Uncertainty in Neural Networks:
Approximately Bayesian Ensembling

Tim Pearce[1], **Felix Leibfried**[2], **Alexandra Brintrup**[1], **Mohamed Zaki**[1], **Andy Neely**[1]
[1]University of Cambridge, [2]PROWLER.io

Abstract

Understanding the uncertainty of a neural network's (NN) predictions is essential for many purposes. The Bayesian framework provides a principled approach to this, however applying it to NNs is challenging due to large numbers of parameters and data. Ensembling NNs provides an easily implementable, scalable method for uncertainty quantification, however, it has been criticised for not being Bayesian. This work proposes one modification to the usual process that we argue does result in approximate Bayesian inference; regularising parameters about values drawn from a distribution which can be set equal to the prior. A theoretical analysis of the procedure in a simplified setting suggests the recovered posterior is centred correctly but tends to have underestimated marginal variance, and overestimated correlation. However, two conditions can lead to exact recovery. We argue that these conditions are partially present in NNs. Empirical evaluations demonstrate it has an advantage over standard ensembling, and is competitive with variational methods.

Interactive demo: teapearce.github.io.

1 Introduction

Neural networks (NNs) are the current dominant force within machine learning, however, quantifying the uncertainty of their predictions is a challenge. This is important for many real-world applications (Bishop, 1994) as well as in auxiliary ways; to drive exploration in reinforcement learning (RL), for active learning, and to guard against adversarial examples (Smith and Gal, 2018; Sünderhauf et al., 2018)

A principled approach to modelling uncertainty is provided by the Bayesian framework. Bayesian Neural Networks (BNNs) model the parameters of a NN as probability distributions computed via Bayes rule (MacKay, 1992). Whilst appealing, the large number of parameters and data points used with modern NNs renders many Bayesian inference techniques that work well in small-scale settings infeasible, e.g. MCMC methods.

Ensembling provides an alternative way to estimate uncertainty: it aggregates the estimates of multiple individual NNs, trained from different initialisations and sometimes on noisy versions of the training data. The variance of the ensemble's predictions may be interpreted as its uncertainty. The intuition is simple: predictions converge to similar results where data has been observed, and will be diverse elsewhere. The chief attraction is that the method scales well to large parameter and data settings, with each individual NN implemented in precisely the usual way.

Whilst ensembling has proven empirically successful (Tibshirani, 1996; Lakshminarayanan et al., 2017; Osband et al., 2016), the absence of connection to Bayesian methodology has drawn critics and inhibited uptake, e.g. Gal (2016) [p. 27].

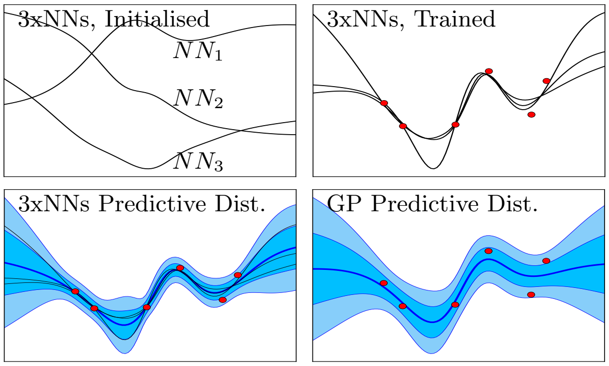


Figure 1: An ensemble of NNs, starting from different initialisations and trained with the proposed modification, produce a predictive distribution approximating that of a GP. This improves with number of NNs.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1.1 Contribution

This paper proposes, analyses and tests one modification to the usual NN ensembling process, with the purpose of examining how closely the resulting procedure aligns with Bayesian inference. The modification regularises parameters about values drawn from an anchor distribution, which can be set to be equal to the prior distribution. We name this procedure *anchored ensembling* - see figure 1 for an illustration. This falls into a family of little known Bayesian inference methods, *randomised MAP sampling* (RMS) (section 2.2).

Our first contributions do not specifically consider NNs; we derive an abstracted version of RMS in parameter space rather than output space. (This abstraction later allows us to propose RMS for classification tasks for the first time.) Under the assumption that the joint parameter likelihood and prior obey a multivariate normal distribution, we show that it is always possible to design an RMS procedure to recover the true posterior.

This design requires knowing the parameter likelihood covariance a priori, which is infeasible except in the simplest models. We propose a workaround that results in an approximation of the posterior. In general this approximation has correct mean but underestimated variance and overestimated correlation. However, two conditions lead to an exact recovery: 1) perfectly correlated parameters, 2) parameters whose marginal likelihood variance is infinite ('extrapolation parameters').

We proceed by considering the applicability of RMS to NNs. We discuss the appropriateness of assumptions used in the theoretical analysis, and argue that the two conditions leading to exact recovery of the posterior are partially present in NNs. We postulate this as the reason that predictive posteriors produced by anchored ensembling appear very similar to those by exact Bayesian methods in figures 4, 6, 7 & 8.

The performance of anchored ensembling is assessed experimentally on regression, image classification, sentiment analysis and RL tasks. It provides an advantage over standard ensembling procedures, and is competitive with variational methods.

2 Background

2.1 Bayesian Neural Networks

A variety of methods have been developed to perform Bayesian inference in NNs. Variational inference (VI) has received much attention both explicitly modelling parameters with distributions (Graves, 2011; Hernández-Lobato and Adams, 2015) and also implicitly through noisy optimisation procedures - MC Dropout (Gal and Ghahramani, 2015), Vadam (Khan et al., 2018). Correlations between parameters are often ignored - mean-field VI (MFVI).

Other inference methods include: Hamiltonian Monte Carlo (HMC), a MCMC variant which provides 'gold standard' inference but at limited scalability (Neal, 1997); The Laplace method fits a multivariate normal distribution to the posterior (Ritter et al., 2018). Whilst ensembling is generally seen as a non-Bayesian alternative, Duvenaud et al. (2016) interpreted it, with early stopping, as approximate inference. Aside from *doing* Bayesian inference, recent works have begun exploring prior design in BNNs, e.g. Pearce et al. (2019).

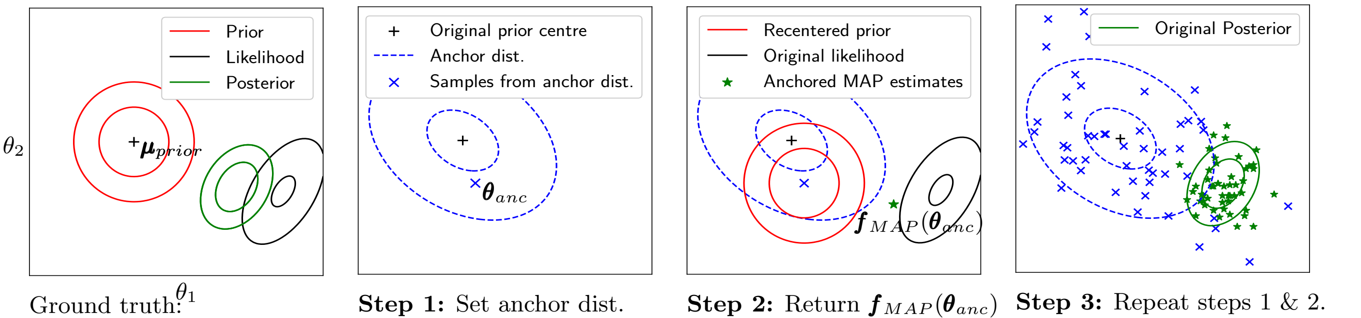
BNNs of infinite width converge to GPs (Neal, 1997). Analytical kernels exist for NNs with certain activation functions, including sigmoidal (Error Function, ERF) (Williams, 1996), Rectified Linear Unit (ReLU) (Cho and Saul, 2009), and leaky ReLU (Tsuchida et al., 2018). Whilst GPs scale superlinearly with data (though see (Wang et al., 2019)), they provide a convenient method for doing exact inference on small problems. In this paper we use these GPs as 'ground truth' predictive distributions to compare to wide NNs. In section 5, we benchmark the ReLU GP on UCI datasets.

2.2 Randomised MAP Sampling

Recent work in the Bayesian community, and independently in the RL community, has begun to explore a novel approach to Bayesian inference. Roughly speaking, it exploits the fact that adding a regularisation term to a loss function returns a maximum a posteriori (MAP) parameter estimate - a point estimate of the Bayesian posterior. Injecting noise into this loss, either to targets or regularisation term, and sampling repeatedly (i.e. ensembling), produces a *distribution* of MAP solutions mimicking that of the true posterior. This can be an efficient method to sample from high-dimensional posteriors (Gu et al., 2007; Chen and Oliver, 2012; Bardsley et al., 2014).

Whilst it is possible to specify a noise injection that produces exact inference in linear regression, there is difficulty in transferring this idea to more complex settings, such as NNs or classification. Directly applying the noise distribution from linear regression to NNs has had some empirical success, despite not reproducing the true posterior (Lu and Van Roy, 2017; Osband et al., 2018) (section 3.2). A more accurate, though more computationally demanding solution, is to wrap the optimisation step in an MCMC procedure (Bardsley, 2012; Bardsley et al., 2014).

These works have been proposed under several names including randomise-then-optimise, randomised prior functions, and ensemble sampling. We refer to this family of procedures *randomised MAP sampling* (RMS).



Ground truth: Analytical Bayesian Inference

Step 1: Set anchor dist. as $\mathcal{N}(\boldsymbol{\mu}_{anc}, \boldsymbol{\Sigma}_{anc})$ from eq. 4 & 5. Sample $\boldsymbol{\theta}_{anc}$ once.

Step 2: Return $\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})$ from eq. 3 with $\boldsymbol{\theta}_{anc}$ sampled from step one.

Step 3: Repeat steps 1 & 2. The distribution of $\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})$ recovers original posterior.

Figure 2: Demonstration of (exact) RMS in a 2D parameter space.

3 RMS Theoretical Results

This section presents several novel results. We first derive a general form of RMS by analysing the procedure in parameter space, using the simplifying assumption that both prior and parameter likelihood are multivariate normal distributions. This is an abstraction compared to previous works. Appendix A contains definitions and proofs in full.

If the parameter likelihood covariance is known a priori, we show how RMS can be designed to recover the true posterior. In general, this will not be known, and we propose a practical workaround requiring knowledge only of the prior distribution.

This workaround no longer guarantees exact recovery of the posterior. We derive results specifying in what ways the estimated RMS posterior is in general biased, including underestimated marginal variance, and overestimated correlation coefficient. We discover two special conditions that lead to an exact recovery.

The appropriateness of the normal assumption in nonlinear models for general data likelihoods will be discussed in section 4, when we consider applying this RMS scheme with workaround to NNs.

3.1 Parameter-Space Derivation

Consider multivariate normal prior and parameter likelihood distributions, $P(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{prior})$, $P_{\boldsymbol{\theta}}(\mathcal{D}|\boldsymbol{\theta}) \propto \mathcal{N}(\boldsymbol{\mu}_{like}, \boldsymbol{\Sigma}_{like})$. We make a distinction between two forms of likelihood: *data likelihood*, which is defined on the domain of the target variable, and *parameter likelihood*, which is specified in parameter space. (See definition 1.)

From Bayes rule the posterior, also normal, is,

$$\mathcal{N}(\boldsymbol{\mu}_{post}, \boldsymbol{\Sigma}_{post}) \propto \mathcal{N}(\boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{prior})\mathcal{N}(\boldsymbol{\mu}_{like}, \boldsymbol{\Sigma}_{like}) \tag{1}$$

The MAP solution is simply $\boldsymbol{\theta}_{\mathrm{MAP}} = \boldsymbol{\mu}_{post}$,

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{like}^{-1}\boldsymbol{\mu}_{like} + \boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{prior}^{-1}\boldsymbol{\mu}_{prior}, \tag{2}$$

where $\boldsymbol{\Sigma}_{post} = (\boldsymbol{\Sigma}_{like}^{-1} + \boldsymbol{\Sigma}_{prior}^{-1})^{-1}$. In RMS we assume availability of a mechanism for returning $\boldsymbol{\theta}_{\mathrm{MAP}}$, and are interested in injecting noise into eq. 2 so that a *distribution* of $\boldsymbol{\theta}_{\mathrm{MAP}}$ solutions are produced, matching the true posterior distribution.

A practical choice of noise source is the mean of the prior, $\boldsymbol{\mu}_{prior}$, since a modeller has full control over this value. Let us replace $\boldsymbol{\mu}_{prior}$ with some noisy random variable, $\boldsymbol{\theta}_{anc}$. This is the same place as a hyperprior over $\boldsymbol{\mu}_{prior}$, though with a subtly different role. Denote $\boldsymbol{f}_{\mathrm{MAP}}(\boldsymbol{\theta}_{anc})$ a function that takes as input $\boldsymbol{\theta}_{anc}$ and returns the resulting MAP estimate,

$$\boldsymbol{f}_{\mathrm{MAP}}(\boldsymbol{\theta}_{anc}) = \boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{like}^{-1}\boldsymbol{\mu}_{like} + \boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{prior}^{-1}\boldsymbol{\theta}_{anc}. \tag{3}$$

Accuracy of this procedure hinges on selection of an appropriate distribution for $\boldsymbol{\theta}_{anc}$, which we term the *anchor distribution*. The distribution that will produce the true posterior can be found by setting $\mathbb{E}[\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})] = \boldsymbol{\mu}_{post}$ and $\mathbb{V}ar[\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})] = \boldsymbol{\Sigma}_{post}$.

Theorem 1. *In order that,* $P(\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})) = P(\boldsymbol{\theta}|\mathcal{D})$, *the required distribution of* $\boldsymbol{\theta}_{anc}$ *is also multivariate normal,* $P(\boldsymbol{\theta}_{anc}) = \mathcal{N}(\boldsymbol{\mu}_{anc}, \boldsymbol{\Sigma}_{anc})$, *where,*

$$\boldsymbol{\mu}_{anc} = \boldsymbol{\mu}_{prior} \tag{4}$$

$$\boldsymbol{\Sigma}_{anc} = \boldsymbol{\Sigma}_{prior} + \boldsymbol{\Sigma}_{prior}\boldsymbol{\Sigma}_{like}^{-1}\boldsymbol{\Sigma}_{prior}. \tag{5}$$

Figure 2 provides a demonstration of the RMS algorithm in 2D parameter space.

3.2 Comparison to Prior Work

Previous work on RMS (Lu and Van Roy, 2017; Osband et al., 2019) was motivated via linear regression. Noting that the MAP solution is given by,

$$\boldsymbol{\theta}_{\mathrm{MAP}} = (\frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_{prior}^{-1})^{-1}(\frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{y} + \boldsymbol{\Sigma}_{prior}^{-1}\boldsymbol{\mu}_{prior}), \tag{6}$$

these works added Gaussian noise to $\boldsymbol{\mu}_{prior}$, *in addition to* adding noise to $\mathbf{y}$, either by additive Gaussian noise or bootstrapping. Eq. 6 is a special case of our own derivation, substituting $\boldsymbol{\Sigma}_{like}^{-1} = 1/\sigma_\epsilon^2\mathbf{X}^T\mathbf{X}$ into eq. 2.

3.3 Practical Workaround: General Case

The previous section showed how to design an RMS procedure that will precisely recover the true Bayesian posterior. Unfortunately, in eq. 5 one must specify the parameter likelihood covariance in order to set the anchor distribution. For most models, this is infeasible.

A practical workaround is to simply ignore the second term in eq. 5 and set $\boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$. Using RMS with this anchor distribution will not generally lead to an exact recovery of the true posterior, however the resulting distribution can be considered an approximation of it. Corollary 1.1 derives this RMS approximate posterior in terms of the true posterior.

Corollary 1.1. *Set $\boldsymbol{\mu}_{anc} := \boldsymbol{\mu}_{prior}$ and $\boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$. The RMS approximate posterior is $P(\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})) = \mathcal{N}(\boldsymbol{\mu}_{post}, \boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{prior}^{-1}\boldsymbol{\Sigma}_{post})$.*

Proof sketch. This follows similar working to theorem 1, but instead of enforcing $\mathbb{E}[\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})] = \boldsymbol{\mu}_{post}, \mathbb{V}ar[\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})] = \boldsymbol{\Sigma}_{post}$ and solving for $\boldsymbol{\mu}_{anc}, \boldsymbol{\Sigma}_{anc}$, we now enforce $\boldsymbol{\mu}_{anc} := \boldsymbol{\mu}_{prior}$, $\boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$ and solve for $\mathbb{E}[\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})], \mathbb{V}ar[\boldsymbol{f}_{MAP}(\boldsymbol{\theta}_{anc})]$.

This corollary shows that the means of the two distributions are aligned, although the covariances are not. Next we state several results quantifying how the RMS approximate posterior covariance differs compared to the true posterior covariance. All results assume multivariate normal prior and parameter likelihood. They can be observed in figure 3 (A).

Lemma 1.1. *When $\boldsymbol{\mu}_{anc} := \boldsymbol{\mu}_{prior}$, $\boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$ the RMS approximate posterior will in general underestimate the marginal variance compared to the true posterior, $\mathbb{V}ar[\boldsymbol{f}_{MAP}(\theta_{anc})] < \mathbb{V}ar[\theta|\mathcal{D}]$.*

Proof sketch. $\boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{prior}^{-1}\boldsymbol{\Sigma}_{post}$ can be rearranged as $\boldsymbol{\Sigma}_{post} - \boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{like}^{-1}\boldsymbol{\Sigma}_{post}$. The second term will be positive definite, so the diagonal entry is positive, and hence, $\text{diag}(\boldsymbol{\Sigma}_{post} - \boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{like}^{-1}\boldsymbol{\Sigma}_{post})_i < \text{diag}(\boldsymbol{\Sigma}_{post})_i$.

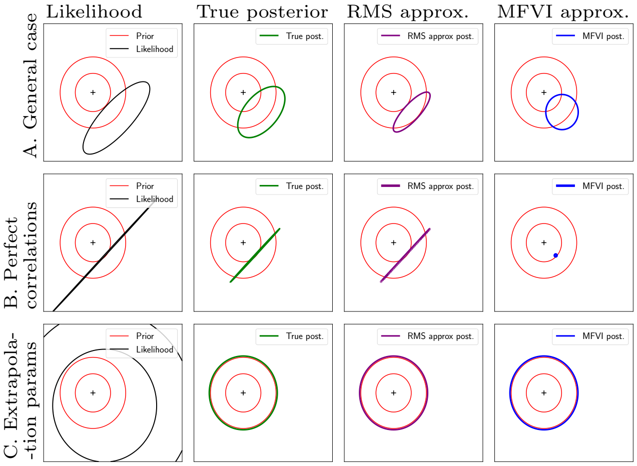


Figure 3: Examples of the RMS approximate posterior when $\boldsymbol{\mu}_{anc} := \boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$. MFVI also shown.

Lemma 1.2. *Additionally assume the prior is isotropic. When $\boldsymbol{\mu}_{anc} := \boldsymbol{\mu}_{prior}$, $\boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$ the eigenvectors (or 'orientation') of the RMS approximate posterior equal those of the true posterior.*

Proof sketch. $\boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{prior}^{-1}\boldsymbol{\Sigma}_{post} = 1/\sigma_{prior}^2\boldsymbol{\Sigma}_{post}^2$. Squaring a matrix only modifies eigenvalues not eigenvectors. As does multiplying by a constant.

Theorem 2. *Additionally assume the prior is isotropic. For a two parameter model, when $\boldsymbol{\mu}_{anc} := \boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$, the RMS approximate posterior will in general overestimate the magnitude of the true posterior parameter correlation coefficient, $|\rho|$. If $|\rho| = 1$, then it will recover it precisely.*

Proof sketch. We compute the individual entries resulting from the required 2 × 2 matrix multiplications.

We were unable to generalise theorem 2 beyond a two parameter model, but numerical examples (appendix B.1) suggest that it holds for higher dimensionality.

3.4 Practical Workaround: Special Cases

Having described the covariance bias that in general will be present in the RMS approximate posterior, we now give two special conditions under which there is no bias, and the true posterior is exactly recovered. Illustrations of these cases are shown in figure 3 (B, C).

Theorem 3. *For extrapolation parameters (def. 2 - parameters which do not affect data likelihood but may affect new predictions) of a model, setting $\boldsymbol{\mu}_{anc} := \boldsymbol{\mu}_{prior}$, $\boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$, means the marginal RMS approximate posterior equals that of the marginal true posterior.*

Proof sketch. We show that the required matrix multiplications, $\boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{prior}^{-1}\boldsymbol{\Sigma}_{post}$, do not affect rows corresponding to extrapolation parameters.

Theorem 4. *Set $\boldsymbol{\mu}_{anc} := \boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{anc} := \boldsymbol{\Sigma}_{prior}$. The RMS approximate posterior will exactly equal the true posterior, $\boldsymbol{\Sigma}_{post}$, when all eigenvalues of a scaled version of $\boldsymbol{\Sigma}_{post}$ (scaled such that the prior equals the identity matrix) are equal to either 0 or 1. This corresponds to posteriors that are a mixture of perfectly correlated and perfectly uncorrelated parameters.*

Proof sketch. We are searching for solutions to $\boldsymbol{\Sigma}_{post} = \boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{prior}^{-1}\boldsymbol{\Sigma}_{post}$. Applying a scaling, $\boldsymbol{\Sigma}_{post}' = \boldsymbol{\Sigma}_{prior}^{-1/2}\boldsymbol{\Sigma}_{post}\boldsymbol{\Sigma}_{prior}^{-1/2}$, results in a slightly simpler equation to find a solution to, $\boldsymbol{\Sigma}_{post}' = \boldsymbol{\Sigma}_{post}'^2$. Results for idempotent matrices tell us that if $\boldsymbol{\Sigma}_{post}'$ is singular with all eigenvalues equal to 0 or 1, this will be a solution.

To provide intuition behind theorem 4 consider a two

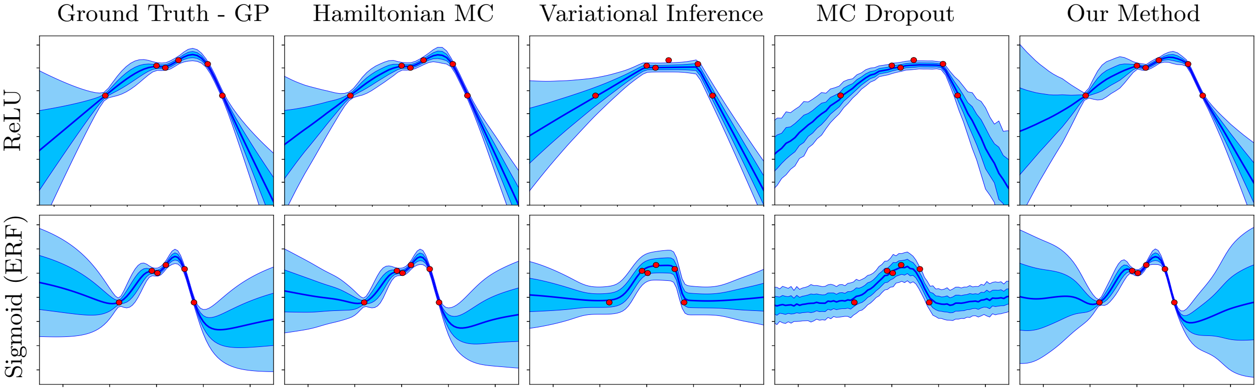


Figure 4: Predictive distributions produced by various inference methods (columns) with varying activation functions (rows) in single-layer NNs on a toy regression task.

parameter model. If parameters are perfectly correlated, the effect on the data likelihood of an increase in the first can be exactly compensated for by a change in the second. If the region over which this applies is large relative to the prior, the likelihood is a line of negligible width. This leads to a posterior of negligible width spanning the prior. Examples in appendix B.1 show what combinations of parameters this holds for.

This section's proofs show that if these two conditions exist, RMS makes a precise recovery. In practise, one would expect to see an increasingly accurate RMS approximation as these conditions are approached.

4 RMS for Neural Networks

We now apply RMS with practical workaround to NNs. We will refer to this as 'anchored ensembling'.

First, we define the NN loss function to be optimised that corresponds to RMS. We then discuss the validity of the RMS procedure in the context of NNs, given the assumptions made. Finally we consider some matters arising in implementation of the scheme. Appendix, algorithm 1 details the full procedure.

4.1 Loss Function

Consider a NN containing parameters, $\boldsymbol{\theta}$, making predictions, $\hat{\mathbf{y}}$, with H hidden nodes and N data points. If the prior is given by $P(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{prior})$, maximising the following returns MAP parameter estimates. (See appendix A.1 for the standard derivation.)

$$\boldsymbol{\theta}_{MAP} = \text{argmax}_{\boldsymbol{\theta}} \log(P_{\mathcal{D}}(\mathcal{D}|\boldsymbol{\theta})) - \frac{1}{2}||\boldsymbol{\Sigma}_{prior}^{-1/2} \cdot (\boldsymbol{\theta} - \boldsymbol{\mu}_{prior})||_2^2 \tag{7}$$

When $\boldsymbol{\mu}_{prior} = \mathbf{0}$, this is standard L2 regularisation. In order to apply RMS we instead replace $\boldsymbol{\mu}_{prior}$ with some random variable $\boldsymbol{\theta}_{anc}$. To use the practical form of RMS, we will draw $\boldsymbol{\theta}_{anc} \sim \mathcal{N}(\boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{prior})$.

Conveniently, no parametric form of data likelihood has yet been specified. For a regression task assuming homoskedastic Gaussian noise of variance σ_ϵ^2, MAP estimates are found by minimising,

$$Loss_j = \frac{1}{N}||\mathbf{y} - \hat{\mathbf{y}}_j||_2^2 + \frac{1}{N}||\boldsymbol{\Gamma}^{1/2} \cdot (\boldsymbol{\theta}_j - \boldsymbol{\theta}_{anc,j})||_2^2. \tag{8}$$

We have defined a diagonal regularisation matrix, $\boldsymbol{\Gamma}$, where the i^{th} diagonal element is the ratio of data noise of the target variable to prior variance for parameter θ_i, $\text{diag}(\boldsymbol{\Gamma})_i = \sigma_\epsilon^2/\sigma_{prior_i}^2$. Note a subscript has been introduced, $j \in \{1...M\}$, with the view of an ensemble of M NNs, each with a distinct draw of $\boldsymbol{\theta}_{anc}$.

For classification tasks, cross entropy is normally maximised, which assumes a multinomial data likelihood,

$$= -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log \hat{y}_{n,c,j} + \frac{1}{N}||\boldsymbol{\Gamma}^{1/2} \cdot (\boldsymbol{\theta}_j - \boldsymbol{\theta}_{anc,j})||_2^2, \tag{9}$$

where y_c is the label for class $c \in \{1...C\}$. Here, $\text{diag}(\boldsymbol{\Gamma})_i = 1/2\sigma_{prior_i}^2$.

4.2 Validity of RMS in NNs

Theory derived to motivate and analyse RMS assumed a simplified setting of multivariate normal parameter likelihoods. This section discusses this assumption, then considers the prevalence of special conditions (section 3.4) that would lead to a close approximation of the true posterior.

4.2.1 Normal Distribution

Earlier proofs assumed parameter likelihoods follow a multivariate normal distribution. We provide two justifications for using this assumption in NNs.

1) Other approximate Bayesian methods incorporate similar assumptions into their methodologies. MFVI commonly fits a factorised normal distribution to the

posterior. The Laplace approximation fits a multivariate normal distribution to the mode of a MAP solution.

2) In figure 5 we visualise conditional parameter likelihoods for actual NNs trained on regression and classification tasks. After training, a parameter is randomly selected, and all others are frozen. The chosen parameter is varied over a small range and the data likelihood calculated at each point. Hence conditional distributions may be plotted. The plots suggest that thinking of local modes as approximately normally distributed is not unreasonable for the purpose of analysis.

This justifies modelling a single mode of the parameter likelihood as multivariate normal. However, the parameter space of a NN is likely to contain many such modes, with each member of an anchored ensemble ending up at a different one. We believe that many of these modes would be exchangeable, for example arising from parameter symmetries. In this case we believe that MAP solutions would also be exchangeable.

Empirically we did not observe this multimodality being problematic - plots such as figure 8 show predictive posteriors with low bias compared to the true posterior.

4.2.2 Presence of Special Cases

Setting the anchor distribution equal to the prior leads to an RMS approximate posterior that, in general, has underestimated variance and overestimated correlation.

Figures 4, 6 & 8 show predictive distributions for anchored ensembles that very closely approximate the true Bayesian posterior, with little sign of bias. This demands an answer to why, rather than if, anchored ensembling performs such accurate inference in these examples. We believe the reason is the presence of the two special conditions that can lead to exact recovery.

It should be straightforward to see that extrapolation parameters (definition 2) exist in the figures. Many hidden nodes will be dead across the range which contains data. Their corresponding final layer weight then has no effect on the data likelihood, but they do affect predictions outside of the training data.

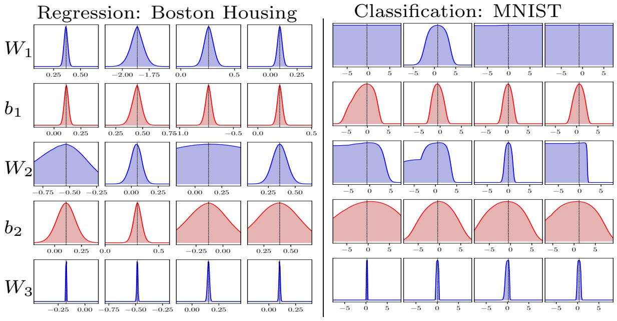


Figure 5: Empirical plots of conditional likelihoods for 4 randomly sampled parameters in two-layer NNs.

It is more difficult to see that perfect correlations also exist, and we provide a numerical example illustrating this in appendix B.3. Essentially it relies on two hidden nodes becoming live in between the same two data points. The associated final layer weights are then perfectly correlated. Whether these special conditions exist beyond fully-connected NNs is something tested indirectly in later experiments with CNNs.

One obvious way to further encourage these conditions is to increase the width of the NN, creating more parameters and an increasing probability of strong correlations. See also a study of multicollinearity in NNs (Cheng et al., 2018) [7.1].

4.3 Implementation Practicalities

How many NNs to use in an RMS ensemble? A large number of samples (and therefore NNs) would be required to fully capture the posterior parameter distributions. By contrast, if one thinks of each NN as an iid sample from a posterior *predictive* distribution, a much smaller number are required, given output dimensionality is typically small. Note this is unaffected by input dimension. Our experiments in section 5 used 5-10 NNs per ensemble, delivering good performance on tasks ranging from 1-10 outputs. See also figure 8. This results in anchored ensembles scaling by $\mathcal{O}(MN)$.

Should the NNs be initialised at anchor points? It is convenient to draw parameter initialisations from the anchor distribution, and regularise directly around these initialised values, however, we found decoupling initialisations from anchor points benefited experiments.

5 Experiments

This section shares high-level findings from experiments. Further details and hyperparameter settings are given in appendix E. Appendix C additionally includes two RL experiments; one testing uncertainty-aware agents for model-free RL, and one applying anchored ensembles to noisy environments for model-based RL. Code is available online (github/TeaPearce). Also see our interactive demo.

5.1 Qualitative Tests

We first examine anchored ensembles on toy problems to gain intuition about its behaviour compared to popular approximate inference and ensembling methods.

Figure 4 compares popular Bayesian inference methods in single-layer NNs for ReLU and sigmoidal nonlinearities. GP and HMC produce 'gold standard' Bayesian inference, and we judge the remaining methods, which are scalable approximations, to them. Both

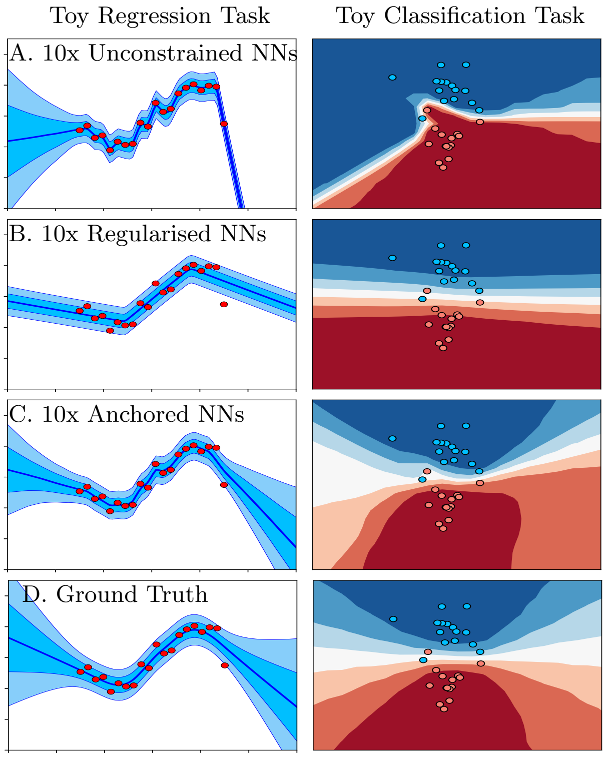


Figure 6: Comparison of NN ensemble loss choices.

MFVI (with a factorised normal distribution) and MC dropout do a poor job of capturing interpolated uncertainty. This is a symptom of the posterior approximation ignoring parameter correlations - see also figure 3 which shows MFVI failing to capture correlations in the posterior. This was explored in Foong et al. (2019).

Figure 6 contrasts anchored ensembles trained on eq. 8 & 9, with NN ensembles using standard loss functions, either with no regularisation term ('unconstrained', $\mathbf{\Gamma} = \mathbf{0}$), or regularised around zero ('regularised', $\boldsymbol{\theta}_{anc,j} = \mathbf{0}$). Regularised produces poor results since it encourages all NNs to the same single solution and diversity is reduced. Unconstrained is also inappropriate - although it produces diversity, no notion of prior is maintained and it overfits the data.

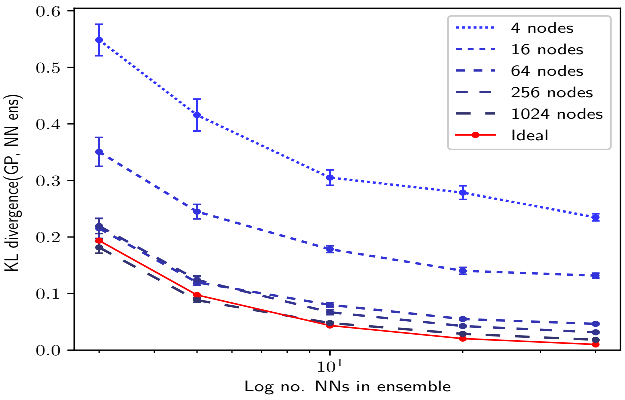


Figure 7: Difference in predictive distributions of an anchored ensemble and a ReLU GP as a function of width and number of NNs. Mean ±1 standard error.

Figure 8 shows the predictive distribution improving with number of NNs compared to a ReLU GP, however it appears a small residual difference remains.

5.2 Convergence Behaviour

To assess how precisely anchored ensembling performs Bayesian inference on a real dataset, we compared its predictive distribution with that of an exact method (ReLU GP) on the Boston housing dataset. Figure 7 quantifies the difference when varying both the width of the NN, and number of NNs in the ensemble. KL divergence between the two predictive distributions was measured and found to decrease as both NN width and number of NNs was increased. As in figure 8 a small amount of residual difference remains even for 40xNNs of $1,024$ nodes.

5.3 UCI Regression Benchmarks

In order to compare anchored ensembles against popular approximate inference methods, we used a standard BNN benchmark. This assesses uncertainty quality for UCI regression tasks on data drawn from the same distribution as the training data (Hernández-Lobato and Adams, 2015). We also implemented the ReLU GP to assess the performance limit on these datasets.

Table 1 lists our results. We include results reported for Deep Ensembles (Lakshminarayanan et al., 2017), which is considered the state-of-the-art ensemble method. Appendix C.3 provides a full comparison with other approximate Bayesian methods including Probabilistic Backpropagation, MC Dropout, and Stochastic Gradient HMC.

Ordering results according to the level of estimated data noise, $\hat{\sigma}_{\epsilon}^2$, shows a clear pattern - anchored ensembles perform best in datasets with low data noise, surpassing both Deep Ensembles and all approximate inference methods listed in appendix C.3. This may be due to an increased importance of interpolation uncertainty when data noise is low, which anchored ensembles models well. On other datasets, the method is also competitive (the Deep Ensemble implementation used additional complexity to capture heteroskedastic uncertainty and has an advantage on higher data noise datasets).

5.4 Out-of-Distribution Classification

We now test on classification tasks, for out-of-distribution (OOD) data, with complex NN architectures, and compare against other ensemble methods.

An uncertainty-aware NN should make predictions of decreasing confidence as it is asked to predict on data

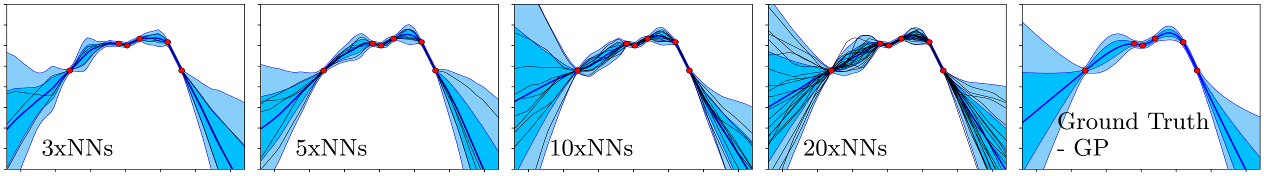


Figure 8: The predictive distribution of an anchored ensemble approaches that of a ReLU GP.

Table 1: NLL regression benchmark results. See appendix C for RMSE and variants of our method. Mean ±1 standard error.

	$\hat{\sigma}_\epsilon^2$	Deep Ens. *State-Of-Art*	Anch. Ens. *Our Method*	ReLU GP[1] *Gold Standard*
		High Epistemic Uncertainty		
Energy	1e-7	1.38 ± 0.22	**0.96 ± 0.13**	0.86 ± 0.02
Naval	1e-7	-5.63 ± 0.05	**-7.17 ± 0.03**	-10.05 ± 0.02
Yacht	1e-7	1.18 ± 0.21	**0.37 ± 0.08**	0.49 ± 0.07
		Equal Epistemic & Aleatoric Uncertainty		
Kin8nm	0.02	-1.20 ± 0.02	**-1.09 ± 0.01**	-1.22 ± 0.01
Power	0.05	**2.79 ± 0.04**	2.83 ± 0.01	2.80 ± 0.01
Concrete	0.05	3.06 ± 0.18	**2.97 ± 0.02**	2.96 ± 0.02
Boston	0.08	**2.41 ± 0.25**	2.52 ± 0.05	2.45 ± 0.05
		High Aleatoric Uncertainty		
Protein	0.5	**2.83 ± 0.02**	2.89 ± 0.01	*2.88 ± 0.00
Wine	0.5	**0.94 ± 0.12**	**0.95 ± 0.01**	0.92 ± 0.01
Song	0.7	**3.35 ± NA**	3.60 ± NA	**3.62 ± NA

[1] For comparison only (not a scalable method). * Trained on 10,000 rows of data. ** Trained on 20,000 rows of data, tested on 5,000 data points.

further from the distribution seen during training. To test this, we report the proportion of high confidence predictions (defined as a softmax output class being ≥ 90%) made by various ensemble systems - unconstrained, regularised, and anchored (as in section 5.1).

We trained on three different datasets, using a NN architecture appropriate to each: 1) Fashion MNIST image classification; 3 fully-connected layers of 100 hidden nodes. 2) IMDb movie review sentiment classification; embedding + 1D convolution + fully-connected layer. 3) CIFAR-10 image classification; convolutional NN (CNN) similar to VGG-13 (9 million parameters).

The confidence of predictions on novel data categories not seen during training was assessed. Table 2 shows example OOD images shown to the NNs trained on CIFAR-10. Edge refers to two CIFAR classes held out during training (ships, dogs). Appendix E provides OOD examples for other datasets.

The three tables show similar patterns. Whilst all methods predict with similar confidence on the training data, confidence differs greatly for other data categories, with anchored ensembles generally producing the most conservative predictions. This gap increases for data drawn further from the training distribution. Encouragingly, we observe similar (though less extreme) behaviour to that in the toy examples of figure 6.

Table 2: Proportion of predictions that were high confidence on out-of-distribution data, e.g. a single regularised NN trained on CIFAR-10 made high confidence predictions 54% of the time when asked to predict on MNIST. Mean over five runs (three for CIFAR).

CIFAR-10 Image Classification, VGG-13 CNN

	Accuracy	Train	Edge	Fashion	MNIST	Scramble	Invert	Noise
1xNNs Reg.	81.6%	0.671	0.466	0.440	0.540	0.459	0.324	0.948
5xNNs Uncons.	85.0%	0.607	0.330	0.208	0.275	0.175	0.209	0.380
5xNNs Reg.	86.1%	0.594	0.296	0.219	0.188	**0.106**	0.153	0.598
5xNNs Anch.	85.6%	0.567	**0.258**	**0.184**	**0.149**	0.134	**0.136**	**0.118**
10xNNs Anch.	86.0%	0.549	0.256	0.119	0.145	0.122	0.124	0.161

IMDb Text Sentiment Classification, Embedding+CNN

	Accuracy	Train	Reuters	Rand. 1	Rand. 2	Rand. 3
1xNNs Reg.	85.3%	0.637	0.119	0.153	0.211	0.326
5xNNs Uncons.	89.1%	0.670	0.102	0.141	0.100	0.075
5xNNs Reg.	87.1%	0.612	0.051	0.091	0.076	0.055
5xNNs Anch.	87.7%	0.603	**0.049**	**0.075**	**0.061**	**0.009**

Fashion MNIST Image Classification, Fully-Connected NN

	Accuracy	Train	Edge	CIFAR	MNIST	Distort	Noise
1xNN Reg.	86.8 %	0.660	0.584	0.143	0.160	0.429	0.364
5xNNs Uncons.	89.0 %	0.733	0.581	0.301	0.104	0.364	0.045
5xNNs Reg.	87.8 %	0.634	**0.429**	0.115	0.072	0.342	0.143
5xNNs Anch.	88.0 %	0.631	0.452	**0.065**	**0.041**	**0.246**	**0.006**

6 Conclusion

This paper proposed, analysed, and tested a modification to the usual NN ensembling process that results in approximate Bayesian inference - regularising parameters around values drawn from a prior distribution.

Under simplifying assumptions, we derived an abstracted form of RMS motivating this. We analysed a practical RMS variant to understand the bias of its approximate posterior. Two special conditions were shown to lead to recovery of the true posterior: perfectly correlated parameters and extrapolation parameters. We discussed the validity of applying RMS to NNs, arguing that these two special conditions are partially present in NNs.

On regression benchmarking experiments, state-of-the-art performance was achieved on 3/10 datasets - outperforming popular approximate inference methods. On image and text classification tasks, anchored ensembles were shown to be more robust than alternative ensemble methods.

Acknowledgements

Thanks to all anonymous reviewers for their helpful comments and suggestions. The lead author was funded through EPSRC (EP/N509620/1) and partially accommodated by the Alan Turing Institute. The model-based RL experiments were run during an internship at PROWLER.io. Thanks to Nicolas Anastassacos for collaborating on an early version of the paper, and Ayman Boustati and Ahmed Al-Ali for helpful discussions.

References

Bardsley, J. M. (2012). MCMC-based image reconstruction with uncertainty quantification. *SIAM Journal on Scientific Computing*, 34(3):1316–1332.

Bardsley, J. M., Solonen, A., Haario, H., and Laine, M. (2014). Randomize-Then-Optimize: A Method for Sampling from Posterior Distributions in Nonlinear Inverse Problems. *SIAM Journal on Scientific Computing*, 36(4).

Bishop, C. (1994). Novelty detection and neural network validation. *IEE Proceedings - Vision, Image, and Signal Processing*, 141(4):217.

Chen, Y. and Oliver, D. S. (2012). Ensemble Randomized Maximum Likelihood Method as an Iterative Ensemble Smoother. *International Association for Mathematical Geosciences*, (D):1–26.

Cheng, X., Khomtchouk, B., Matloff, N., and Mohanty, P. (2018). Polynomial Regression As an Alternative to Neural Nets. In *arXiv:1806.06850*.

Cho, Y. and Saul, L. K. (2009). Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems 22*.

Chua, K., Calandra, R., McAllister, R., and Levine, S. (2018). Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. In *NeurIPS*.

Dearden, R., Friedman, N., and Russell, S. (1998). Bayesian Q-learning. In *American Association for Artificial Intelligence (AAAI)*.

Duvenaud, D., Maclaurin, D., and Adams, R. P. (2016). Early Stopping as Nonparametric Variational Inference. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, pages 1070–1077.

Foong, A. Y. K., Li, Y., Hernández-Lobato, J. M., and Turner, R. E. (2019). 'In-Between' Uncertainty in Bayesian Neural Networks. In *Workshop on Uncertainty and Robustness in Deep Learning, ICML*.

Gal, Y. (2016). *Uncertainty in Deep Learning*. PhD thesis.

Gal, Y. and Ghahramani, Z. (2015). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*.

Graves, A. (2011). Practical Variational Inference for Neural Networks. *Advances in Neural Information Processing Systems*, pages 1–9.

Gu, Y., Oliver, D. S., and Oklahoma, U. (2007). An Iterative Ensemble Kalman Filter for Multiphase Fluid Flow Data Assimilation. *SPE Journal 12(4)*.

Hernández-Lobato, J. M. and Adams, R. P. (2015). Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*.

Khan, M. E., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. (2018). Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam. In *ICML*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *31st Conference on Neural Information Processing Systems*.

Lu, X. and Van Roy, B. (2017). Ensemble Sampling. In *31st Conference on Neural Information Processing Systems*.

MacKay, D. J. C. (1992). A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472.

Mukhoti, J., Stenetorp, P., and Gal, Y. (2018). On the Importance of Strong Baselines in Bayesian Deep Learning. In *Bayesian Deep Learning Workshop, Neural Information Processing Systems (NeurIPS)*, pages 1–4.

Neal, R. M. (1997). *Bayesian Learning for Neural Networks*. PhD thesis.

Osband, I., Aslanides, J., and Cassirer, A. (2018). Randomized Prior Functions for Deep Reinforcement Learning. In *32nd Conference on Neural Information Processing Systems (NIPS 2018)*.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep Exploration via Bootstrapped DQN. In *Advances in neural information processing systems*, pages 1–18.

Osband, I., Russo, D., Wen, Z., and Van Roy, B. (2019). Deep Exploration via Randomized Value Functions. *JMLR*.

Pearce, T., Tsuchida, R., Zaki, M., Brintrup, A., and Neely, A. (2019). Expressive Priors in Bayesian Neural Networks: Kernel Combinations and Periodic Functions. In *Uncertainty in Artifical Intelligence (UAI)*.

Pedersen, M. S., Baxter, B., Templeton, B., Rishøj, C., Theobald, D. L., Hoegh-rasmussen, E., Casteel, G., Gao, J. B., Dedecius, K., Strim, K., Christiansen, L., Hansen, L. K., Wilkinson, L., He, L., Bar, M., Winther, O., Sakov, P., Hattinger, S., Petersen, K. B., and Rishø j, C. (2008). The Matrix Cookbook. *Matrix*, M:1–71.

Ritter, H., Botev, A., and Barber, D. (2018). A Scalable Laplace Approximation for Neural Networks. In *ICLR*, pages 1–15.

Smith, L. and Gal, Y. (2018). Understanding Measures of Uncertainty for Adversarial Example Detection. In *Uncertainty in Artifical Intelligence (UAI)*.

Sünderhauf, N., Brock, O., Scheirer, W., Hadsell, R., Fox, D., Leitner, J., Upcroft, B., Abbeel, P., Burgard, W., Milford, M., and Corke, P. (2018). The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37:405–420.

Tibshirani, R. (1996). A Comparison of Some Error Estimates for Neural Network Models. *Neural Computation*, 8:152–163.

Tsuchida, R., Roosta-Khorasani, F., and Gallagher, M. (2018). Invariance of Weight Distributions in Rectified MLPs. In *Proceedings of the 35th International Conference on Machine Learning*.

Wang, K. A., Pleiss, G., Gardner, J. R., Tyree, S., Weinberger, K. Q., and Wilson, A. G. (2019). Exact Gaussian Processes on a Million Data Points. In *Neural Information Processing Systems*.

Williams, C. K. I. (1996). Computing with infinite networks. In *Advances in Neural Information Processing Systems 9*.