is zero almost everywhere.* As a consequence, our result shows that applying parametric bootstrap on DNNs with ReLU activations yields an optimization of the particles which does not degrade the quality of the approximation.

© A. Authors.

2. Bootstrap Ensembles

In regression, observations y are assumed to be a realization of a latent function $f(\mathbf{x};\theta)$ corrupted by a noise term ϵ:

$$y = f(\mathbf{x};\theta) + \epsilon \tag{1}$$

Given n input-output training pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1 \dots n\}$, the objective is to estimate θ. In Bayesian inference, this problem is formulated as a transformation of a prior belief $p(\theta)$ into a posterior distribution by means of the likelihood function $p(\mathcal{D}|\theta)$. This is achieved by applying Bayes rule: $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\,p(\theta)}{p(\mathcal{D})}$, where the model evidence $p(\mathcal{D})$ denotes the probaility of data when model parameters are marginalized out.

In a strictly frequentist setting, one can obtain a *maximum a posteriori* estimate (MAP) by maximizing: $\theta_* = \arg\max_\theta \left[\log p(\mathcal{D}|\theta) + \log p(\theta)\right]$, where $\log p(\theta)$ is interpreted as a regularization term. In parametric bootstrap (Efron and Tibshirani, 1993), data replicates are created by sampling from a parametric distribution that is fitted to the data, typically by means of maximum likelihood (ML). Then, a different model is fit to each replicate, and the ensemble of trained models is used to calculate statistics of interest. In this work, we use the likelihood model as the resampling distribution, in order to reflect the assumptions of the Bayesian model.

Consider a vectorization of model parameters $\theta \in \mathbb{R}^m$. For most of this work, we assume a Gaussian model for the noise, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and a Gaussian prior over θ:

$$p(\mathcal{D}|\theta) = \prod_{\mathbf{x},y\in\mathcal{D}} \mathcal{N}(y; f(\mathbf{x};\theta), \sigma^2) \quad \text{and} \quad p(\theta) = \mathcal{N}(0, \alpha^2 I_m) \tag{2}$$

For each likelihood component, we have exactly one observation y, which is also the ML estimate for the mean parameter of a density with the same shape. In the bootstrap scheme, the label of each data-point is resampled as: $\tilde{y} \sim \mathcal{N}(y, \sigma^2)$, where $\tilde{y}$ denotes a perturbed version of the originally observed label y. We shall denote the entire dataset of perturbed labels as $\tilde{\mathcal{D}}$, such that $(\mathbf{x}, \tilde{y}) \in \tilde{\mathcal{D}}$.

In the Bayesian treatment of the model, variability is also encoded in the prior distribution. We shall capture this behavior by introducing a perturbation on the prior, so that each perturbed model is attracted to a different prior sample. Considering the prior of Equation (2), we create a perturbed version by resampling parameter components as follows:

$$p(\theta;\tilde{\theta}) = \mathcal{N}(\tilde{\theta}, \alpha^2 I_m), \quad \text{where} \quad \tilde{\theta} \sim \mathcal{N}(0, \alpha^2 I_m) \tag{3}$$

The perturbed distribution $p(\theta;\tilde{\theta})$ depends on $\tilde{\theta}$, which has been sampled from the original Gaussian prior. The combined resampling results in the following perturbed joint log-likelihood:

$$\log \tilde{p}(\mathcal{D}, \theta) = \log p(\tilde{\mathcal{D}}|\theta) + \log p(\theta;\tilde{\theta}) \tag{4}$$

Along these lines, we propose a gradient ascent scheme which operates on a set of particles $\{\theta^{(1)}, \dots, \theta^{(k)}\}$, where $\theta^{(i)} \in \mathbb{R}^m$ for $1 \le i \le k$. We effectively maximize different realizations of $\log \tilde{p}(\mathcal{D}, \theta)$; note that this process can be trivially parallelized. Each particle is then attracted to a different sample of the prior and a different perturbation of the data.

If q is the empirical distribution of the particles, we hope that q can serve as an approximation to the Bayesian posterior. The distribution q is implicitly initialized to the

prior, as we have $\theta^{(i)} \leftarrow \tilde{\theta} \sim p(\theta)$. This is a sensible choice, as any samples away from the support of the prior would have very low probability under the posterior. Notice that we make no other assumptions regarding the shape of q; we only know q implicitly through its samples.

3. Analysis of Perturbed Gradients

We shall investigate the effect of the optimization of particles on the implicit distribution q. Let $\mathrm{KL}[q||p]$ be the divergence between the approximating distribution $q(\theta)$ and the posterior $p(\theta|\mathcal{D})$. The update on an individual particle is described by the transformation:

$$\tau(\theta) = \theta + h\nabla \log \tilde{p}(\mathcal{D}, \theta) \tag{5}$$

Assume that $\theta \sim q$; then the transformation τ induces a change in q so that $\tau(\theta) \sim q_\tau$. It is desirable that the updated distribution q_τ is closer to the true posterior. Thus, the derivative of the KL divergence along the direction induced by τ has to be negative. Then for a gradient step h small enough, τ should decrease the KL divergence and the following difference should be negative:

$$\delta_h = \mathrm{KL}[q_\tau||p] - \mathrm{KL}[q||p]$$

Because the approximating distribution q is arbitrary, we shall take advantage of the fact that the KL divergence remains invariant under parameter transformations (Amari and Nagaoka, 2000). By applying the inverse transformation τ^{-1}, we have the following equivalent difference:

$$\delta_h = \mathrm{KL}[q||p_{\tau^{-1}}] - \mathrm{KL}[q||p] = \mathrm{E}_q[\log p(\theta|\mathcal{D}) - \log p_{\tau^{-1}}(\theta|\mathcal{D})] \tag{6}$$

where $p_{\tau^{-1}}$ denotes the transformed posterior density by τ^{-1}, which can be expanded as follows (Bishop, 2006):

$$p_{\tau^{-1}}(\theta|\mathcal{D}) = p(\tau(\theta)|\mathcal{D}) \det\{I_m + h \operatorname{Hess} \log \tilde{p}(\mathcal{D}, \theta)\} \tag{7}$$

After substituting $p_{\tau^{-1}}$ in (6), we can calculate the directional derivative of the KL along the direction of τ by considering the following limit:

$$\lim_{h \to 0} \frac{\delta_h}{h} = -\mathrm{E}_q[\nabla \log p^\top \nabla \log \tilde{p} + \operatorname{tr}\{\operatorname{Hess} \log \tilde{p}\}] = \nabla_\tau \mathrm{KL}[q||p] \tag{8}$$

where:

$$\nabla \log p^\top \nabla \log \tilde{p} = \lim_{h \to 0} \frac{\log p(\mathcal{D}, \theta + h\nabla \log \tilde{p}) - \log p}{h}$$
$$\operatorname{tr}\{\operatorname{Hess} \log \tilde{p}\} = \lim_{h \to 0} \frac{\log \det\{I_m + h \operatorname{Hess} \log \tilde{p}\}}{h}$$

To keep notation concise, we refer to the joint log-densities $\log p(\mathcal{D}, \theta)$ and $\log \tilde{p}(\mathcal{D}, \theta)$ simply as $\log p$ and $\log \tilde{p}$, respectively. The first of the two limits above is the directional derivative towards the gradient $\nabla \log \tilde{p}$. In a gradient ascent scheme, it is expected to have a positive value which gradually approaches zero over the course of optimization.

Ideally, the directional derivative in (8) should stay negative (or zero) as $\log \tilde{p}(\mathcal{D}, \theta)$ is maximized. The conditions under which this is true are reflected in the following proposition:

Proposition 1 *Let $\log \tilde{p}(\mathcal{D}, \theta)$ be a perturbed Bayesian model, and q an arbitrary distribution that approximates the true posterior $p(\theta|\mathcal{D})$. The transformation $\tau(\theta) = \theta + h\nabla \log \tilde{p}(\mathcal{D}, \theta)$ induces a change of measure such that the directional derivative $\nabla_\tau \mathrm{KL}[q||p]$ is non-positive if:*

$$\mathrm{E}_q[\nabla \log p^\top \nabla \log \tilde{p}] \geq -\mathrm{E}_q[\mathrm{tr}\{\mathrm{Hess} \log \tilde{p}\}] \tag{9}$$

Proof The result is given by a simple manipulation of (8), and by $\nabla_\tau \mathrm{KL}[q||p] < 0$. ■

The inequality in (9) is not always satisfied, as the Hessian can contain negative numbers in its diagonal; e.g., the second derivatives for θ near local maxima should be negative.

As an example where the inequality in (9) is violated, consider a convex *unperturbed* joint log-likelihood, i.e. different particles $\theta \sim q$ optimize $\log p(\mathcal{D}, \theta)$. Eventually, the different gradients $\nabla \log p$ would approach zero for any θ. The directional derivative expectation would also approach zero, as all points converge to the same maximum. The directional derivative of the KL divergence would tend to be positive, implying that further application of the transformation τ results in poorer approximation of the true posterior.

In the general case, it is rather difficult to reason precisely about the value of $\nabla_\tau \mathrm{KL}[q||p]$. Nevertheless, we conjecture that the introduction of a perturbation makes the inequality in (9) less likely to be violated. We demonstrate this effect for certain kinds of prior and likelihood in the following section.

Gradient analysis for Gaussian prior and likelihood

Let $f(\mathbf{x}; \theta)$ be the output of a nonlinear model (i.e. a DNN) and let $\theta \in \mathbb{R}^m$ be a vectorization of its parameters including weight and bias terms. We shall consider a Gaussian prior $\mathcal{N}(0, \alpha^2 I_m)$ and a likelihood function of the form $\mathcal{N}(f(\mathbf{x}; \theta), \sigma^2)$.

Let $\mathcal{N}(\tilde{\theta}, \alpha^2 I_m)$ denote a perturbed version of the prior, where $\tilde{\theta} \sim \mathcal{N}(0, \alpha^2 I_m)$. Let the perturbed version of the data be $\tilde{\mathcal{D}}$, where for all $(\mathbf{x}, y) \in \mathcal{D}$ and $(\mathbf{x}, \tilde{y}) \in \tilde{\mathcal{D}}$ we have $\tilde{y} = y + \tilde{y}_0$ and $\tilde{y}_0 \sim \mathcal{N}(0, \sigma^2)$. The perturbed version of the log-likelihood is:

$$\log \tilde{p}(\mathcal{D}, \theta) = -\sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{(f(\mathbf{x}; \theta) - \tilde{y})^2}{2\sigma^2} - \sum_{j=1}^{m} \frac{(\theta_j - \tilde{\theta}_j)^2}{2\alpha^2} \tag{10}$$

For the gradient and the Hessian trace of the perturbed log-likelihood above, we have:

$$\nabla \log \tilde{p}(\mathcal{D}, \theta) = \nabla \log p(\mathcal{D}, \theta) + \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{\tilde{y}_0}{\sigma^2} \nabla f(\mathbf{x}; \theta) + \frac{\tilde{\theta}_j}{\alpha^2} \tag{11}$$

$$\mathrm{tr}\{\mathrm{Hess} \log \tilde{p}(\mathcal{D}, \theta)\} = \mathrm{tr}\{\mathrm{Hess} \log p(\mathcal{D}, \theta)\} + \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \left(\frac{\tilde{y}_0}{\sigma^2} \sum_{j=1}^{m} \partial_{\theta_j^2} f(\mathbf{x}; \theta) \right) \tag{12}$$

During the optimization process, each particle θ is associated with a particular random perturbation. We have not made any specific assumptions regarding the approximating distribution q, therefore the random variable $\theta \sim q$ and the perturbations $\tilde{y}_0$ and $\tilde{\theta}$ are mutually independent.

We leverage this mutual independence and we exploit certain properties of the Gaussian assumptions, in order to further develop Eq. (9) into the theorem that follows.

Theorem 2 *Let $\log \tilde{p}(\mathcal{D}, \theta)$ be a perturbed Bayesian nonlinear model with prior $\mathcal{N}(0, \alpha^2 I_m)$ and likelihood $\mathcal{N}(f(\mathbf{x}; \theta), \sigma^2)$, with perturbations $\tilde{y} \sim \mathcal{N}(f(\mathbf{x}; \theta), \sigma^2)$ and $\tilde{\theta} \sim \mathcal{N}(0, \alpha^2 I_m)$. Let q be an arbitrary distribution that approximates the true posterior $p(\theta|\mathcal{D})$. The transformation $\tau(\theta)$ induces a change of measure such that the directional derivative $\nabla_\tau \mathrm{KL}[q||p]$ is non-positive if:*

$$E_{q,\tilde{y}_0,\tilde{\theta}} \left[\|\nabla \log \tilde{p}\|_2^2 \right] \geq E_q \left[\sum_{\mathbf{x},y \in \mathcal{D}} \left(\frac{f(\mathbf{x}; \theta) - y}{\sigma^2} \sum_{j=1}^{m} \partial_{\theta_j^2} f(\mathbf{x}; \theta) \right) \right] \tag{13}$$

Proof We can calculate the expectation of the KL directional derivative w.r.t. $\tilde{y}_0$, $\tilde{\theta}$ by noticing that $\mathrm{E}_{\tilde{y}_0}[\tilde{y}_0] = 0$ and $\mathrm{E}_{\tilde{\theta}}[\tilde{\theta}] = 0$:

$$\mathrm{E}_{\tilde{y}_0,\tilde{\theta}} \left[\nabla_\tau \mathrm{KL}[q||p] \right] = -\mathrm{E}_q [\nabla \log p^\top \nabla \log p + \mathrm{tr}\{\mathrm{Hess} \log p\}] \tag{14}$$

We next express the gradient norm $\|\nabla \log p\|_2^2$ in terms of the expectation $\mathrm{E}_{\tilde{y}_0,\tilde{\theta}} \left[\|\nabla \log \tilde{p}\|_2^2 \right]$ (see Lemma 1 in Appendix A.3). The same quadratic terms appear in both the norm expectation and the Hessian, thus the inequality simplifies to Eq. (13). See details in Section A.3 of the supplement. ■

Theorem 2 involves the expectation of the perturbed gradient norm. This should be a positive value that approaches zero, as the optimization converges. The theorem demands that this positive value is larger in expectation than a summation involving the diagonal second derivatives of $f(\mathbf{x}; \theta)$.

The fraction term in r.h.s. of (13) may have large absolute values if the particles $\theta \sim q$ are far from the posterior, but so will the perturbed gradient norm. However, the difference $f(\mathbf{x}; \theta) - y$ is not evaluated in absolute value; if the data are reasonably approximated, any discrepancies are averaged out. Nevertheless, it is rather difficult to reason about the magnitude of this difference in the general case.

Remarks on linear and piecewise-linear models The second-order derivative term in r.h.s. of (13) represents the curvature of the learned regression model. This term can be further simplified for certain families of functions. For a linear model, f is linear w.r.t. the parameters; this means that the directional derivative of KL divergence is *guaranteed* to be non-positive. This result can be extended to functions that are only piecewise-linear w.r.t. their parameters. A popular DNN design choice in the recent years involves ReLU activations, which are known to produce piecewise-linear models. It can be easily seen that the Hessian of a ReLU network is defined almost everywhere and its diagonal contains zeros. The set for which the Hessian is not defined has zero measure; for the same set the gradient is not defined either, but this has little effect on the usability of ReLU models.

As a final remark for models for which $\partial_{\theta_j^2} f(\mathbf{x}; \theta) = 0$, the KL divergence would only decrease over the course of optimization, until its directional derivative finally becomes zero. That does not guarantee that the KL divergence is optimized, as the derivative would have to be zero towards *any* direction.

Example – Regression ReLU network We consider a 8-layer DNN with 50 ReLU nodes per layer, featuring prior $\theta \sim \mathcal{N}(0, I_m)$, where $m = 18000$, and likelihood $y|\theta \sim \mathcal{N}(f(x), 0.1)$.

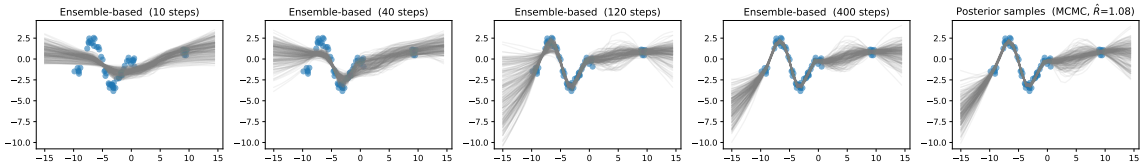

Figure 1: Regression on a 8-layer DNN with 50 ReLU nodes – State of 200 ensemble-based particles at different optimization stages.

As a point of reference for this example, we use samples of the Metropolis-Hastings algorithm featuring a Gaussian proposal with variance 0.01. We generated 200 samples by performing 10 restarts and having kept one sample every 20000 steps, after discarding the first 40000 samples. As a convergence diagnostic, we report the average $\hat{R}$ statistic (Gelman and Rubin, 1992) on the predictive models; the current choice of parameters has resulted in average $\hat{R} = 1.08$, which is a strong indication of convergence.

Figure 1 shows the particles given by the ensemble scheme at different stages of the optimization. As a reference, we use samples of the Metropolis-Hastings algorithm. As the optimization progresses, the distribution of predictive models improves until it reasonably approximates the MCMC result.

4. Conclusions

We have employed variational arguments to establish a connection between a certain kind of ensemble learning and Bayesian inference beyond linear regression with a Gaussian likelihood, for which this connection was already known. In particular, we interpreted the particles associated to perturbed versions of the joint log-likelihood as samples from a distribution approximating the posterior over model parameters. We then derived conditions under which any optimization steps of these particles yields an improvement of the divergence between the approximate and the actual posterior. Remarkably, the conditions do not require any particular form for the approximation and they are purely geometrical, involving first and second derivatives of the modeled function with respect to model parameters, giving insights on the behavior of ensemble learning on a number of interesting models. In particular, we showed how these results can be applied to DNNs with ReLU activations to establish that the optimization of the particles is guaranteed to have no detrimental contribution to the KL divergence between the approximate and the actual posterior.

We believe that this work makes a significant step in broadening the scope of ensemble methods for quantification of uncertainty. However, we are considering further extensions of our analysis to include more general classes of models, likelihoods, and priors, as well as investigations on how to use the criterion in Proposition 1 to derive novel particle-based variational inference methods.

References

S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical monographs*. Oxford University Press, 2000.

Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 1st ed. 2006. corr. 2nd printing 2011 edition, August 2006. ISBN 0387310738.

T. G. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems. First International Workshop, MCS2000, Cagliari, Italy*, pages 1–15. Springer-Verlag, 2000.

B. Efron and R Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

Bradley Efron. Bayesian inference and the parametric bootstrap. *Ann. Appl. Stat.*, 6(4): 1971–1997, 12 2012. doi: 10.1214/12-AOAS571.

A. Gelman and D.R. Rubin. A single series from the gibbs sampler provides a false sense of security. *Bayesian Statistics*, (4):625–631, 1992.

L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(10):993–1001, October 1990. ISSN 0162-8828. doi: 10.1109/34.58871.

Tom Heskes. Practical confidence and prediction intervals. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 176–182. MIT Press, 1997.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017.

Michael A. Newton and Adrian E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48, 1994.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems 31*, pages 8617–8629. Curran Associates, Inc., 2018.

Tim Pearce, Nicolas Anastassacos, Mohamed Zaki, and Andy Neely. Bayesian inference with anchored ensembles of neural networks, and application to reinforcement learning. In *ICML 2018: Workshop on Exploration in Reinforcement Learning*, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger,

editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–3

Supplementary Material: Bootstrap Ensembles as Variational Inference

Anonymous Authors
Anonymous Institution

Appendix A. Derivations and Proofs

Let $\log p(\mathcal{D}, \theta)$ be the joint log-likelihood of a model with likelihood $\mathcal{N}(f(\mathbf{x}; \theta), \sigma^2)$ and prior $\mathcal{N}(0, \alpha^2 I_m)$. We consider a perturbation of the prior $\tilde{\theta} \sim \mathcal{N}(0, \alpha^2 I_m)$, and the data $\tilde{\mathcal{D}}$ such that for all $(\mathbf{x}, y) \in \mathcal{D}$ and $(\mathbf{x}, \tilde{y}) \in \tilde{\mathcal{D}}$ we have $\tilde{y} = y + \tilde{y}_0$, where $\tilde{y}_0 \sim \mathcal{N}(0, \sigma^2)$. Then the perturbed version of the log-likelihood will be:

$$\log \tilde{p}(\mathcal{D}, \theta) = -\sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{(f(\mathbf{x}; \theta) - \tilde{y})^2}{2\sigma^2} - \sum_{j=1}^{m} \frac{(\theta_j - \tilde{\theta}_j)^2}{2\alpha^2} \tag{1}$$

A.1. Expectation of perturbed gradient

The components of the perturbed gradient will be:

$$\begin{aligned} \frac{\partial}{\partial \theta_j} \log \tilde{p}(\mathcal{D}, \theta) &= -\sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{f(\mathbf{x}; \theta) - \tilde{y}}{\sigma^2} \partial_{\theta_j} f(\mathbf{x}; \theta) - \frac{\theta_j - \tilde{\theta}_j}{\alpha^2} \\ &= \frac{\partial}{\partial \theta_j} \log p(\mathcal{D}, \theta) + \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{\tilde{y}_0}{\sigma^2} \partial_{\theta_j} f(\mathbf{x}; \theta) + \frac{\tilde{\theta}_j}{\alpha^2} \end{aligned} \tag{2}$$

It is easy to see that for the gradient of a perturbed log-likelihood, its expectation with respect to the perturbation will be equal to the unperturbed gradient:

$$\mathrm{E}_{\tilde{y}_0, \tilde{\theta}}[\nabla \log \tilde{p}(\mathcal{D}, \theta)] = \nabla \log p(\mathcal{D}, \theta) \tag{3}$$

A.2. Expectation of perturbed Hessian

For the trace we only need the diagonal components of the perturbed Hessian; by differentiating Equation (2) by θ_j we get:

$$\mathrm{tr}\{\mathrm{Hess} \log \tilde{p}(\mathcal{D}, \theta)\} = \sum_{j=1}^{m} \frac{\partial^2}{\partial \theta_j^2} \log p(\mathcal{D}, \theta) + \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \left(\frac{\tilde{y}_0}{\sigma^2} \sum_{j=1}^{m} \partial_{\theta_j^2} f(\mathbf{x}; \theta) \right) \tag{4}$$

Since $\mathrm{E}_{\tilde{y}_0}[\tilde{y}_0] = 0$, we have:

$$\mathrm{E}_{\tilde{y}_0, \tilde{\theta}}[\mathrm{tr}\{\mathrm{Hess} \log \tilde{p}(\mathcal{D}, \theta)\}] = \mathrm{tr}\{\mathrm{Hess} \log p(\mathcal{D}, \theta)\} \tag{5}$$

© A. Authors.

A.3. Proof of Theorem 2

Before proving Theorem 2, we shall review the following lemma:

Lemma 1 *Let $\log \tilde{p}(\mathcal{D}, \theta)$ be a perturbed Bayesian non-linear model as in (1). For the perturbation distributions we assume: $\tilde{y}_0 \sim \mathcal{N}(0, \sigma^2)$ and $\tilde{\theta} \sim \mathcal{N}(0, \alpha I_m)$. Then, for arbitrary θ we have:*

$$\|\nabla \log p\|_2^2 = \mathrm{E}_{\tilde{y}_0, \tilde{\theta}} \left[\|\nabla \log \tilde{p}\|_2^2 \right] + \frac{1}{\sigma^2} \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \|\nabla f(\mathbf{x}; \theta)\|_2^2 + \frac{m}{\alpha^2} \tag{6}$$

Proof From (2), we have the following for the original gradient:

$$\nabla \log p(\mathcal{D}, \theta) = - \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{\tilde{y}_0}{\sigma^2} \nabla f(\mathbf{x}; \theta) - \frac{\tilde{\theta}}{\alpha^2} + \nabla \log \tilde{p}(\mathcal{D}, \theta) \tag{7}$$

We consider the following joint expectation with respect to $\tilde{y}_0$ and $\tilde{\theta}$:

$$\begin{aligned}
&\mathrm{E}_{\tilde{y}_0, \tilde{\theta}}[\nabla \log p(\mathcal{D}, \theta)^\top \nabla \log p(\mathcal{D}, \theta)] \\
&= \mathrm{E}_{\tilde{y}_0, \tilde{\theta}} \left[\left(- \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{\tilde{y}_0}{\sigma^2} \nabla f(\mathbf{x}; \theta) - \frac{\tilde{\theta}}{\alpha^2} + \nabla \log \tilde{p}(\mathcal{D}, \theta) \right)^\top \left(- \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{\tilde{y}_0}{\sigma^2} \nabla f(\mathbf{x}; \theta) - \frac{\tilde{\theta}}{\alpha^2} + \nabla \log \tilde{p}(\mathcal{D}, \theta) \right) \right] \\
&= \mathrm{E}_{\tilde{y}_0, \tilde{\theta}} \left[\sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \frac{\tilde{y}_0^2}{\sigma^4} \|\nabla f(\mathbf{x}; \theta)\|_2^2 + \sum_{j=1}^{m} \frac{\tilde{\theta}_j^2}{\alpha^4} + \|\nabla \log \tilde{p}\|_2^2 \right]
\end{aligned} \tag{8}$$

Note that the terms of the polynomial that we have omitted are zero in expectation, because we have $\mathrm{E}[\tilde{y}_0] = 0$ and $\mathrm{E}[\tilde{\theta}] = 0$. Also since we have $\mathrm{E}[\tilde{y}_0^2] = \sigma^2$ and $\mathrm{E}[\tilde{\theta}_j] = \alpha^2$, the expectation becomes:

$$\nabla \log p(\mathcal{D}, \theta)^\top \nabla \log p(\mathcal{D}, \theta) = \frac{1}{\sigma^2} \sum_{\mathbf{x}, \tilde{y} \in \tilde{\mathcal{D}}} \|\nabla f(\mathbf{x}; \theta)\|_2^2 + \frac{m}{\alpha^2} + \mathrm{E}_{\tilde{y}_0, \tilde{\theta}} \left[\|\nabla \log \tilde{p}\|_2^2 \right] \tag{9}$$

Now we can move to the main theorem. ■

Theorem 2 *Let $\log \tilde{p}(\mathcal{D}, \theta)$ be a perturbed Bayesian non-linear model with prior $\mathcal{N}(0, \alpha^2 I_m)$ and likelihood $\mathcal{N}(f(\mathbf{x}; \theta), \sigma^2)$, with perturbations $\tilde{y}_0 \sim \mathcal{N}(f(\mathbf{x}; \theta), \sigma^2)$ and $\tilde{\theta} \sim \mathcal{N}(0, \alpha^2 I_m)$. Let q be an arbitrary distribution that approximates the true posterior $p(\theta|\mathcal{D})$. The transformation $\tau(\theta)$ will induce a change of measure such that the directional derivative $\nabla_\tau KL[q||p]$ is non-positive if:*

$$E_{q, \tilde{y}_0, \tilde{\theta}} \left[\|\nabla \log \tilde{p}\|_2^2 \right] \geq E_q \left[\sum_{\mathbf{x}, y \in \mathcal{D}} \left(\frac{f(\mathbf{x}; \theta) - y}{\sigma^2} \sum_{j=1}^{m} \partial_{\theta_j^2} f(\mathbf{x}; \theta) \right) \right]$$

Proof

According to Theorem 1, the directional derivative $\nabla_\tau \mathrm{KL}[q||p]$ is non-positive if the following holds:

$$\mathrm{E}_q[\nabla \log p^\top \nabla \log \tilde{p}] \geq -\mathrm{E}_q[\mathrm{tr}\{\mathrm{Hess} \log \tilde{p}\}]$$

For a non-linear model with Gaussian prior and likelihood, the gradient and the Hessian trace of the perturbed log-likelihood $\log \tilde{p}$ will have expectations:

$$\mathrm{E}_{\tilde{y}_0,\tilde{\theta}}[\nabla \log \tilde{p}] = \nabla \log p \tag{10}$$

$$\mathrm{E}_{\tilde{y}_0,\tilde{\theta}}[\mathrm{tr}\{\mathrm{Hess} \log \tilde{p}\}] = \mathrm{tr}\{\mathrm{Hess} \log p\} \tag{11}$$

Derivations for the expectations above can be found in Sections A.1 and A.2 of the supplementary material. Also, we can expand $\nabla \log \tilde{p}$ in the following inner product using (2):

$$\nabla \log p^\top \nabla \log \tilde{p} = \nabla \log p^\top \left(\nabla \log p + \sum_{\mathbf{x},\tilde{y} \in \tilde{\mathcal{D}}} \frac{\tilde{y}_0}{\sigma^2} \nabla f(\mathbf{x};\theta) + \frac{\tilde{\theta}}{\alpha^2} \right)$$

$$\mathrm{E}_{\tilde{y}_0,\tilde{\theta}}[\nabla \log p^\top \nabla \log \tilde{p}] = \nabla \log p^\top \nabla \log p$$

If we consider the joint expectation with respect to $\theta \sim q$, $\tilde{y}_0$ and $\tilde{\theta}$, the condition specified by Theorem 1 can be approximated as follows:

$$\begin{aligned} \mathrm{E}_{q,\tilde{y}_0,\tilde{\theta}}[\nabla \log p^\top \nabla \log \tilde{p}] &\geq -\mathrm{E}_{q,\tilde{y}_0,\tilde{\theta}}[\mathrm{tr}\{\mathrm{Hess} \log \tilde{p}\}] \\ \mathrm{E}_q[\nabla \log p^\top \nabla \log p] &\geq -\mathrm{E}_q[\mathrm{tr}\{\mathrm{Hess} \log p\}] \end{aligned} \tag{12}$$

Finally, if we use Lemma 1 on Equation (12) and we expand the Hessian,we obtain:

$$\begin{aligned} &\mathrm{E}_q \left[\frac{1}{\sigma^2} \sum_{\mathbf{x},\tilde{y} \in \tilde{\mathcal{D}}} \|\nabla f(\mathbf{x};\theta)\|_2^2 + \frac{m}{\alpha^2} + \mathrm{E}_{\tilde{y}_0,\tilde{\theta}} \left[\|\nabla \log \tilde{p}\|_2^2 \right] \right] \\ &\geq -\mathrm{E}_q \left[-\frac{1}{\sigma^2} \sum_{\mathbf{x},y \in \mathcal{D}} \|\nabla f(\mathbf{x};\theta)\|_2^2 - \frac{m}{\alpha^2} - \sum_{\mathbf{x},y \in \mathcal{D}} \left(\frac{f(\mathbf{x};\theta) - y}{\sigma^2} \sum_{j=1}^{m} \partial_{\theta_j^2} f(\mathbf{x};\theta) \right) \right] \end{aligned}$$

which easily simplifies to the condition required for non-positive $\nabla_\tau \mathrm{KL}[q||p]$ in Theorem 2. ■