Radial Bayesian Neural Networks: Beyond Discrete Support In Large-Scale Bayesian Deep Learning

Sebastian Farquhar
OATML, University of Oxford

Michael A. Osborne
MLRG, University of Oxford

Yarin Gal
OATML, University of Oxford

Abstract

We propose Radial Bayesian Neural Networks (BNNs): a variational approximate posterior for BNNs which scales well to large models. Unlike scalable Bayesian deep learning methods like deep ensembles that have discrete support (assign exactly zero probability almost everywhere in weight-space) Radial BNNs maintain full support: letting them act as a prior for continual learning and avoiding the *a priori* implausibility of discrete support. Our method avoids a sampling problem in mean-field variational inference (MFVI) caused by the so-called 'soap-bubble' pathology of multivariate Gaussians. We show that, unlike MFVI, Radial BNNs are robust to hyperparameters and can be efficiently applied to challenging real-world tasks without needing ad-hoc tweaks and intensive tuning: on a real-world medical imaging task Radial BNNs outperform MC dropout and deep ensembles.

1 INTRODUCTION

The most effective scalable methods for Bayesian deep learning have a significant shortcoming: they learn an approximate posterior distribution that has discrete support over the weight-space—the probability assigned to almost all possible weights is exactly zero. This is true of methods like MC dropout [Gal and Ghahramani, 2015], but also to samples from stochastic gradient Markov Chain Monte Carlo (MCMC) [Welling and Teh, 2011], or deep ensembles [Lakshminarayanan et al., 2016] where finitely many samples appear in the empirical distribution. This is implausible *a priori*: from an epistemic perspective assigning zero posterior probability almost everywhere is pathological overconfidence. But overconfidence is also unhelpful, these distributions are unsuitable as a data-dependent prior in continual learning—once a prior is exactly zero, no amount of data can update it.

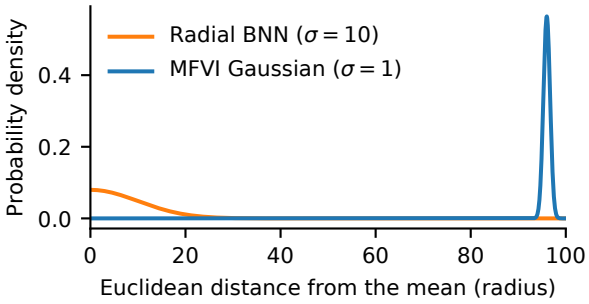

Figure 1: MFVI uses a multivariate Gaussian approximate posterior whose probability mass is tightly clustered at a fixed radius from the mean depending on the number of dimensions—the 'soap-bubble'. In our Radial BNN, samples from the approximate posterior are more reflective of the mean. This helps training by reducing gradient variance. (Plotted p.d.f. is based on dimensionality of a 3x3 conv layer with 64 channels.)

Some variational inference methods *do* learn approximate posteriors with full support over the weight-space. 'Mean-field' variational inference (which assumes independent weight distributions) is fast and has linear time complexity in the number of parameters [Hinton and van Camp, 1993, Graves, 2011, Blundell et al., 2015]. Unfortunately, MFVI struggles in practice for tasks larger than roughly the scale of MNIST and is sensitive to hyperparameters [Wu et al., 2019]. Tuning hyperparameters is a barrier to using MFVI for larger models where each iteration could take days. To make MFVI work, researchers often resort to ad-hoc tweaks to the loss or optimization process which sidestep the variational inference arguments that motivate the approach in the first place! (See §4.1.2.) Other research relaxes MFVI's independence assumption by introducing expensive techniques which are tractable for small networks with only thousands of parameters

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

and low-dimensional problems (e.g., MNIST) [Louizos and Welling, 2016, Sun et al., 2017, 2019, Oh et al., 2019]. **What is missing is robust, scalable inference for BNNs that maintains full support.**

In this paper we identify a sampling problem at the heart of MFVI's failures—typical samples from the multivariate Gaussian approximate posterior used in MFVI are unrepresentative of the most-probable weights, and this problem gets worse for larger networks [Bishop, 2006]. Probability mass in a multivariate Gaussian is clustered in a narrow 'soap-bubble' far from the mean (see Figure 1).[1] Unless the approximate posterior distribution is very tight, samples tend to be distant from each other. This leads to exploding gradient variance whenever the posterior becomes broad, and prevents MFVI from actually fitting to the loss. We demonstrate this in §5.

Therefore, we propose an alternative approximate posterior distribution without a 'soap-bubble'. The Radial BNN defines a simple approximate posterior distribution in a hyperspherical space corresponding to each layer, and then transforms this distribution into the coordinate system of the weights. The typical samples from this distribution tend to come from areas of high probability density. We show that the Radial BNN can be sampled efficiently in weight-space, without needing explicit coordinate transformations, and derive an analytic expression for the loss that makes training as fast and as easy to implement as MFVI.

We establish the robustness and performance of Radial BNNs using a Bayesian medical imaging task identifying diabetic retinopathy in 'fundus' eye images [Leibig et al., 2017], using models with ∼15M parameters and inputs with ∼230,000 dimensions, in §4.1 (see Figure 2). Radial BNNs are more robust to hyperparameter choice than MFVI and that Radial BNNs outperform the current state-of-the-art Monte-Carlo (MC) dropout and deep ensembles on this task.

In addition, because Radial BNN approximate posteriors have *full support* over the weight-space, they can be used as a prior for further inference. We show this in §4.2 using a continual learning setting [Kirkpatrick et al., 2017, Nguyen et al., 2018], where a sequence of approximate posteriors are used as a prior to avoid catastrophic forgetting. While we do not solve continual learning, we use the problem setting to demonstrate the potential to find rich data-dependent priors.

[1]We refer readers to the Appendix A for more detail on this phenomenon. Intuitively, the issue arises because the space expands with the polynomial r^D in the radius r in D dimensions, while the p.d.f. of the Gaussian falls exponentially. At the origin, the polynomial term is small, at infinity the exponential term is small, and almost all the probability mass lies in a narrow band in between.

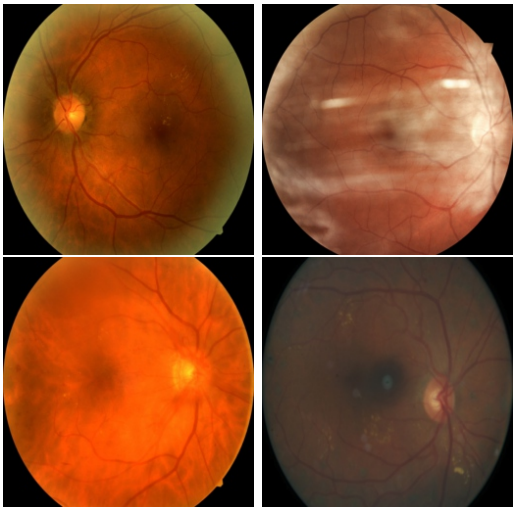

Figure 2: Examples from retinopathy dataset. Top L: healthy eye. Top R: healthy eye with camera artefacts. Bottom: diseased eyes. The chance that bad images cause misdiagnosis makes uncertainty-aware models vital. Input dimension 334x bigger than MNIST.

2 PRIOR WORK

Instead of point estimates, Bayesian neural networks (BNNs) place a parameterized distribution over each weight in a neural network [MacKay, 1992, Neal, 1995]. Many efficient approximations have been proposed to estimate the posterior distribution over those weights including mean-field variational inference [Hinton and van Camp, 1993, Graves, 2011, Blundell et al., 2015], Monte Carlo (MC) dropout [Gal and Ghahramani, 2015], stochastic gradient Markov Chain MC [Welling and Teh, 2011] and others. Deep ensembles [Lakshminarayanan et al., 2016] have also been proposed as a way to learn a distribution over the weights (although the connection to the posterior remains unclear).

Unfortunately, existing robust methods, scalable to large models and datasets, do not learn a posterior with full support over the weight-space. Monte Carlo dropout learns a parameterized distribution based on Bernouilli random variables, which only represent discrete points in weight-space. Methods like deep ensembles and SG-MCMC instead produce a finite number of samples and estimate the predictive distribution using the an empirical distribution with discrete support. These sorts of discrete distributions are epistemically pathological—they represent implausible overconfidence. This makes them unsuitable as a prior for further inference.

Mean-field variational inference offers a fully supported distribution over the weights. It sets an approximate posterior distribution, $q_{\boldsymbol{\theta}}$, over each weight, $\mathbf{w}$, in the

network—an independent Gaussian (where $\boldsymbol{\theta}$ is $\{\boldsymbol{\mu} \cup \boldsymbol{\sigma}\}$). It then optimizes a lower-bound on the marginal likelihood which tries to find the approximate posterior with the smallest KL-divergence to the true posterior. This loss can be interpreted as balancing predictive accuracy on the data ($\mathbf{y}$ and $\mathbf{X}$), the entropy of the posterior, and the cross-entropy between the prior and posterior. In the common case of a unit multivariate Gaussian prior and an approximate posterior $\mathcal{N}(\mu_i, \sigma_i^2)$ over the weights w_i, the negative evidence lower bound (ELBO) objective is:

$$\mathcal{L}_{\text{MFVI}} = \overbrace{\sum_i \frac{1}{2}\left[\sigma_i^2 + \mu_i^2\right]}^{\text{prior cross-entropy}} - \overbrace{\sum_i \log[\sigma_i]}^{\substack{\text{approximate-}\\ \text{posterior entropy}}} - \overbrace{\mathbb{E}_{\mathbf{w}\sim q_\theta(\mathbf{w})}\left[\log p(\mathbf{y}|\mathbf{w}, \mathbf{X})\right]}^{\text{data likelihood}}. \quad (1)$$

In practice, training BNNs with MFVI is difficult. For example, Wu et al. [2019] argue that it is sensitive to initialization and priors. Others worry that the mean-field approximation is too constraining. Louizos and Welling [2016], Sun et al. [2017] and Oh et al. [2019] have all introduced richer variational distributions which permit correlations between weights to be learned by the BNN; Sun et al. [2019] instead perform inference in function-space. Unfortunately, these methods are considerably more computationally expensive than MFVI and have only been demonstrated on problems *at MNIST scale or below*. Wu et al. [2019] instead see the variance of ELBO estimates as the problem and introduce a deterministic alternative. We agree that this is a crucial problem, but offer a simpler and cheaper alternative solution which does not require extra assumptions about the distribution of activations. Note that Osawa et al. [2019] present scalable inference for MFVI using variational online Gauss-Newton methods [Khan et al., 2018]. However, this method relies on several significant approximations to make estimates of the Hessian tractable and the performance of the method lags significantly behind deep ensembles, which we compare to.

We note that there is a superficial similarity between our method and Oh et al. [2019], insofar as they also make use of a hyperspherical coordinate system for variational inference. However, they use this coordinate system over each row in their weight matrix, rather than the whole layer, and introduce an expensive posterior distribution (von Mises-Fisher) to explicitly model weight correlations within rows, whereas we do *not* seek to learn any correlations between parameters in the hyperspherical space. That is, their method uses a different technique to solve a different problem.

3 METHOD

A well-known property of multivariate Gaussians in high dimensions is that the probability mass concentrates in a 'soap-bubble'—a narrow shell at a radius determined by the variance and the number of dimensions (e.g., [Bishop, 2006, Betancourt, 2018]). This has the consequence that almost all samples from the distribution are very distant from the mean. All else equal, we might expect this to lead to predictions and losses from multiple samples of the weights which are less correlated with each other than if the samples were near to each other in weight-space. Moreover, the distance of typical samples from the approximate posterior over each layer from the mean is $\sim \sigma\sqrt{D}$, for standard deviation parameter σ and the number of parameters in the layer D, for the domain of large D typically found in modern neural networks.[2] We anticipate (and demonstrate in §5) that the distance between samples from the MFVI approximate posterior makes the gradient estimator of the log-likelihood term of the loss in Equation (1) have a large variance, which makes optimization difficult.

3.1 The Radial BNN Posterior

A 'soap-bubble' arises when, for large D, the probability density function over the radius from the mean is sharply peaked at a large distance from the mean (see Figure 1). Therefore, we pick a probability distribution which cannot have this property. We can easily write down a probability density function which cannot have a 'soap-bubble' by explicitly modelling the radius from the mean. The hyperspherical coordinate system suits our needs: the first dimension is the radius and the remaining dimensions are angles. We pick the simplest practical distribution in hyperspherical coordinates with no soap bubble:

- In the radial dimension: $r = |\tilde{r}|$ for $\tilde{r} \sim \mathcal{N}(0, 1)$.
- In the angular dimensions: uniform distribution over the hypersphere—all directions equally likely.

A critical property is that it is easy to sample this distribution *in the weight-space coordinate system*—we wish to avoid the expense of explicit coordinate transformations when sampling from the approximate posterior. Instead of sampling the posterior distribution directly, we use the local reparameterization trick Rezende et al. [2014], Kingma et al. [2014], and sample the noise distribution instead. This is similar to Graves [2011],

[2] To calculate the distance of typical samples from the mean we imagine an isotropic posterior. Our posteriors are not isotropic, but the pattern is similar.

Method	Architecture	# Params	Epoch Train Time (m)	ROC-AUC for different percent data referred to experts 0%	10%	20%	30%
MC-dropout	[Leibig et al., 2017]	∼21M	-	92.7±0.3%	93.8±0.3%	94.7±0.3%	95.6±0.3%
MC-dropout	VGG-16	∼15M	5.6	93.0±0.04%	94.1±0.05%	94.5±0.05%	95.1±0.07%
MFVI	VGG-16*	∼15M	16.0	63.6±0.13%	63.5±0.09%	63.5±0.09%	62.6±0.10%
MFVI w/ tweaks	VGG-16*	∼15M	16.0	93.9±0.04%	94.4±0.05%	95.4±0.04%	96.4±0.05%
Radial BNN	VGG-16*	∼15M	16.2	**94.3±0.04%**	**95.3±0.06%**	**96.1±0.06%**	**96.8±0.04%**
Deep Ensemble	3xVGG-16	∼45M	16.8†	93.9±0.04%	96.0±0.05%	96.6±0.04%	97.2±0.04%
Radial Ensemble	3xVGG-16*	∼45M	48.6†	**94.5±0.05%**	**97.9±0.04%**	**98.0±0.03%**	**98.1±0.03%**

Table 1: Diabetic Retinopathy Prescreening: Our Radial BNN outperforms SOTA MC-dropout and is able to scale to model sizes that MFVI cannot handle without ad-hoc tweaks (see §4.1.2). Even with tweaks, Radial BNN still outperforms. Deep Ensembles outperform a single Radial BNN at estimating uncertainty, but are worse than an ensemble of Radial BNNs with the same number of parameters. ± indicates bootstrapped standard error from 100 resamples of the test data. VGG-16* model has fewer channels so that # of parameters is the same as non-Bayesian model. † 3x single model train time. Could be in parallel.

Blundell et al. [2015] who sample their weights

$$\mathbf{w} \coloneqq \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}_{\text{MFVI}}, \tag{2}$$

where $\boldsymbol{\epsilon}_{\text{MFVI}} \sim \mathcal{N}(0, \mathbf{I})$. In order to sample from the Radial BNN posterior we make a small modification:

$$\mathbf{w}_{\text{radial}} \coloneqq \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \frac{\boldsymbol{\epsilon}_{\text{MFVI}}}{\|\boldsymbol{\epsilon}_{\text{MFVI}}\|} \cdot r, \tag{3}$$

which works because dividing a multi-variate Gaussian random variable by its norm provides samples from a direction uniformly selected from the unit hypersphere [Muller, 1959, Marsaglia, 1972]. As a result, sampling from our posterior is nearly as cheap as sampling from the MFVI posterior. The only extra steps are to normalize the noise, and multiply by a scalar Gaussian random variable.

3.2 Evaluating the Objective

To use the our approximate posterior for variational inference we must be able to estimate the ELBO loss. The Radial BNN posterior does not change how the expected log-likelihood is estimated, using mini-batches of datapoints and MC integration.

The KL divergence between the approximate posterior and prior can be written:

$$\begin{aligned} KL\big(q(\mathbf{w}) \parallel p(\mathbf{w})\big) &= \int q(\mathbf{w}) \log[q(\mathbf{w})] d\mathbf{w} \\ &\quad - \int q(\mathbf{w}) \log\big[p(\mathbf{w})\big] d\mathbf{w} \\ &= \mathcal{L}_{\text{entropy}} - \mathcal{L}_{\text{cross-entropy}}. \end{aligned} \tag{4}$$

We estimate the cross-entropy term using MC integration, just by taking samples from the posterior and averaging their log probability under the prior. We find that this is low-variance in practice, and is often done for MFVI as well [Blundell et al., 2015].

We can evaluate the entropy of the posterior analytically. We derive the entropy term in Appendix B:

$$\mathcal{L}_{\text{entropy}} = -\sum_i \log[\sigma_i] + \text{const.} \tag{5}$$

where i sums over the weights. This is, up to a constant, the same as when using an ordinary multivariate Gaussian in MFVI. (For sake of completeness, we also derive the constant terms in the Appendix.) In Appendix C, we also provide a derivation of the cross-entropy loss term in the case where the prior is a Radial BNN. This is useful in continual learning (see §4.2) where we use the posterior from training one model as a prior when training another.

Code implementing Radial BNNs can be found at `https://github.com/SebFar/radial_bnn`.

3.3 Computational Complexity

Training Radial BNNs has the same computational complexity as MFVI—$\mathcal{O}(D)$, where D is the number of weights in the model. In contrast, recent non-mean-field extensions to VI like Louizos and Welling [2016] and Sun et al. [2017] have higher time complexities. For example, Louizos and Welling [2016] uses a pseudo-data approximation which reduces their complexity to $\mathcal{O}(D + M^3)$ where M is a pseudo-data count. But even for MNIST, they use M up to 150 which becomes computationally expensive (this is their largest experiment). Sun et al. [2017] have the same complexity as Louizos and Welling [2016], depending on similar approximations and consider a maximum input dimension of only 16—over 16,000 times smaller than the input dimension of the task we address in §4.1. In practice, the comparison between our Radial BNNs and MFVI is even more favorable. Our method is more robust to hyperparameters, allowing hyperparameters to be selected for training/inference speed to still achieve good accuracy.

4 EXPERIMENTS

Our work is focused on large datasets and big models, which is where the most exciting application for deep learning are. That is where complicated variational inference methods that try to learn weight covariances become intractable, and where the 'soap-bubble' pathology emerges.

We address this head-on in §4.1. We show that on a large-scale diabetic retinopathy diagnosis image classification task: our radial posterior is *more accurate*, has *better calibrated uncertainty*, and is *more robust* to hyperparameters than MFVI with a multivariate Gaussian and therefore requires significantly fewer iterations and less experimenter time. In this setting, we have ∼260,000 input dimensions and use a model with ∼15M parameters. This is orders of magnitude larger than most other VI work, has been heavily influenced by the experimental settings used to evaluate performance on UCI datasets by Hernández-Lobato and Adams [2015] with between 4 and 16 input dimensions and using fewer than 2000 parameters.[3]

In §4.2, we show that we can use the posterior from variational inference with Radial BNNs as a prior when learning future tasks. We demonstrate this using a continual learning problem [Kirkpatrick et al., 2017] on FashionMNIST [Xiao et al., 2017]. We show significantly improved performance relative to the MFVI-based Variational Continual Learning (VCL) introduced by Nguyen et al. [2018].

4.1 Diabetic Retinopathy Prescreening

We perform classification on a dataset of 'fundus' images taken of the back of retinas in order to diagnose diabetic retinopathy [Kaggle, 2015] building on Leibig et al. [2017] and Filos et al. [2019]. Diabetic retinopathy is graded in five stages, where 0 is healthy and 4 is the worst. Following Leibig et al. [2017], we distinguish the healthy (classes 0 and 1) from those that require medical observation and attention (2, 3, and 4). Images (512x512) include left and right eyes separately, which are not considered as a pair by the models, and come from two different camera technologies in many different physical locations. Model uncertainty is used to identify badly-taken or confusing images which could be used to refer affected patients to experts for more detailed examination.

[3]Radial BNNs, like MFVI, do not match the performance of some of the more expensive methods on the UCI datasets. We would not expect it to—our method is specifically designed for models with high-dimensional weight-space, not for the artificial constraints of the experimental settings used on the UCI evaluations. See Appendix E for details.

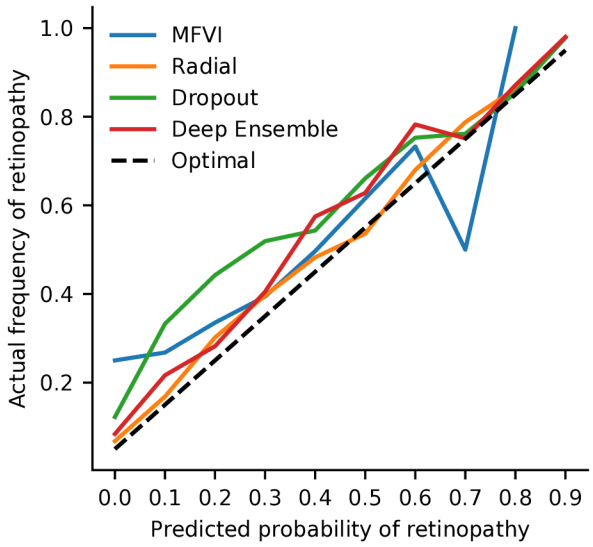

Figure 3: Radial BNN is almost perfectly calibrated, compared with MC dropout and deep ensembles (overconfident) and ordinary MFVI without ad-hoc tweaks which is not well calibrated. X-axis labels are the lower-bound of each range (e.g., 0.0 is 0.0-0.1).

4.1.1 Performance and Calibration

In Table 1 we compare the classification area under the curve (AUC) of the receiver operating characteristic of predicted classes (higher is better).[4] We consider the model performance under different thresholds for referring data to experts. At 0%, the model makes predictions about all data. At 30%, the 30% of images about which the model is least confident are referred to experts and do not get scored for the model—the AUC should therefore become higher if the uncertainties are well-calibrated. We show that our Radial BNN outperforms MFVI by a wide margin, and even outperforms MC dropout. While the deep ensemble is better at estimating uncertainty than a single Radial BNN, it has three times as many parameters. An ensemble of Radial BNNs outperforms deep ensembles at all levels of uncertainty. Radial BNN models trained on this dataset also show empirical calibration that is closer to optimal than other methods (see Figure 3).

The model hyperparameters were all selected individually by Bayesian optimization using ten runs. Full hyperparameters and search strategy, preprocessing, and architecture are provided in Appendix D.1. We include both the original MC dropout results from Leibig et al. [2017] as well as our reimplementation using the same model architecture as our Radial BNN model. The only difference between the MC dropout and Radial BNN/MFVI architectures is that we use more channels for MC dropout, so that the number of

[4]We use AUC because classes are unbalanced (mostly healthy): accuracy gives distorted picture of performance.

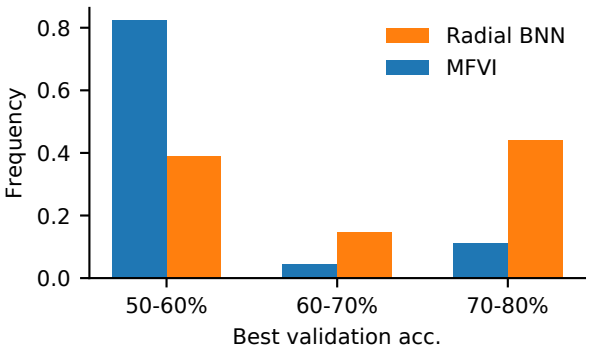

Figure 4: Radial BNN posterior is more robust to hyperparameters on a downsampled version of the retinopathy dataset. Over 80% of configurations for the MFVI baseline learned almost nothing. 4 times more Radial BNNs had good accuracies than MFVI models.

parameters is the same in all models. We estimate the standard error of the AUC using bootstrapping.

4.1.2 MFVI Tweaks

In some cases, researchers have been able to get MFVI to work by applying various ad-hoc tweaks to the training process. Here, we evaluate the performance of these tweaks and in §5 we explain how the success of these tweaks aligns with our hypothesis that MFVI suffers from a sampling problem which Radial BNNs fix.

One approach to making MFVI work is to pre-train the means of the model using the ordinary log-likelihood loss and to switch partway through training to the ELBO loss, initializing the weight variances at this point with a very small value approximating a deterministic neural network. (E.g., [Nguyen et al., 2018] who initialize with a variance of 10^{-6} after pre-training the means.) If one trains to convergence, the weight variances will tend to grow bigger than their tiny initialization, which destroys model performance in MFVI, so one must also employ early stopping.[5] If we perform all these ad-hoc tweaks we are indeed able to get acceptable performance on our diabetic retinopathy dataset (see Table 1), though still worse than Radial BNNs. But the tweaks mean that the learned distribution can certainly not be regarded as an approximate posterior based on optimizing the ELBO. Moreover, the tweaks amount to approximating a deterministic network.

4.1.3 Robustness

The radial posterior was more robust to hyperparameter variation (Figure 4). We assess robustness on a downsampled version of the diabetic retiopathy dataset (256x256) using a smaller model with a similar architecture to VGG-16, but which trained to convergence in about a tenth the time and had only ∼1.3M parameters. We randomly selected 86 different runs from plausible optimizer, learning rate, learning rate decay, batch size, number of variational samples per forward pass, and initial variance. 82% of hyperparameters tried for the MFVI baseline resulted in barely any improvement over randomly guessing, compared with 39% for the radial posterior. 44% of configurations for our radial posterior reached good AUCs, compared with only 11% for MFVI. This is despite the fact that we did allow models to pre-train the means using a negative log-likelihood loss for one epoch before begining ELBO training, a common tweak to improve MFVI.

[5] Other authors, e.g., Fortunato et al. [2018], achieve a similar result just by ignoring the KL to the prior in the loss so that the weight variances tend to shrink to overfit the training data.

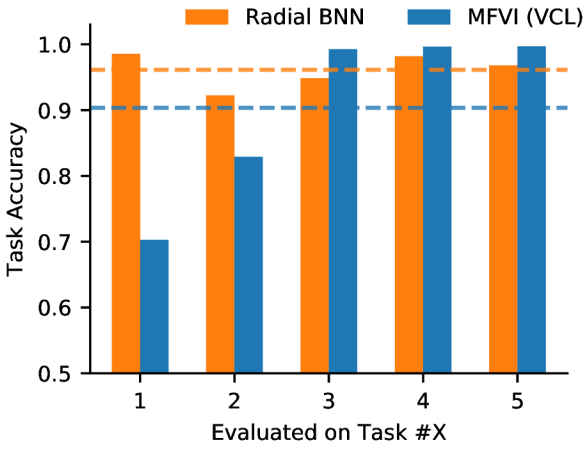

Figure 5: Models are trained on the five FashionMNIST tasks in sequence using the posterior from the previous task to remember *all* earlier tasks. Here we show performance of the final model, on all tasks. MFVI (as used in VCL) gradually forgets tasks—the final model's accuracy is worse the older the tasks get. Our Radial BNN preserves information and is still good at the first task. Average accuracy shown by the dotted line.

4.2 Continual Learning

Continual learning is a problem setting where a sequence of tasks must be learned separately while a single model is carried from one task to the next but all data are discarded [Kirkpatrick et al., 2017]. This is hard because neural networks tend to exhibit 'catastrophic forgetting' and lose performance on previous tasks. A number of authors have proposed prior-focused Bayesian approaches to continual learning in which the posterior at the end of learning a task becomes a prior when learning the next task [Kirkpatrick et al., 2017, Zenke et al., 2017, Chaudhry et al., 2018, Nguyen et al., 2018, Farquhar and Gal, 2018b, Ritter et al., 2018]. In the case of exact Bayesian updating,

this ought to balance the information learned from the datasets of each tasks. But for approximate methods, we have no such guarantee. The better the posterior approximation, the better we might expect such prior-focused Bayesian approaches to work.

Variational Continual Learning (VCL), by Nguyen et al. [2018], applies MFVI to learning the posterior. Here, we use VCL as a problem setting to evaluate the quality of the posterior. Note that we do not aim to solve the continual learning problem, but rather to demonstrate the improvement offered by Radial BNNs to the posterior approximation. A good posterior estimate should work as an effective prior and prevent forgetting. This setting is particularly relevant to variational inference, as other methods for estimating uncertainty in neural networks (such as Monte-Carlo dropout [Gal and Ghahramani, 2015] or ensembles [Lakshminarayanan et al., 2016]) cannot be straightforwardly used during training as a prior because the posteriors they learn assign zero probability to almost all weight values.

We consider a sequence of five tasks known as Split FashionMNIST [Nguyen et al., 2018, Farquhar and Gal, 2018a]. FashionMNIST is a dataset of images of items of clothing or attire (shoes, t-shirts, handbags etc.) [Xiao et al., 2017]. The first task is to classify the first two classes of FashionMNIST, then the next two etc. We examine a multi-headed model [Chaudhry et al., 2018, Farquhar and Gal, 2018a] in order to evaluate the quality of the posterior, although this is a limited version of continual learning. The models are BNNs with four hidden layers with 200 weights in each (∼250k parameters). We perform an extensive grid search over hyperparameters. Full hyperparameters and a more thorough description of the experimental settings, as well as results for the single-headed continual learning setting, are in Appendix D.2.

The Radial BNN approximate posterior acts as a better prior, showing that it learns the true posterior better (Figure 5). Radial BNNs maintain good accuracy on old tasks even after training on all five tasks. In contrast, the MFVI posterior gets increasingly less accurate on old tasks as training progresses. The MFVI posterior approximation is not close enough to the true posterior to carry the right information to the next task.

5 ANALYSING RADIAL BNNs

Why is it that Radial BNNs offer improved performance relative to MFVI? In §3 we observed that the multivariate Gaussian distribution typically used in MFVI features a 'soap-bubble'—almost all of the probability mass is clustered at a radius proportional to $\sigma\sqrt{D}$ from the mean in the large D limit (illustrated in Figure 1). This has two consequences in larger models. First, unless the weight variances are very small, a typical sample from the posterior has a high L_2 distance from the means. Second, because the mass is distributed uniformly over the hypersphere that the 'soap-bubble' clusters around, each sample from the multivariate Gaussian has a high expected L_2 distance from every other sample (similarly proportional to $\sigma\sqrt{D}$). This means that as σ and D grow, samples from the posterior are very different from each other, which we might expect to result in high gradient variance.

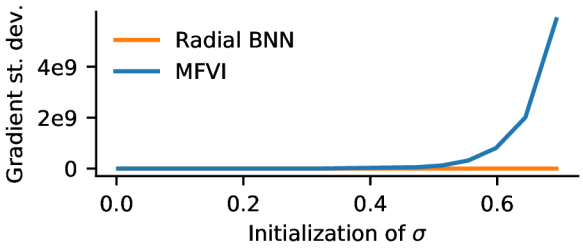

Figure 6: The variance of gradient estimates in the standard MFVI posterior explodes as the weight variance parameter grows.

In contrast, in Radial BNNs the expected distance between samples from the posterior is independent of D for the dimensionality typical of neural networks. The expected L_2 distance between samples from a unit hypersphere rapidly tends to $\sqrt{2}$ as the number of dimensions increases. Since the radial dimension is also independent of D, the expected L_2 distance between samples from the Radial BNN is independent of D. This means that, even in large networks, samples from the Radial BNN will tend to be more representative of each other. As a result, we might expect that the gradient variance is less of a problem.

Indeed, this is exactly what we find. In Figure 6, we show that for the standard MFVI posterior in a 3x3 conv layer with 512 channels, the variance of initial gradients explodes after the weight standard deviation exceeds roughly 0.3. This matters because, for MFVI with a unit Gaussian prior, the KL-divergence term of the loss is minimized by $\sigma_i = 1$—well within the region where gradient noise has exploded.

We can track this effect as it kicks in during training. In Figure 7 we show a sample training run on the down-sampled version of the diabetic retinopathy dataset using an MFVI and Radial BNN with the same hyperparameters. Pathologically, the *training* accuracy falls for MFVI after about 150 epochs (top graph). The critical moment corresponds to the point where the training process begins to optimize the prior cross-entropy term of the loss, sacrificing the negative log-likelihood term (middle graph). We can further show that this corresponds to the point where the standard deviation of the negative log-likelihood term of the gradient begins

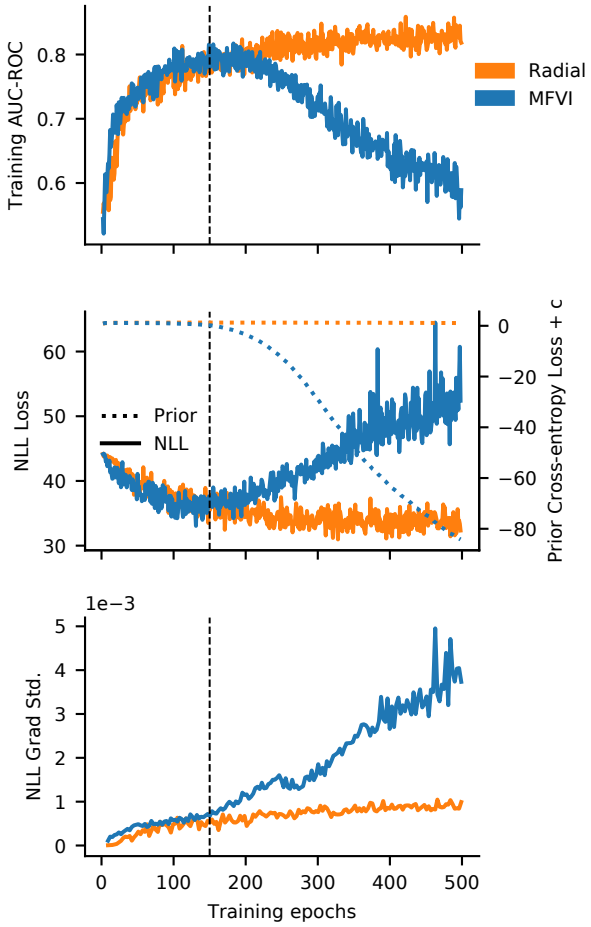

Figure 7: We can track the deterioration of the MFVI training dynamics. **Top**: After ∼150 epochs (dashed line) *training* set performance degrades for MFVI while Radial continues to improve. **Middle**: for MFVI, the NLL term of the loss *increases* during training, but the prior cross-entropy term falls faster so the overall loss continues to fall. **Bottom**: The standard deviation of the NLL gradient estimator grows sharply for MFVI after about 150 epochs. This coincides with the moment where the loss is optimized by minimizing the prior cross-entropy and sacrificing the NLL.

to sharply increase for MFVI. Meanwhile, the prior cross-entropy term is computed analytically, so its variance does not grow as the values of σ increase during training from their tiny initializations.

That is, MFVI fails because the high variance of the negative log-likelihood (NLL) term of the loss causes the optimizer to improve the cross-entropy term at the expense of the NLL term. For Radial BNNs, however, the NLL gradient variance stays low throughout training.

In Appendix F we offer further analysis of the failure of MFVI demonstrating that a biased but low-variance estimator of the gradient (using a truncated posterior approximation) improves training in MFVI.

5.1 Avoiding the Pathology in MFVI

Based on this analysis, we can see why Radial BNNs fix a sampling problem in MFVI. But this also helps explain why the ad-hoc tweaks which researchers have been using for MFVI have been successful. These tweaks chiefly serve to keep the weight variance low. Researchers initialize with small variances [Blundell et al., 2015, Fortunato et al., 2017, Nguyen et al., 2018]. Sometimes they adapt the loss function to remove or reduce the weight of the KL-divergence term, which reduces the pressure on weight variances to grow [Fortunato et al., 2018]. Other times researchers pretrain the means with just the NLL loss, which makes it possible to stop training after relatively little training on the ELBO loss, which stops the variances from growing too much [Nguyen et al., 2018]. Another approach, which we have not seen tried, would be to use a very tight prior, effectively enforcing the desire to have a basically deterministic network (a prior inversely proportional to $\sqrt{D}$ would balance the 'soap-bubble' variance). However, this sort of very tight prior is not compatible with the use of data-dependent priors in sequential learning.

For most of these tweaks, the resulting network is not fully optimizing the ELBO. This does not necessarily make the resulting network useless—after all, the ELBO is only a bound on the actual model evidence, and other methods like Deep Ensembles work surprisingly well despite not necessarily estimating the model posterior at all. However, if we have a theoretically principled way to fix our sampling problems without resorting to ad-hoc tweaks, then we should prefer that. Radial BNNs offer exactly that theoretically principled fix.

6 Discussion

Bayesian neural networks need to scale to large models in order to reach their full potential. Until now, researchers who want BNNs at scale needed to accept a posterior distribution which has zero probability mass almost everywhere—an implausible and problematic assumption. At the same time, MFVI requires increasingly demanding ad-hoc tweaks in order to work in anything but small models. We show why MFVI faces a serious gradient estimation problem which gets worse in high dimensions. Based on this motivation, we introduce Radial BNNs. This alternative variational inference posterior approximation is simple to implement, computationally fast, robust to hyperparameters, and scales to large models. Radial BNNs outperform other efficient BNN methods, and have the potential to craft data-dependent priors for use in applications like continual learning.

Acknowledgements

We would like to especially acknowledge Alexander Lvovsky for spotting an error in an early version of a proof and Lewis Smith for useful conversations and suggestion of the experiment in Appendix F. We would also like to thank Milad Alizadeh, Joost van Amersfoort, Gregory Farquhar, Angelos Filos, and Andreas Kirsch for valuable discussions and/or comments on drafts. In addition, we gratefully thank the Alan Turing Institute and Google for their donation of computing resources. Lastly, we thank the EPSRC for their support of Sebastian Farquhar through the Centre for Doctoral Training in Cyber Security at the University of Oxford.

References

Michael Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*, July 2018. 3

Chris Bishop. Introduction. In *Pattern Recognition and Machine Learning*. Springer, 2006. 1, 3, A

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. *Proceedings of the 32nd International Conference on Machine Learning*, 37:1613–1622, 2015. 1, 2, 3.1, 3.2, 5.1, D.1

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. *ECCV*, 2018. 4.2, D.2, D.3

Sebastian Farquhar and Yarin Gal. Towards Robust Evaluations of Continual Learning. *Lifelong Learning: A Reinforcement Learning Approach Workshop ICML*, May 2018a. 4.2, D.2, D.3

Sebastian Farquhar and Yarin Gal. A Unifying Bayesian View of Continual Learning. *Bayesian Deep Learning Workshop at NeurIPS*, 2018b. 4.2

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim G J Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. Benchmarking Bayesian Deep Learning with Diabetic Retinopathy Diagnosis. *Bayesian Deep Learning Workshop at NeurIPS*, 2019. 4.1

Meire Fortunato, Charles Blundell, and Oriol Vinyals. Bayesian Recurrent Neural Networks. *arXiv preprint*, 2017. 5.1

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Ian Osband, Alex Graves, Vlad Mnih, Remi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy Networks for Exploration. *ICLR*, 2018. 5, 5.1

Yarin Gal. Uncertainty in Deep Learning. *PhD Thesis*, 2016. B

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of the 33rd International Conference on Machine Learning*, 48: 1050–1059, 2015. 1, 2, 4.2, 2

Alex Graves. Practical Variational Inference for Neural Networks. *Neural Information Processing Systems*, 2011. 1, 2, 3.1

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CVPR*, 7(3):171–180, 2016. D.2

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. *Proceedings of the 32nd International Conference on Machine Learning*, 2015. 4

Geoffrey Hinton and Drew van Camp. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory*, 1993. 1, 2

Kaggle. Diabetic Retinopathy Detection, 2015. 4.1

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and Scalable Bayesian Deep Learning by Weight-Perturbation in Adam. *International Conference on Machine Learning*, 2018. 2

Diederik P. Kingma, Maz Welling, and Soumith Chintala. Auto-Encoding Variational Bayes. *International Conference on Learning Representations*, 2014. 3.1, B

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. 1, 4, 4.2

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. 2016. 1, 2, 4.2

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1), December 2017. 1, 4.1, 4.1.1, D.1

Christos Louizos and Max Welling. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. *International Conference on Machine Learning*, pages 1708–1716, 2016. 1, 2, 3.3, 2

David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 1992. 2

George Marsaglia. Choosing a Point from the Surface of a Sphere. *The Annals of Mathematical Statistics*, 43 (2):645–646, April 1972. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177692644. 3.1

Angel Muleshkov and Tan Nguyen. Easy Proof of the Jacobian for the N-Dimensional Polar Coordinates. *Pi Mu Epsilon Journal*, 14:269–273, 2016. B

Mervin E. Muller. A Note on a Method for Generating Points Uniformly on N-dimensional Spheres. *Commun. ACM*, 2(4):19–20, April 1959. ISSN 0001-0782. doi: 10.1145/377939.377946. 3.1

Radford M. Neal. *Bayesian Learning for Neural Networks*. PhD Thesis, University of Toronto, 1995. 2

Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational Continual Learning. *International Conference on Learning Representations*, 2018. 1, 4, 4.1.2, 4.2, 5.1, D.2

ChangYong Oh, Efstratios Gavves, and Max Welling. BOCK : Bayesian Optimization with Cylindrical Kernels. *Proceedings of The 35th International Conference on Machine Learning*, pages 3868–3877, June 2018. 6

Changyong Oh, Kamil Adamczewski, and Mijung Park. Radial and Directional Posteriors for Bayesian Neural Networks. *Bayesian Deep Learning Workshop at NeurIPS*, February 2019. 1, 2

Kazuki Osawa, Siddharth Swaroop, Anirudh Jain, Runa Eschenhagen, Richard E. Turner, Rio Yokota, and Mohammad Emtiyaz Khan. Practical Deep Learning with Bayesian Principles. *arXiv:1906.02506 [cs, stat]*, June 2019. 2

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the Convergence of Adam and Beyond. *International Conference on Learning Representations*, February 2018. D.2

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *Proceedings of The 31st International Conference on Machine Learning*, 2014. 3.1, B

Hippolyt Ritter, Aleksandar Botev, and David Barber. A Scalable Laplace Approximation for Neural networks. page 15, 2018. 4.2

Shengyang Sun, Changyou Chen, and Lawrence Carin. Learning Structured Weight Uncertainty in Bayesian Neural Networks. *Artificial Intelligence and Statistics*, pages 1283–1292, 2017. 1, 2, 3.3, 2

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational Bayesian Neural Networks. *International Conference on Learning Representations*, 2019. 1, 2, 2

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. *Proceedings of the 28th International Conference on Machine Learning*, 2011. 1, 2

Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E Turner, Jose Miguel Hernandez-Lobato, and Alexander L Gaunt. Deterministic Variational Inference for Robust Bayesian Neural Networks. *International Conference on Learning Representations*, 2019. 1, 2, 2

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv*, August 2017. 4, 4.2

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual Learning Through Synaptic Intelligence. *Proceedings of the 34th International Conference on Machine Learning*, pages 3987–3995, 2017. 4.2