SAMPLING-BASED INFERENCE FOR LARGE LINEAR MODELS, WITH APPLICATION TO LINEARISED LAPLACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale linear models are ubiquitous throughout machine learning, with contemporary application as surrogate models for neural network uncertainty quantification; that is, the linearised Laplace method. Alas, the computational cost associated with Bayesian linear models constrains this method's application to small networks, small output spaces and small datasets. We address this limitation by introducing a scalable sample-based Bayesian inference method for conjugate Gaussian multi-output linear models, together with a matching method for hyperparameter (regularisation) selection. Furthermore, we use a classic feature normalisation method (the g-prior) to resolve a previously highlighted pathology of the linearised Laplace method. Together, these contributions allow us to perform linearised neural network inference with ResNet-18 on CIFAR100 (11M parameters, 100 output dimensions × 50k datapoints) and with a U-Net on a high-resolution tomographic reconstruction task (2M parameters, 251k output dimensions).

1 INTRODUCTION

The linearised Laplace method, originally introduced by Mackay (1992), has received renewed interest in the context of uncertainty quantification for modern neural networks (NN) (Khan et al., 2019; Immer et al., 2021b; Daxberger et al., 2021a). The method constructs a surrogate Gaussian linear model for the NN predictions, and uses the error bars of that linear model as estimates of the NN's uncertainty. However, the resulting linear model is very large; the design matrix is sized number of parameters by number of datapoints times number of output classes. Thus, both the primal (weight space) and dual (observation space) formulations of the problem are intractable. This restricts the method to small network or small data settings. Moreover, the method is sensitive to the choice of regularisation strength for the linear model (Immer et al., 2021a; Antorán et al., 2022c). Motivated by linearised Laplace, we study inference and hyperparameter selection in large linear models.

To scale inference and hyperparameter selection in Gaussian linear regression, we introduce a sample-based Expectation Maximisation (EM) algorithm. It interleaves E-steps, where we infer the model's posterior distribution over parameters given some choice of hyperparameters, and M-steps, where the hyperparameters are improved given the current posterior. Our contributions here are two-fold:

1. We improve posterior sampling for large-scale conjugate Gaussian models using a novel sample-then-optimize objective, which we use to approximate the E-step.
2. We introduce a method for hyperparameter selection that requires only access to posterior samples, and not the full posterior distribution. This forms our M-step.

Combined, these allow us to perform inference and hyperparameter selection by solving a series of quadratic optimisation problems using iterative optimisation, and thus avoiding an explicit cubic cost in any of the problem's properties. Our method readily extends to non-conjugate settings, such as classification problems, through the use of the Laplace approximation. In the context of linearised NNs, our approach also differs from previous work in that it avoids instantiating the full NN Jacobian matrix, an operation requiring as many backward passes as output dimensions in the network.

We demonstrate the strength of our inference technique in the context of the linearised Laplace procedure for CIFAR100 (100 classes × 50k datapoints) image classification using an 11M parameter ResNet-18, and on a high-resolution (251k pixel) tomographic reconstruction (regression) task with

a 2M parameter U-Net. In tackling both of these, we encounter a pathology in the M-step of the procedure first highlighted by Antorán et al. (2022c): the standard objective therein is ill-defined when the NN contains normalisation layers. Rather than using the solution proposed in Antorán et al. (2022c), which involves introducing more hyperparameters, we show that a standard feature-normalisation method, the g-prior (Zellner, 1986; Minka, 2000), resolves this pathology. For the tomographic reconstruction task, the regression problem requires a dual-form formulation of our E-step; interestingly, we show that this is equivalent to an optimisation viewpoint on Matheron's rule (Journel & Huijbregts, 1978; Wilson et al., 2020), a connection that we believe to be novel.

2 Conjugate Gaussian regression and the EM algorithm

We study Bayesian conjugate Gaussian linear regression with multidimensional outputs, where we observe inputs $x_1, \dots, x_n \in \mathbb{R}^d$ and corresponding outputs $y_1, \dots, y_n \in \mathbb{R}^m$. We model these as

$$y_i = \phi(x_i)\theta + \eta_i, \tag{1}$$

where $\phi\colon \mathbb{R}^d \mapsto \mathbb{R}^m \times \mathbb{R}^{d'}$ is a known embedding function. The parameters θ are assumed sampled from $\mathcal{N}(0, A^{-1})$ with an unknown precision matrix $A \in \mathbb{R}^{d' \times d'}$, and for each $i \leq n$, $\eta_i \sim \mathcal{N}(0, B_i^{-1})$ are additive noise vectors with precision matrices $B_i \in \mathbb{R}^{m \times m}$ relating the m output dimensions.

Our goal is to infer the posterior distribution for the parameters θ given our observations, under the setting of A of the form $A = \alpha I$, for $\alpha > 0$, most likely to have generated the observed data. For this, we will use the iterative procedure of Mackay (1992). It alternates computing the posterior for θ, denoted Π, for a given choice of A, and updating A, until the pair (A, Π) converge to a locally optimal setting. This corresponds to an EM algorithm (Bishop, 2006).

Henceforth, we will use the following stacked notation: we write $Y \in \mathbb{R}^{nm}$ for the concatenation of $y_1, \dots, y_n$; $\mathrm{B} \in \mathbb{R}^{nm \times nm}$ for a block diagonal matrix with blocks $B_1, \dots, B_n$ and $\Phi = [\phi(X_1); \dots ; \phi(X_n)]^T \in \mathbb{R}^{nm \times d'}$ for the embedded design matrix. We write $M =: \Phi^T \mathrm{B} \Phi$. Additionally, for a vector v and a PSD matrix G of compatible dimensions, $\|v\|_G^2 = v^T G v$.

With that, the aforementioned EM algorithm starts with some initial $A \in \mathbb{R}^{d' \times d'}$, and iterates:

- (E step) Given A, the posterior for θ, denoted Π, is computed exactly as
$$\Pi = \mathcal{N}(\bar{\theta}, H^{-1}) \quad \text{where} \quad H = M + A \quad \text{and} \quad \bar{\theta} = H^{-1}\Phi^T \mathrm{B} Y. \tag{2}$$
- (M step) We lower bound the log-probability density of the observed data, i.e. the evidence for the model with posterior Π and precision A' as (derivation in Appendix B.2)
$$\log p(Y; A') \geq -\tfrac{1}{2}\|\bar{\theta}\|_{A'}^2 - \tfrac{1}{2}\log\det(I + A'^{-1}M) + C =: \mathcal{M}(A'), \tag{3}$$
for C independent of A'. We choose an A that improves this lower bound.

Limited scalability The above inference and hyperparameter selection procedure for Π and A is futile when both d' and nm are large. The E-step requires the inversion of a $d' \times d'$ matrix and the M-step evaluating its log-determinant, both cubic operations in d'. These may be rewritten to instead yield a cubic dependence on nm (as in Section 3.3), but under our assumptions, that too is not computationally tractable. Instead, we will pursue a stochastic approximation to this EM-procedure.

3 Evidence maximisation using stochastic approximation

We now present our main contribution, a stochastic approximation to MacKay's iterative algorithm, presented in the previous section; an M-step requiring access only to samples from Π, and a method for obtaining approximate posterior samples through stochastic optimisation for the E-step.

3.1 Hyperparameter selection using posterior samples (M-step)

For now, assume that we have an efficient method of obtaining samples $\zeta_1, \dots, \zeta_k \sim \Pi^0$ at each step, where Π^0 is a zero mean version of Π, and access to $\bar{\theta}$, the mean of Π. Evaluating the first order optimality condition for $\mathcal{M}$ (see Appendix B.3) yields that the optimal choice of A satisfies

$$\|\bar{\theta}\|_A^2 = \mathrm{Tr}\left\{H^{-1}M\right\} =: \gamma, \tag{4}$$

where the quantity γ is the effective dimension of the regression problem. It can be interpreted as the number of directions in which the weights θ are strongly determined by the data. Setting $A = \alpha I$ for $\alpha = \gamma / \|\bar{\theta}\|^2$ yields a contraction step converging towards the optimum of $\mathcal{M}$ (Mackay, 1992).

Computing γ directly requires the inversion of H, a cubic operation. We instead rewrite γ as an expectation with respect to Π^0 using Hutchinson (1990)'s trick, and approximate it using samples as

$$\gamma = \operatorname{Tr}\{H^{-1}M\} = \operatorname{Tr}\{H^{-\frac{1}{2}}MH^{-\frac{1}{2}}\} = \mathbb{E}[\zeta_1^T M \zeta_1] \approx \tfrac{1}{k}\textstyle\sum_{j=1}^{k} \zeta_j^T \Phi^T \mathrm{B} \Phi \zeta_j =: \hat{\gamma}. \tag{5}$$

We then select $\alpha = \hat{\gamma}/\|\bar{\theta}\|^2$. We have thus avoided the explicit cubic cost of computing the log-determinant in $\mathcal{M}$ or inverting H. Due to the block structure of B, $\hat{\gamma}$ may be computed in order n vector-matrix products.

3.2 Reliably sampling from the linear model's posterior using SGD (E-step)

Now we turn to sampling from $\Pi^0 = \mathcal{N}(0, H^{-1})$. It is known (also, shown in Appendix C.1) that for $\mathcal{E} \in \mathbb{R}^{nm}$ the concatenation of $\varepsilon_1, \ldots, \varepsilon_n$ with $\varepsilon_i \sim \mathcal{N}(0, B_i^{-1})$ and $\theta^0 \sim \mathcal{N}(0, A^{-1})$, the minimiser of the following loss is a random variable ζ with distribution Π^0:

$$L(z) = \tfrac{1}{2}\|\Phi z - \mathcal{E}\|_{\mathrm{B}}^2 + \tfrac{1}{2}\|z - \theta^0\|_A^2. \tag{6}$$

This is known as the sample-then-optimise method (Papandreou & Yuille, 2010). We may obtain a posterior sample by optimising this quadratic loss for a given sample pair $(\mathcal{E}, \theta^0)$. Examining L:

- The first term is data dependent, and corresponds to scaled squared error in fitting $\mathcal{E}$ as a linear combination of Φ. Its gradient requires stochastic approximation for large datasets.
- The second term, a regulariser centred at θ^0, does not depend on the data. Thus, its gradient can be computed exactly at every optimisation step.

Predicting $\mathcal{E}$, i.e. random noise, well from Φ is hard. Due to this, the variance of a mini-batch estimate of the gradient of $\|\Phi z - \mathcal{E}\|_{\mathrm{B}}^2$ may be large. Instead, for $\mathcal{E}$ and θ^0 defined as before, we propose the following alternative loss, equal to L up to an additive constant independent of z:

$$L'(z) = \tfrac{1}{2}\|\Phi z\|_{\mathrm{B}}^2 + \tfrac{1}{2}\|z - \theta^n\|_A^2 \quad \text{with} \quad \theta^n = \theta^0 + A^{-1}\Phi^T \mathrm{B}\mathcal{E}. \tag{7}$$

The mini-batch gradients of L' and L are equal in expectation (see Appendix C.1). However, in L' the randomness from the noise samples $\mathcal{E}$ and the prior sample θ^0 both feature within the regularisation term—the gradient of which can be computed exactly—rather than in the data-dependent term. This may lead to a lower minibatch gradient variance. To see this, consider the variance of the stochastic gradient estimators for the data dependent terms. At $z \in \mathbb{R}^{d'}$, for $j \sim \mathrm{Unif}(\{1, \ldots, n\})$, these are

$$\hat{g} = n\phi(X_j)^T(\phi(X_j)z - \varepsilon_j) \quad \text{and} \quad \hat{g}' = n\phi(X_j)^T\phi(X_j)z \tag{8}$$

for L and L', respectively. Direct calculation, presented in Appendix C.2, shows that

$$\tfrac{1}{n}\left[\mathrm{Var}\hat{g} - \mathrm{Var}\hat{g}'\right] = \mathrm{Var}(\Phi^T\mathrm{B}\mathcal{E}) - 2\mathrm{Cov}(\Phi^T\mathrm{B}\Phi z, \Phi^T\mathrm{B}\mathcal{E}) =: \Delta. \tag{9}$$

Note that both $\mathrm{Var}\hat{g}$ and $\mathrm{Var}\hat{g}'$ are $d' \times d'$ matrices. We impose an order on these by considering their traces: we prefer the new gradient estimator $\hat{g}'$ if the sum of its per-dimension variances is lower than that of $\hat{g}$; that is if $\operatorname{Tr}\Delta > 0$. We analyse two key settings:

- At initialisation, taking $z = \theta^0$ (or any other initialisation independent of $\mathcal{E}$),
$$\operatorname{Tr}\Delta = \operatorname{Tr}\{\Phi^T\mathrm{B}\mathbb{E}[\mathcal{E}\mathcal{E}^T]\mathrm{B}\Phi\} - \operatorname{Tr}\{\Phi^T\mathrm{B}\Phi\mathbb{E}[\theta^0\mathcal{E}^T]\mathrm{B}\Phi\} = \operatorname{Tr}\{M\} > 0, \tag{10}$$
where we used that $\mathbb{E}[\mathcal{E}\mathcal{E}^T] = \mathrm{B}^{-1}$ and since $\mathcal{E}$ is zero mean and independent of θ^0, we have $\mathbb{E}[\theta^0\mathcal{E}^T] = \mathbb{E}\theta^0\mathbb{E}\mathcal{E}^T = 0$. Thus the new objective L' is preferred.
- At convergence, that is, when $z = \zeta$, a more involved calculation presented in Appendix C.3 shows that L' is preferred if
$$2\alpha\gamma > \operatorname{Tr} M. \tag{11}$$
This is satisfied if α is large relative to the eigenvalues of M (see Appendix C.4), that is, when the parameters are not strongly determined relative to the prior.

When L' is preferred both at initialisation and at convergence, we expect it to have lower variance for most minibatches throughout training. Even if the proposed objective L' is not preferred at convergence, it may still be preferred for most of the optimisation, before the noise is fit well enough.

3.3 DUAL FORM OF E-STEP: MATHERON'S RULE AS OPTIMISATION

The minimiser of L and L' is $\zeta = H^{-1}(A\theta^0 + \Phi^T \mathrm{B}\mathcal{E})$. The dual (kernelised) form of this is

$$\zeta = \theta^0 + A^{-1}\Phi^T(B^{-1} + \Phi A^{-1}\Phi^T)^{-1}(\varepsilon - \Phi\theta^0), \tag{12}$$

which is known in the literature as Matheron's rule (Wilson et al., 2020, see Appendix D). Evaluating (12) requires solving a mn-dimensional quadratic optimisation problem, which may be preferable to the primal problem for $mn < d'$; however, this dual form cannot be minibatched over observations. For small n, we can solve this optimisation problem using iterative full-batch quadratic optimisation algorithms (e.g. conjugate gradients), significantly accelerating our sample-based EM iteration.

4 NN UNCERTAINTY QUANTIFICATION AS LINEAR MODEL INFERENCE

Consider the problem of m-output prediction. Suppose that we have trained a neural network of the form $f\colon \mathbb{R}^{d'} \times \mathbb{R}^d \mapsto \mathbb{R}^m$, obtaining weights $\bar{w} \in \mathbb{R}^{d'}$ by optimising a loss of the form

$$\mathcal{L}(f(w, \cdot)) = \textstyle\sum_{i=1}^n \ell(y_i, f(w, x_i)) + \mathcal{R}(w) \tag{13}$$

over $w \in \mathbb{R}^{d'}$, where ℓ is a data fit term (a negative log-likelihood) and $\mathcal{R}$ is a regulariser. We first look at how to use linearised Laplace to quantify uncertainty for the network predictions $f(\bar{w}, \cdot)$. We then present the g-prior, a feature normalisation that resolves a certain pathology in the linearised Laplace method when the network f contains normalisation layers.

4.1 THE LINEARISED LAPLACE METHOD

The linearised Laplace method consists of two consecutive approximations, the latter of which is necessary only if ℓ is non-quadratic (that is, the likelihood is non-Gaussian):

1. We take a first-order Taylor expansion of f around $\bar{w}$, yielding the surrogate model
$$h(\theta, x) = f(\bar{w}, x) + \phi(x)(\theta - \bar{w}) \;\text{ for }\; \phi(x) = \nabla_w f(\bar{w}, x). \tag{14}$$
This is an affine model in the features $\phi(x)$ given by the network Jacobian at x.
2. We approximate the loss of the linear model $\mathcal{L}(h(\theta, \cdot))$ with a quadratic, and treat it as a negative log-density for the parameters θ, yielding a Gaussian posterior of the form
$$\mathcal{N}(\bar{\theta},\ (\nabla^2_\theta \mathcal{L})(h(\bar{\theta}, \cdot))^{-1}). \tag{15}$$
Direct calculation shows that $(\nabla^2_\theta \mathcal{L})(h(\bar{\theta}, \cdot) = (A + \Phi^T \mathrm{B}\Phi) = H$, for $\nabla^2_w \mathcal{R}(\bar{w}) = A$ and B a block diagonal matrix with blocks $B_i = \nabla^2_{\hat{y}_i} l(y_i, \hat{y}_i)$ evaluated at $\hat{y}_i = h(\bar{\theta}, x_i)$.

We have thus recovered a conjugate Gaussian multi-output linear model. We treat A as a learnable parameter thereafter.[1] In practice, we depart from the above procedure in two ways:

- We use the neural network output $f(\bar{w}, \cdot)$ as the predictive mean, rather than the surrogate model mean $h(\bar{\theta}, \cdot)$. Nonetheless, we still need to compute $\bar{\theta}$ to use within the M-step of the EM procedure: this is simply the minimiser of $\mathcal{L}(h(\theta, \cdot))$ over $\theta \in \mathbb{R}^{d'}$.
- We compute the loss curvature B_i at predictions $\hat{y}_i = f(\bar{w}, x_i)$ in place of $h(\bar{\theta}, x_i)$, since the latter would change each time the regulariser A is updated, requiring expensive re-evaluation.

Both of these departures are recommended within the literature (Antorán et al., 2022c).

4.2 ON THE COMPUTATIONAL ADVANTAGE OF SAMPLE-BASED PREDICTIONS

The linearised Laplace predictive posterior at an input x is $\mathcal{N}(f(\bar{w}, x), \phi(x)H^{-1}\phi(x)^T)$. Even given H^{-1}, evaluating this naïvely requires instantiating $\phi(x)$, at a cost of m vector-Jacobian products (i.e. backward passes) which is prohibitive for large m. However, expectations of any function $r : \mathbb{R}^m \mapsto \mathbb{R}$ under the predictive posterior can be approximated using only samples from Π^0 as

$$\mathbb{E}[r] \approx \tfrac{1}{k}\textstyle\sum_{j=1}^k r(\hat{g}_j) \quad \text{for} \quad \hat{g}_j = f(\bar{w}, x) + \phi(x)\zeta_j \quad \text{with} \quad \zeta_1, \ldots, \zeta_k \sim \Pi^0, \tag{16}$$

which requires only k Jacobian-vector products. In practice, we find k much smaller than m suffices.

[1] The EM procedure from Section 2 is for the conjugate Gaussian-linear model, where it carries guarantees on non-decreasing model evidence, and thus convergence to a local optimum. These guarantees do not hold for non-conjugate likelihood functions, e.g., the softmax-categorical, where the Laplace approximation is necessary.

4.3 FEATURE EMBEDDING NORMALISATION: THE DIAGONAL G-PRIOR

The embedding function $\phi(\cdot) = \nabla_w f(\bar{w}, \cdot)$ used in the linearised Laplace method yields features with potentially arbitrary scales across the d' weight dimensions due to symmetries and indeterminacies in neural networks. This leads to the dimensions of the embeddings having arbitrarily unequal weight under an isotropic prior, that is, when we consider the distribution of $\phi(x)\theta^0$ for $\theta^0 \sim \mathcal{N}(0, \alpha^{-1}I)$.

There are two natural solutions: either normalise the features by their (empirical) second moment, that is use embeddings ϕ' given by

$$\phi'(x) = \phi(x)\,\mathrm{diag}(s) \quad \text{for} \quad s \in \mathbb{R}^{d'} \quad \text{given by} \quad s_i = [\Phi^T \mathrm{B} \Phi]_{ii}^{-1/2}, \tag{17}$$

or likewise scale the prior, setting $A = \alpha\,\mathrm{diag}(s^2)$. The latter formulation is a diagonal version of what is known as the g-prior (Zellner, 1986). When the linearised network contains normalisation layers, such as batchnorm (that is, all modern networks), the g-prior is not just useful, but essential. Antorán et al. (2022c) show that normalisation layers lead to indeterminacies in the Jacobians, that in turn lead to an ill-defined model evidence objective. They propose learning separate regularisation parameters for each normalised layer of the network. While fixing the pathology, this increases the complexity of model evidence optimisation. However, as we show in Appendix E, the g-prior cancels these indeterminacies, allowing for the use of a single regularisation parameter. Given this, we recommend using the diagonalised g-prior as default for linearised Laplace implementations.

5 DEMONSTRATION: SAMPLE-BASED LINEARISED LAPLACE INFERENCE

We demonstrate our linear model inference and hyperparameter selection approach on the problem of estimating the uncertainty in NN predictions with the linearised Laplace method. First, in Section 5.1, we perform an ablation analysis on the different components of our algorithm on MNIST using small LeNet-style CNNs. In this setting, full-covariance Laplace inference (that is, exact linear model inference), is tractable, allowing us to evaluate the quality of our approximations. We then demonstrate our method at scale on CIFAR100 classification with a ResNet18 in Section 5.2, and the dual formulation of our method on CT image reconstruction using U-Net in Section 5.3. We focus both on marginal and joint uncertainty calibration and computational cost.

For all experiments, our method does not need to store covariance matrices H^{-1}, compute their log-determinants, or instantiate Jacobian matrices $\phi(x)$, all of which have hindered previous linearised Laplace implementations. We interact with NN Jacobians only through Jacobian-vector and vector-Jacobian products, which have the same asymptotic computational and memory costs as a NN forward-pass (Novak et al., 2022). Unless otherwise specified, we use the g-prior, resulting in a scalar regularisation parameter. We present a summary of our method in Algorithm 1, an illustrative example in Figure 1, and full algorithmic detail in Appendix F.

5.1 ABLATION STUDY: LENET ON MNIST

We evaluate our approach on MNIST m=10 class image classification, where exact linearised Laplace inference is tractable. The training set consists of n=60000 observations and we employ 3 LeNet-style CNNs of increasing size: "LeNetSmall" (d'=14634), "LeNet" (d'=29226) and "LeNetBig"

Algorithm 1: Sampling-based linearised Laplace hyperparameter learning and inference

Inputs: initial $\alpha > 0$; $k, k' \in \mathbb{N}$, number of samples for stochastic EM and prediction, respectively.

Compute g-prior scaling vector s as in (17)

Sample random regularisers $\theta_1^n, \ldots, \theta_k^n$ per (7)

while α *has not converged* **do**

- Find posterior mode $\bar{\theta}$ by optimising linear model loss $\mathcal{L}(h(\theta, \cdot))$, given in (13)
- Draw posterior samples $\zeta_1 \ldots \zeta_k$ by optimising objective L' with $\theta_1^n, \ldots, \theta_k^n$
- Estimate effective dimension $\hat{\gamma}$ per (5) using samples $\zeta_1 \ldots \zeta_k$
- Update prior precision $\alpha \leftarrow \hat{\gamma} / \|\bar{\theta}\|_2^2$

Sample k' random regularisers $\theta_1^{n'}, \ldots, \theta_{k'}^{n'}$ using optimised α

Draw corresponding posterior samples $\zeta_1', \ldots, \zeta_{k'}'$ using loss L'

Output: posterior samples $\zeta_1', \ldots, \zeta_k'$

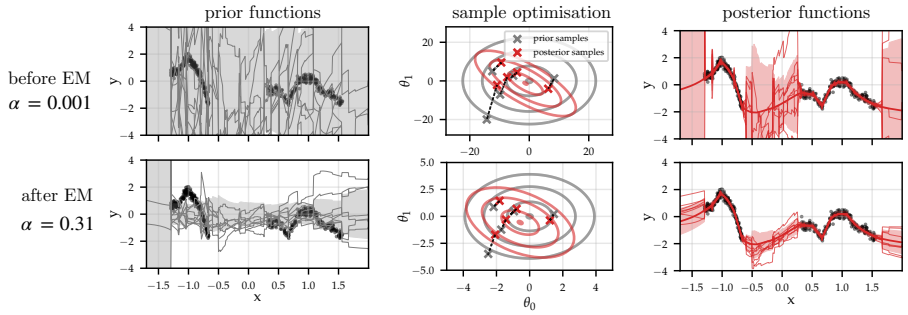

Figure 1: Illustration of our procedure for a fully connected NN on the toy dataset of Antorán et al. (2020). Top: prior function samples present large std-dev. (left). When these samples are optimised (middle shows a 2D slice of weight space), the resulting predictive errorbars are larger than the marginal target variance (right). Bottom: after EM, the std-dev. of prior functions roughly matches that of the targets (left), the overlap between prior and posterior is maximised leading to shorter sample trajectories (center), and the predictive errorbars are qualitatively more appealing (right).

(d'=46024). The latter is the largest model for which we can store the covariance matrix on an A100 GPU. We draw samples and estimate posterior modes using SGD with Nesterov momentum (full details in Appendix G). We use 5 seeds for each experiment, and report the mean and std. error.

Comparing sampling objectives We first compare the proposed objective L' with L, the one standard in the literature, using LeNet, with results shown in Figure 2. We draw exact samples: $(\zeta_j^\star)_{j\leq k}$ through matrix inversion and assess sample fidelity in terms of normalised squared distance to these exact samples $\|\zeta-\zeta^\star\|_2^2/\|\zeta^\star\|_2^2$. All runs share a prior precision of $\alpha \approx 5.5$ obtained with EM iteration. The effective dimension is $\hat{\gamma} \approx 1300$. Despite condition (11) not being satisfied ($2 \times 5.5 \times 1300 < 29k$), the proposed objective converges to more accurate samples even when using a 16-times smaller batch size (left plot). The right side plots relate sample error to categorical symmetrised-KL (sym. KL) and logit Wasserstein-2 (W2) distance between the sampled and exact lin. Laplace predictive distributions on the test-set. Both objectives' prediction errors stop decreasing below a sample error of $\approx$0.5 but, nevertheless, the proposed loss L' reaches lower a prediction error.

Fidelity of sampling inference We benchmark our method using 64 samples, against approximate methods based on the weight point-estimate (MAP), a diagonal lin. Laplace covariance, and a KFAC estimate of the covariance (Martens & Grosse, 2015; Ritter et al., 2018), implemented with the `Laplace` library, against the ground truth lin. Laplace predictive posterior. As standard, we compute categorical predictive distributions with the probit approximation (Daxberger et al., 2021a). All methods use the same layerwise prior precision obtained with 5 steps of full-covariance EM iteration. For all three LeNet sizes, the sampled approximation presents the lowest categorical sym. KL and logit W2 distance to the exact lin. Laplace pred. posterior (Figure 3, LHS). Surprisingly, while the fidelity of competing approximations degrades with model size, that of sampling increases.

Accuracy of sampling hyperparameter selection We compare our sampling EM iteration with 16 samples to the full-covariance EM on LeNet. Figure 3, right, shows that for a single precision hyperparameter, both approaches converge in about 3 steps to the same value. In this setting, the diagonal covariance approximation diverges, and KFAC converges to a biased solution. In the

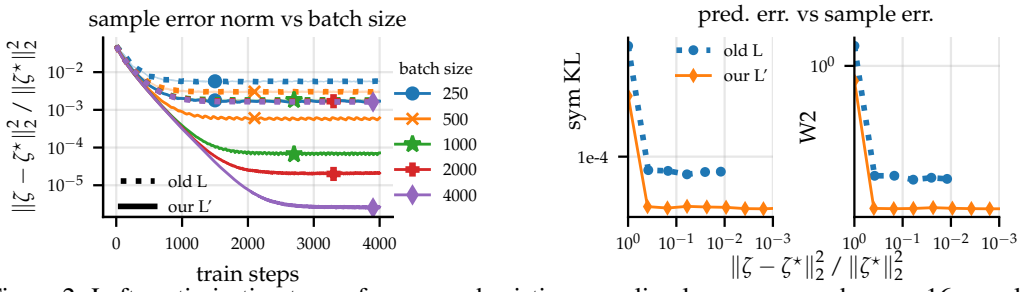

Figure 2: Left: optimisation traces for new and existing sampling losses averaged across 16 samples and 5 seeds. Right: for batch-size=1000, traces of sample-error vs distance to exact linear predictions.

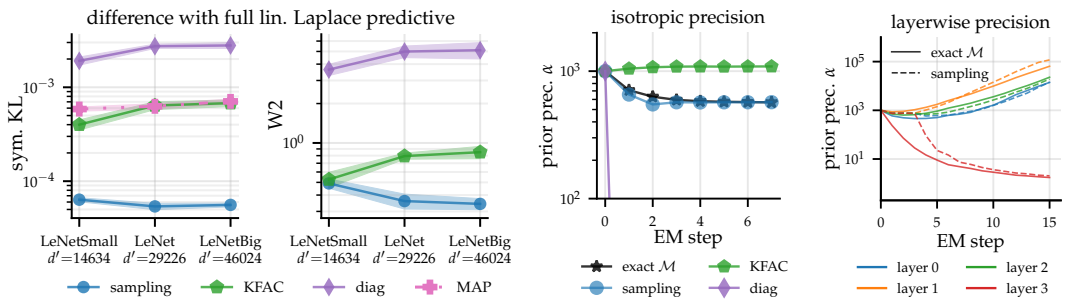

Figure 3: Left: similarity to exact lin. Laplace predictions on MNIST test-set for different approximate methods applied to NNs of increasing size. Centre right: comparison of EM convergence for a single hyperparameter across approximations. Right: layerwise convergence for exact and sampling methods.

layer-wise case, neither method converges within 15 EM steps. The precisions for all but the final layer grow, revealing a pathology of this prior parametrisation: only the final layer's Jacobian (i.e. final layer activations) are needed to accurately predict the targets; other features are pruned away.

5.2 ResNet18 on CIFAR100

We showcase the scalability of our approach by applying it to CIFAR100 m=100 way image classification. The training set consists of n=50000 observations, and we employ a ResNet-18 model with $d' \approx 11M$ parameters. We perform optimisation using SGD with Nesterov momentum and a linear learning rate decay schedule. Unless specified otherwise, we run 8 steps of EM with 6 samples to select α. We then optimise 64 samples to be used for prediction. We run each experiment on 5 different seeds reporting mean and std. error. Full experimental details are in Appendix G.

Stability and cost of sampling algorithm Figure 4 shows that our sample-based EM converges in 6 steps, even using a single sample. At convergence, $\alpha \approx 10^4$ and $\hat{\gamma} \approx 700$, so $2 \times 700 \times 10^4 = 1.4 \times 10^7 > 1.1 \times 10^7$. Thus, (11) is satisfied: L' is preferred. We use 50 epochs of optimisation for the posterior mode and 20 for sampling. When using 2 samples, the cost of one EM step with our method is 45 minutes on an A100 GPU; for the KFAC approximation, this takes 20 minutes.

Evaluating performance in the face of distribution shift We employ the standard Ovadia et al. (2019) benchmark for evaluating methods' test Log-Likelihood (LL) on increasingly corrupted data sets from Hendrycks & Gimpel (2017). We compare the predictions made with our approach to those from 5-element deep ensembles (Lakshminarayanan et al., 2017), arguably the strongest baseline for uncertainty quantification in deep learning (Ashukha et al., 2020), with point-estimated predictions (MAP), and with a KFAC approximation of the lin. Laplace covariance (Ritter et al., 2018). For the latter, constructing full Jacobian matrices for every test point is computationally intractable, so we use 64 samples for prediction, as suggested in Section 4.2. The KFAC covariance structure leads to fast log-determinant computation, allowing us to learn layer-wise prior precisions (following Immer et al. (2021a)) for this baseline using 10 steps of exact EM. For both lin. Laplace methods, we use the standard probit approximation to the categorical predictive (Daxberger et al., 2021b). Figure 5 shows that for in-distribution inputs, ensembles performs best and KFAC overestimates uncertainty, degrading LL relative to point-estimated predictions. Conversely, our method improves LL. For sufficiently corrupted data, our approach outperforms ensembles, also edging out KFAC, which fares well here due to its consistent uncertainty overestimation.

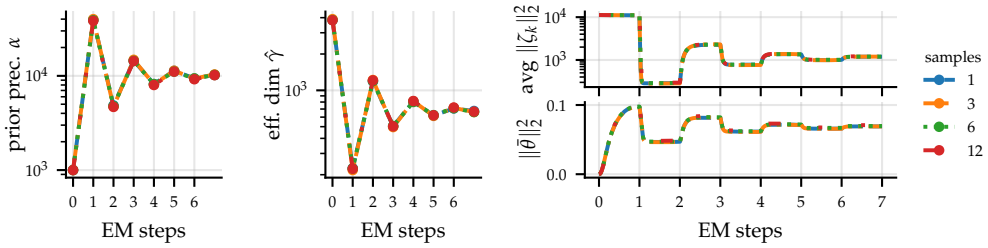

Figure 4: Left: prior precision optimisation traces for ResNet18 on CIFAR100 varying n. samples. Middle: same for the eff. dim. Right: average sample norm and posterior mean norm throughout successive EM steps' SGD runs varying n. samples. Note that traces almost perfectly overlap.

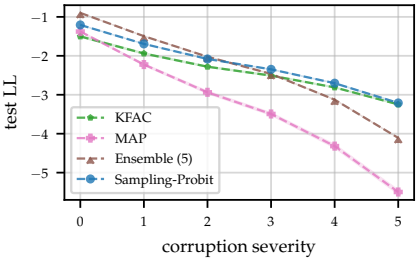

Figure 5: Performance under distribution shift for ResNet18 and CIFAR100.

	κ	MAP	Ensemble (5)	KFAC	Sampling
marginal LL	1	-1.40 ± 0.00	**-0.90 ± 0.00**	-1.12 ± 0.01	-1.07 ± 0.01
joint LL	2	-13.97 ± 0.01	-6.86 ± 0.01	**-4.92 ± 0.04**	-5.14 ± 0.04
	3	-27.89 ± 0.03	-14.17 ± 0.03	-10.83 ± 0.12	**-10.77 ± 0.09**
	4	-41.83 ± 0.03	-22.29 ± 0.04	-19.02 ± 0.22	**-18.04 ± 0.18**
	5	-55.89 ± 0.02	-31.07 ± 0.09	-29.40 ± 0.40	**-26.75 ± 0.26**

Table 1: Comparison of methods' marginal and joint prediction performance for ResNet18 on CIFAR100.

Joint predictions are essential for sequential decision making, but are often ignored in the context of NN uncertainty quantification (Janz et al., 2019). To address this, we replicate the "*dyadic sampling*" experiment proposed by Osband et al. (2022). We group our test-set into sets of κ data points and then uniformly re-sample the points in each set until sets contain τ points. We then evaluate the LL of each set jointly. Since each set only contains κ distinct points, a predictor that models self-covariances perfectly should obtain an LL value at least as large as its marginal LL for all values of κ. We use τ=1000 and repeat the experiment for 10 test-set shuffles. Our setup remains the same as above, except that we use Monte Carlo marginalisation instead of probit, since the latter discards covariance information. Table 1 shows that ensembles make calibrated predictions marginally but their covariance estimates are poor, an observation also made by Osband et al. (2021). Our approach is competitive at all values of κ, preforming best in more challenging large κ cases.

5.3 Tomographic Reconstruction

To demonstrate our approach in dual (kernelised) form, we replicate the setting of Barbano et al. (2022) and Antorán et al. (2022b), where linearised Laplace is used to estimate uncertainty for a tomographic reconstruction output by a U-Net autoencoder. We provide an overview of the problem in Appendix G, but refer to Antorán et al. (2022b) for more details. Whereas Antorán et al. (2022b) use a single step of exact EM on this problem, we use our sample-based variant and run 5 steps. Unless otherwise specified, we use 16 samples for stochastic EM, and 1024 for prediction.

We test on the real-measured μCT dataset of $251k$ pixel scans of a single walnut released by Der Sarkissian et al. (2019b). We train d'=2.97M parameter U-Nets on the m=7680 dimensional observation used by Antorán et al. (2022b) and a twice as large setting m=15360. Here, U-Net's input is clamped to a constant ($n = 1$), and its parameters are optimised to output the reconstructed image. As a result, we do not need mini-batching and can draw samples using Matheron's rule (12). We solve the linear system contained therein using conjugate gradient (CG) iteration implemented with `GPyTorch` (Gardner et al., 2018). We accelerate CG with a randomised SVD preconditioner of rank 400 (alg. 5.6 in Halko et al., 2011). See Appendix G.3 for full experimental details.

Stability and cost of sampling algorithm Figure 6 shows that sample-based EM iteration converges within 4 steps using as few as 2 samples. Table 2 shows the time taken to perform 5 sample-based EM steps for both the m=7680 and m=15360 settings: avoiding explicit estimation of the covariance log-determinant provides us with a two order of magnitude speedup relative to (Antorán et al., 2022c) for hyperparameter learning. By avoiding covariance inversion we obtain an order of magnitude speedup for prediction. Furthermore, while scaling to double the observations m=15360 is intractable with the previous method, it requires only a 25% increase in time using sampling.

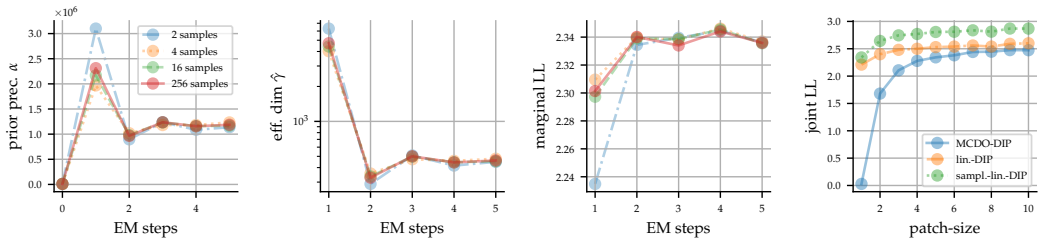

Figure 6: Left 4 plots: traces of prior precision, eff. dim., and marginal LL vs EM steps for tomographic reconstruction task with $m = 7680$. Right: joint test LL vs image patch-size.

Table 2: Tomographic reconstruction: test LL and wall-clock times (A100 GPU) for both target sizes.

	$m = 7680$				$m = 15360$			
	LL		wall-clock time (min.)		LL		wall-clock time (min.)	
Method	marginal	(10×10)	params optim.	prediction	marginal	(10×10)	params. optim.	prediction
MCDO-DIP	0.028	2.474	0	$3'$	0.002	2.762	0	$3'$
lin.-DIP	2.214	2.601	$1260'$	$196'$	–	–	–	–
sampl.-lin.-DIP	**2.341**	**2.869**	$12'$	$14'$	**2.310**	**2.972**	$15'$	$14'$

Predictive performance Figure 7 qualitatively shows that the marginal standard deviation assigned to each pixel by our method aligns with the pixelwise error in the U-Net reconstruction in a fine-grained manner. By contrast, MC dropout (MCDO), the most common baseline for NN uncertainty estimation in CT (Laves et al., 2020; Tölle et al., 2021), places uncertainty more uniformly across large sections of the image. Table 2 shows that the pixelwise LL obtained with our method exceeds that obtained by Antorán et al. (2022b), potentially due to us optimising the prior precision to convergence while the previous work could only afford a single EM step. The rightmost plot in Figure 6 displays joint test LL, evaluated on patches of neighbouring pixels. Our method performs best. MCDO's predictions are poor marginally. They improve when considering covariances, although remaining worse than lin. Laplace. See Appendix H for results for m=15360.

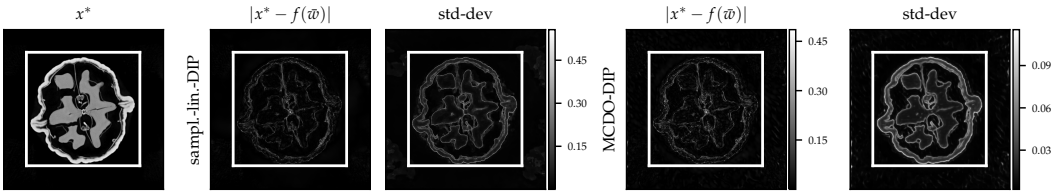

Figure 7: Original 501×501 pixel walnut image and reconstruction error for a m=7680 dimensional observation, along with pixel-wise std-dev obtained with sampling lin. Laplace and MCDO.

6 Related Work

Sample-then-optimise (Papandreou & Yuille, 2010; de G. Matthews et al., 2017) phrases sampling from a conjugate Gaussian-linear model as solving a perturbed quadratic optimisation problem. This has been applied for uncertainty estimation in non-linearised NNs by Osband et al. (2018; 2021), and Pearce et al. (2020), although in this setting it does not draw exact posterior samples. In this work, we show sample-then-optimise to be the primal form of Matheron's rule (Journel & Huijbregts, 1978; Hoffman & Ribak, 1991), a method for updating jointly Gaussian samples into conditional samples, recently repopularised by Wilson et al. (2020).

Linearised neural networks were introduced by Mackay (1992) as an approximation to yield closed-form errorbars for Laplace posteriors. Lawrence (2000) and Ritter et al. (2018) found the Laplace approximation to underperform without the linearisation step. Khan et al. (2019) and Immer et al. (2021b) re-popularised the linearisation step by showing that it improves the quality of uncertainty estimates. Kristiadi et al. (2020) show that the Laplace approximation is sufficient to resolve certain pathologies of point-estimated NNs' predictions. Immer et al. (2021a) and Antorán et al. (2022a;c) explore the linear model's evidence for model selection. Daxberger et al. (2021b) and Maddox et al. (2021) introduce subnetwork and finite differences approaches, respectively, for faster inference with the linearised model. This line of work is intimately related to the neural tangent kernel (Jacot et al., 2018; Lee et al., 2019; Novak et al., 2020) in which NNs are linearised at initialisation.

7 Conclusion

Our work introduced a sample-based approximation to inference and hyperparameter selection in Gaussian linear multi-output models. This allowed us to scale the linearised Laplace method to ResNet18 on CIFAR100, where it was computationally intractable with existing methods. We also demonstrated the strength of the approach on a high resolution tomographic reconstruction task, where it decreases the cost of hyperparameter selection by two orders of magnitude. The uncertainty estimates obtained through our method are well-calibrated not just marginally, but also jointly across predictions. Thus, our work may be of interest in the fields of active and reinforcement learning, where joint predictions are of importance, and computation of posterior samples is often needed.

8 REPRODUCIBILITY STATEMENT

In order to aid reproducibility for our results, we provide the detailed algorithms we use in our two major experiments in Appendix F. We provide full experimental details for all datasets and models used in Appendix G. Therein contained is also a link to an *anonymised code repository*.

REFERENCES

Vincent Adam, Paul E. Chang, Mohammad Emtiyaz Khan, and Arno Solin. Dual parameterization of sparse variational gaussian processes. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 11474–11486, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5fcc629edc0cfa360016263112fe8058-Abstract.html.

Javier Antorán, James Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10620–10634. Curran Associates, Inc., 2020.

Javier Antorán, James Urquhart Allingham, David Janz, Erik Daxberger, Eric Nalisnick, and José Miguel Hernández-Lobato. Linearised laplace inference in networks with normalisation layers and the neural g-prior. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022a. URL https://openreview.net/forum?id=uUH8x-h9zdB.

Javier Antorán, Riccardo Barbano, Johannes Leuschner, José Miguel Hernández-Lobato, and Bangti Jin. A probabilistic deep image prior for computational tomography. *CoRR*, abs/2203.00479, 2022b. doi: 10.48550/arXiv.2203.00479. URL https://doi.org/10.48550/arXiv.2203.00479.

Javier Antorán, David Janz, James Urquhart Allingham, Erik A. Daxberger, Riccardo Barbano, Eric T. Nalisnick, and José Miguel Hernández-Lobato. Adapting the linearised laplace model evidence for modern deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 796–821. PMLR, 2022c. URL https://proceedings.mlr.press/v162/antoran22a.html.

Jordan T. Ash, Cyril Zhang, Surbhi Goel, Akshay Krishnamurthy, and Sham M. Kakade. Anti-concentrated confidence bonuses for scalable exploration. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RXQ-FPbQYVn.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry P. Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Francis R. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.*, 15(1):595–627, 2014. doi: 10.5555/2627435.2627454. URL https://dl.acm.org/doi/10.5555/2627435.2627454.

Daniel Otero Baguer, Johannes Leuschner, and Maximilian Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 36(9): 094004, 2020.

M. Baragatti and D. Pommeret. A study of variable selection using g-prior distribution with ridge parameter. *Computational Statistics & Data Analysis*, 56(6):1920–1934, 2012. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2011.11.017.

Riccardo Barbano, Johannes Leuschner, Maximilian Schmidt, Alexander Denker, Andreas Hauptmann, Peter Maaß, and Bangti Jin. Is deep image prior in need of a good education? *arXiv preprint arXiv:2111.11926*, 2021.

Riccardo Barbano, Javier Antorán, José Miguel Hernández-Lobato, and Bangti Jin. A probabilistic deep image prior over image space. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022. URL https://openreview.net/forum?id=qtFPfwJWowM.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Daniel Sabanés Bové and Leonhard Held. Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387 – 410, 2011. doi: 10.1214/11-BA615.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux–effortless Bayesian deep learning. In M. Ranzato, A. Beygelzimer, P.S. Liang, J.W. Vaughan, and Y. Dauphin (eds.), *Advances in Neural Information Processing Systems*, volume 34, 6–14 Dec 2021a. URL https://papers.nips.cc/paper/2021/hash/a7c9585703d275249f30a088cebba0ad-Abstract.html.

Erik Daxberger, Eric Nalisnick, James U Allingham, Javier Antorán, and Jose Miguel Hernandez-Lobato. Bayesian deep learning via subnetwork inference. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2510–2521. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/daxberger21a.html.

Alexander G. de G. Matthews, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Sample-then-optimize posterior sampling for bayesian linear models. 2017.

Henri Der Sarkissian, Felix Lucka, Maureen van Eijnatten, Giulia Colacicco, Sophia Coban, and Joost Batenburg. Cone-beam x-ray ct data collection designed for machine learning: Samples 1-8. 2019a.

Henri Der Sarkissian, Felix Lucka, Maureen van Eijnatten, Giulia Colacicco, Sophia Bethany Coban, and K. Joost Batenburg. Cone-Beam X-Ray CT Data Collection Designed for Machine Learning: Samples 1-8, 2019b. URL https://doi.org/10.5281/zenodo.2686726. *Zenodo*.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis R. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *J. Mach. Learn. Res.*, 18:101:1–101:51, 2017. URL http://jmlr.org/papers/v18/16-335.html.

Jacob R. Gardner, Geoff Pleiss, Kilian Q. Weinberger, David Bindel, and Andrew Gordon Wilson. GPyTorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 7587–7597. 2018.

Stephen F. Gull. *Bayesian Data Analysis: Straight-line fitting*, pp. 511–518. Springer Netherlands, Dordrecht, 1989. ISBN 978-94-015-7860-8. doi: 10.1007/978-94-015-7860-8_55.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2): 217–288, 2011. doi: 10.1137/090771806. URL https://doi.org/10.1137/090771806.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

James Hensman, Nicoló Fusi, and Neil D. Lawrence. Gaussian processes for big data. In Ann E. Nicholson and Padhraic Smyth (eds.), *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press, 2013. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2389&proceeding_id=29.

Yehuda Hoffman and Erez Ribak. Constrained Realizations of Gaussian Fields: A Simple Algorithm. *The Astrophysical Journal Letters*, 380:L5, October 1991. doi: 10.1086/186160.

M.F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990. doi: 10.1080/03610919008812866. URL https://doi.org/10.1080/03610919008812866.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Mohammad Emtiyaz Khan. Scalable marginal likelihood estimation for model selection in deep learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4563–4573. PMLR, 2021a. URL http://proceedings.mlr.press/v139/immer21a.html.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 703–711. PMLR, 2021b. URL http://proceedings.mlr.press/v130/immer21a.html.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

David Janz, Jiri Hron, Przemyslaw Mazur, Katja Hofmann, José Miguel Hernández-Lobato, and Sebastian Tschiatschek. Successor uncertainties: Exploration and uncertainty in temporal difference learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 4509–4518, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1b113258af3968aaf3969ca67e744ff8-Abstract.html.

A. G. Journel and C. J. Huijbregts. *Mining geostatistics / [by] A. G. Journel and Ch. J. Huijbregts*. Academic Press London ; New York, 1978. ISBN 0123910501.

Mohammad Emtiyaz Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate inference turns deep networks into gaussian processes. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3088–3098, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/b3bbccd6c008e727785cb81b1aa08ac5-Abstract.html.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5436–5446. PMLR, 2020. URL http://proceedings.mlr.press/v119/kristiadi20a.html.

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

Max-Heinrich Laves, Malte Tölle, and Tobias Ortmaier. Uncertainty estimation in medical image denoising with bayesian deep image prior. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pp. 81–96. Springer, 2020.

Neil D. Lawrence. *Variational inference in probabilistic models*. PhD thesis, University of Cambridge, 2000.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8570–8581, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0d1a9651497a38d8b1c3871c84528bd4-Abstract.html.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008. doi: 10.1198/016214507000001337.

David John Cameron Mackay. *Bayesian Methods for Adaptive Models*. PhD thesis, USA, 1992.

Wesley Maddox, Shuai Tang, Pablo G. Moreno, Andrew Gordon Wilson, and Andreas Damianou. Fast adaptation with linearized neural networks. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2737–2745. PMLR, 2021. URL http://proceedings.mlr.press/v130/maddox21a.html.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13132–13143, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html.

Wesley J. Maddox, Gregory W. Benton, and Andrew Gordon Wilson. Rethinking parameter counting in deep models: Effective dimensionality revisited. *CoRR*, abs/2003.02139, 2020. URL https://arxiv.org/abs/2003.02139.

James Martens and Roger B. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2408–2417. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/martens15.html.

Tom Minka. Bayesian linear regression. July 2000. URL https://www.microsoft.com/en-us/research/publication/bayesian-linear-regression/.

George Nicholson, Marta Blangiardo, Mark Briers, Peter J. Diggle, Tor Erlend Fjelde, Hong Ge, Robert J. B. Goudie, Radka Jersakova, Ruairidh E. King, Brieuc C. L. Lehmann, Ann-Marie Mallon, Tullia Padellini, Yee Whye Teh, Chris Holmes, and Sylvia Richardson. Interoperability of Statistical Models in Pandemic Preparedness: Principles and Reality. *Statistical Science*, 37(2):183 – 206, 2022. doi: 10.1214/22-STS854. URL https://doi.org/10.1214/22-STS854.

Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SklD9yrFPS.

Roman Novak, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Fast finite width neural tangent kernel. In *International Conference on Machine Learning*, pp. 17018–17044. PMLR, 2022.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural*

Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pp. 8626–8638, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a7b238ba0f6502e5d6be14424b20ded-Abstract.html.

Ian Osband, Zheng Wen, Mohammad Asghari, Morteza Ibrahimi, Xiyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *CoRR*, abs/2107.08924, 2021. URL https://arxiv.org/abs/2107.08924.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Xiuyuan Lu, and Benjamin Van Roy. Evaluating high-order predictive distributions in deep learning. *arXiv preprint arXiv:2202.13509*, 2022.

Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, Zachary Nado, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pp. 13969–13980, 2019.

George Papandreou and Alan L. Yuille. Gaussian sampling by local perturbations. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta (eds.), *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pp. 1858–1866. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/d09bf41544a3365a46c9077ebb5e35c3-Abstract.html.

Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in neural networks: Approximately bayesian ensembling. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 234–244. PMLR, 2020. URL http://proceedings.mlr.press/v108/pearce20a.html.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006. ISBN 026218253X.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Skdvd2xAZ.

Daniel E. Runcie, Jiayi Qu, Hao Cheng, and Lorin Crawford. Megalmm: Mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome Biology*, 22(1):213, Jul 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02416-w. URL https://doi.org/10.1186/s13059-021-02416-w.

Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, 2001. URL http://jmlr.org/papers/v1/tipping01a.html.

Michael E. Tipping and Anita C. Faul. Fast marginal likelihood maximisation for sparse bayesian models. In Christopher M. Bishop and Brendan J. Frey (eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, AISTATS 2003, Key West, Florida, USA, January 3-6, 2003*. Society for Artificial Intelligence and Statistics, 2003. URL http://research.microsoft.com/en-us/um/cambridge/events/aistats2003/proceedings/123.pdf.

Malte Tölle, Max-Heinrich Laves, and Alexander Schlaefer. A mean-field variational inference approach to deep image prior for inverse problems in medical imaging. In Mattias P. Heinrich, Qi Dou, Marleen de Bruijne, Jan Lellmann, Alexander Schlaefer, and Floris Ernst (eds.), *Medical Imaging with Deep Learning, 7-9 July 2021, Lübeck, Germany*, volume 143 of *Proceedings of Machine Learning Research*, pp. 745–760. PMLR, 2021. URL https://proceedings.mlr.press/v143/tolle21a.html.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Deep image prior. *Int. J. Comput. Vis.*, 128(7):1867–1888, 2020. doi: 10.1007/s11263-020-01303-4. URL https://doi.org/10.1007/s11263-020-01303-4.

James T. Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Efficiently sampling functions from gaussian process posteriors. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10292–10302. PMLR, 2020. URL http://proceedings.mlr.press/v119/wilson20a.html.

David P. Wipf and Srikantan S. Nagarajan. A new view of automatic relevance determination. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis (eds.), *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pp. 1625–1632. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/hash/9c01802ddb981e6bcfbec0f0516b8e35-Abstract.html.

Arnold Zellner. *On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, pp. 233–243. Elsevier Science, 1986.