On Symmetries in Variational Bayesian Neural Nets

Richard Kurle
AWS AI Labs
kurler@amazon.com

Tim Januschowski
AWS AI Labs
tjnsch@amazon.com

Jan Gasthaus
AWS AI Labs
gasthaus@amazon.com

Yuyang Wang
AWS AI Labs
yuyawang@amazon.com

Abstract

Probabilistic inference of Neural Network parameters is challenging due to the highly multi-modal likelihood functions. Most importantly, the permutation invariance of the neurons in the hidden layers renders the likelihood function unidentifiable with a factorial number of equivalent modes, independent of the data. We show that variational Bayesian methods that approximate the resulting highly multi-modal posterior by a uni-modal Gaussian distribution are biased towards approximations with identical (e.g. zero-centred) weights. This is in line with the commonly reported empirical observation that, in contrast to MCMC methods, variational approximations often collapse most weights to the typically zero-centred prior, resulting in severe underfitting. We propose a simple modification to the likelihood function that breaks the symmetry using fixed semi-orthogonal matrices as skip connections in each layer. Initial qualitative results show improved uncertainty estimation and reduced underfitting.

1 Introduction

The probabilistic approach to learning deep neural networks promises several appealing advantages compared to point estimates, s.a. reducing overfitting, estimating epistemic uncertainty [3], and enabling online/continual learning methods [5, 6]. Despite a growing scientific interest, Bayesian neural networks (BNN) are still rarely used in practical applications. Although variational methods s.a. Bayes by Backprop [1] scale well to larger models and datasets in terms of computation, they often result in severe underfitting, especially for small dataset sizes or large models [2]. We hypothesize that underfitting is primarily caused by the entropy term in the evidence lower bound (ELBO) for variational approximations that do not take into account the non-identifiability of the posterior.

Neural networks are known to be invariant wrt. a factorial number of permutations of the neurons in every hidden layer, resulting in an identical likelihood function for different parametrisations. For commonly used priors such as the standard normal distribution or mixtures of zero-centred diagonal Gaussians, the resulting posterior is therefore also a *symmetric mixture* with a factorial number of equivalent modes. We show that a hypothetical *symmetric mixture posterior approximation* that takes into account these symmetries provides a tighter ELBO than a uni-modal approximation, despite an equivalent likelihood. The gap between the corresponding ELBOs is tighter/looser if the symmetric modes of the posterior are overlapping/well separated. A notable special case in which all modes overlap is in case where the distribution over the weights is centred at zero, i.e. collapsing to the prior.

Reducing this additional bias in the ELBO objective is essential to avoid underfitting; this could potentially be achieved through one of the following (non-exclusive) routes:

- likelihood: 'hard' symmetry-breaking by modifying likelihood function.
- prior: 'soft' symmetry-breaking by assigning low probability to symmetric modes.
- variational: approximate the symmetric-mixture posterior.
- objective: modify the ELBO to reduce the bias, e.g. approximate the additional gap between single-mode and symmetric-mixture posterior.

Bayesian Deep Learning workshop, NeurIPS 2021.

In this work, we focus on breaking the permutation and sign-flip symmetries through skip connections with fixed semi-orthogonal matrices. We formulate the permutation invariance problem in Sec. 2, describe approach to break the symmetries in Sec. 3, and show initial results in Sec. 4.

2 Permutation symmetries and variational posterior approximation

Neural networks are known to be invariant wrt. permutations of the neurons in every hidden layer. This implies that the likelihood function of neural networks is non-identifiable and has no global optimum. For maximum likelihood/a posteriori point estimation through stochastic gradient descent, this does not pose a significant problem since any of the equivalent optima suffices. However, the non-identifiability complicates probabilistic inference, especially the variational Bayesian approach.

2.1 Symmetric-mixture posterior

Consider a multi-layer perceptron (MLP) with L hidden layers. Written in pre-activations form, each layer indexed by l, computes the representation

$$h_l = W_l f_l(h_{l-1}) + b_l \tag{1}$$

using weights $W_l \in \mathbb{R}^{N_l \times N_{l-1}}$, biases $b_l \in \mathbb{R}^{N_l}$ and element-wise activation functions f_l. The first activation is the input data, that is, $h_0 := x$ and f_1 is the identity function. This parametrisation is invariant to permutations of the neurons in each hidden layer l. The permutation of the neurons can be written in terms of permutations to the incoming and outgoing weights:

$$\begin{aligned} h_{l+1} &= W_{l+1} f_{l+1}(P_l^T P_l h_l) + b_{l+1} \\ &= \underbrace{W_{l+1} P_l^T}_{W'_{l+1}} f_{l+1}(\underbrace{P_l W_l}_{W'_l} f_l(h_{l-1}) + \underbrace{P_l b_l}_{b'_l}) + b_{l+1}, \end{aligned} \tag{2}$$

where $P_l \in \mathbb{R}^{N_l \times N_l}$ is a permutation matrix for which each row consists of all 'zeros' except a single 'one'. The permutation invariance follows from the activation functions being applied element-wise s.t. the permutation matrix P_l^T in Eq. (2) can be "pulled out" of f_{l+1}. W'_{l+1}, W'_l and b'_l then denote the weights and biases corresponding to one of the equivalent modes. For every hidden layer l, there are $N_l!$ possible permutations, totalling $\prod_{l=1}^{L} N_l!$ equivalent modes.[1]

In the Bayesian approach, we put a prior $p(w)$ over the weights and biases of the MLP and aim to infer the posterior $p(w|\mathcal{D}) \propto p(w)p(\mathcal{D}|w)$ given a dataset $\mathcal{D}$. For common priors such as a Gaussian or a Mixture of Gaussian (MoG) with the mean(s) centred at the origin and diagonal covariance(s), the posterior incurs the factorial number of modes from the likelihood function $p(\mathcal{D}|w)$. The resulting posterior is thus a symmetric mixture distribution consisting of $N = \prod_{l=1}^{L} N_l!$ mixture components

$$p(w|\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} p_n(w|\mathcal{D}). \tag{3}$$

If we are interested in predictions only, it suffices to approximate a single mode $p_n(w|D)$, e.g. through markov chain monte carlo methods. Variational Bayesian methods however compute the entropy of the approximate posterior. In the subsequent section, we describes how the entropy of a variational approximation that does not take into account the model degeneracy causes underfitting.

2.2 Variational Bayesian approximation

Consider a simple diagonal Gaussian variational approximations for the posterior of the weights and biases of the neural network $q_{\theta_0}(w) = \mathcal{N}(w; \mu_0, \Sigma_0)$, $\theta_0 = \{\mu_0, \Sigma_0\}$. Inference amounts to maximizing the ELBO

$$\begin{aligned} \mathcal{L}_{\text{Gauss}}(\mathcal{D}, \theta) = \mathbb{E}_{q_{\theta_0}(w)}[\log p(\mathcal{D}|w) + \log p(w)] \\ - \mathbb{E}_{q_{\theta_0}(w)}[\log q_{\theta_0}(w)]. \end{aligned} \tag{4}$$

[1] Note that, if f_l was linear, the degeneracy would include all invertible transformations. If f_l is symmetric wrt. the origin, flipping the signs of in- and outgoing weights (i.e. a "-1" in P_l) provides 2^{N_l} further sign-flip symmetries. Furthermore, piece-wise linear activations such as ReLU result in continuous scaling symmetries.

Notice again that the likelihood term is invariant wrt. the permutations in $q_{\theta_0}(w)$ as shown in Eq. (2). Similarly, for zero-centred symmetric priors, the term involving the prior is also identical for each of the symmetric modes.

Consider now a *symmetric-MoG* posterior approximation $q_\theta(w) = \frac{1}{N} \sum_{n=1}^{N} q_{\theta_n}(w)$, where $q_{\theta_0}(w)$ is a diagonal Gaussian as before, and the other symmetric modes are given by the permutations from Eq. (2). In the corresponding ELBO objective, the terms related to the model take the posterior expectations over only one of the modes $q_{\theta_0}(w)$:

$$\begin{aligned}\mathcal{L}_{\mathrm{MoG}}(\mathcal{D}, \theta) = \mathbb{E}_{q_{\theta_0}(w)}[\log p(\mathcal{D}|w) + \log p(w)] \\ - \mathbb{E}_{q_\theta(w)}[\log q_\theta(w)].\end{aligned} \tag{5}$$

For the likelihood term, each of the symmetric modes computes an identical function by the definition of the non-identifiability/permutation-invariance in Eq. (2), that is $\mathbb{E}_{q_{\theta_n}(w)}[\log p(\mathcal{D}|w)] = \mathbb{E}_{q_{\theta_0}(w)}[\log p(\mathcal{D}|w)],\ \forall n$. Similarly, $\mathbb{E}_{q_{\theta_n}(w)}[\log p(w)] = \mathbb{E}_{q_{\theta_0}(w)}[\log p(w)],\ \forall n$ in case of the isotropic Gaussian prior or a MoG prior with zero-centred means and diagonal covariance.

Comparing the ELBO corresponding to the diagonal Gaussian and a hypothetical *symmetric-MoG* posterior approximation, we note that they differ only by the entropy term, which can be quantified as

$$\begin{aligned}\mathcal{L}_{\mathrm{MoG}}(\mathcal{D}, \theta) - \mathcal{L}_{\mathrm{Gauss}}(\mathcal{D}, \theta) &= \mathbb{E}_{q_{\theta_0}(w)}[\log q_{\theta_0}(w)] - \mathbb{E}_{q_\theta(w)}[\log q_\theta(w)] \\ &= \mathbb{E}_{q_{\theta_0}(w)}\left[\log q_{\theta_0}(w) - \log \frac{1}{N} \sum_{n=1}^{N} q_{\theta_n}(w)\right] \\ &= \mathrm{KL}\left[q_{\theta_0}(w) \,||\, q_\theta(w)\right].\end{aligned} \tag{6}$$

Here we made use of the symmetry again by taking the expectation in the neg. entropy of the MoG over q_{θ_0} only. The resulting KL divergence is bounded between zero and the log of the number of modes, $0 \leq \mathrm{KL}\left[q_{\theta_0}(w) \,||\, q_\theta(w)\right] \leq \log \prod_{l=2}^{L} N_l!$.

Unsurprisingly, the ELBO corresponding to the more complex *symmetric-MoG* approximation provides a tighter lower bound. The non-identifiability of the neural network likelihood induces an additional gap in the ELBO if the symmetries are not considered in the posterior approximation. This gap between the two corresponding ELBOs can be significant even if the posterior is locally Gaussian, and, most importantly, this gap is not a constant: i) the KL is minimised if the components of the MoG are identical and thus overlap to yield single Gaussian; ii) the KL is maximised if the components of the MoG are well separated. The components of the MoG are identical if they are centred at zero, similar to the prior. It follows that the Gaussian posterior approximation provides an almost equally tight bound compared to the *symmetric-MoG* if most weights collapse to the prior. We therefore hypothesise that the variational Bayesian approach to inference in BNNs causes most weights in the approximate posterior to collapse to the zero-centred prior.

3 Symmetry-breaking through skip connections

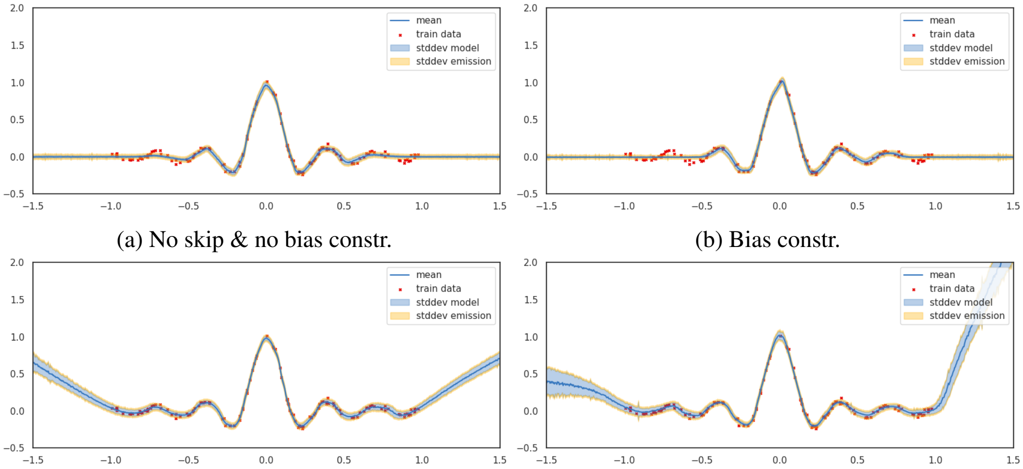
We address the degeneracy/symmetry problem by modifying the likelihood function s.t. the modes are no longer equivalent. Previous work enforces a bias-ordering constraint, $b_l^{(1)} \leq b_l^{(2)} \leq \ldots \leq b_l^{(N_l)}$, by parametrising the log-differences between the scalar biases [8]. However, if the biases take (near) zero values, the degeneracy remains (mostly) intact. Here, we modify each layer to include skip connections with *fixed* matrices O_l:

$$h_l = O_l h_{l-1} + W_l f_l\left(h_{l-1}\right) + b_l. \tag{7}$$

With this simple modification of the likelihood function, it is not possible to permute the neurons/activations, since only the corresponding weight parameters are inferred variables, but the matrices O_l remain fixed. Omitting the biases for simplicity,

$$\begin{aligned}h_l = O_l h_{l-1} + f_l\left(P_l^T P_l W_l h_{l-1}\right) &= O_l h_{l-1} + P_l^T f_l\left(P_l W_l h_{l-1}\right) \\ &\neq P_l^T\left(O_l h_{l-1} + f_l\left(P_l W_l h_{l-1}\right)\right).\end{aligned} \tag{8}$$

The last line would be needed if we want to 'group' P^T with the subsequent layer's weights W_{l+1}.



(a) No skip & no bias constr.

(b) Bias constr.

(c) Orthogonal skip

(d) Ortho. skip & bias constr.

Figure 1: Predictive distribution (1 stddev) for sinc-function regression. The model is an MLP with 5 layers of 32 units and SELU activation functions. The dataset consists of 128 samples (marked red), drawn uniformly in $[-1, 1]$, mapped through the (scaled) sinc-function with additive noise with std. dev 0.2. The stddev is separated in model/epistemic uncertainty given by the randomness of the weights, and emission/aleatoric uncertainty given by the likelihood noise model.

It has been shown previously that models with residual connections break the symmetry and argued that this improves the learning dynamics in ResNets [7]. Since identity residual connections can be used only for layers with the same number of neurons, we instead use *fixed random (semi-) orthogonal* matrices, i.e. non-square matrices for which the rows or columns are orthonormal.

4 Experiments

We trained an MLP with 5 hidden layers of 32 units each on 128 data points from a (scaled) sinc-function with additive noise. We used an isotropic Gaussian prior and a diagonal Gaussian posterior approximation. Inference is performed using the local reparameterisation trick [4] with 64 samples for training and 256 for testing. The model is trained for 200.000 full-batch iterations, with a linear annealing schedule for the KL divergence in the first half of the iterations. The predictive distributions for a standard BNN, a model with a bias ordering constraint, our proposed orthogonal skip connections, as well as the combination of both, are shown in Fig. 1. It can be seen that the baseline results in severe underfitting, while orthogonal skip connections fit the data well and provide better out of distribution uncertainty.

References

[1] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, pages 1613–1622, 2015.

[2] S. Ghosh, J. Yao, and F. Doshi-Velez. Structured variational learning of Bayesian neural networks with horseshoe priors. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1744–1753, 2018.

[3] Alex Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017.

[4] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[5] Richard Kurle, Botond Cseke, Alexej Klushyn, P. V. D. Smagt, and Stephan Günnemann. Continual learning with bayesian neural networks for non-stationary data. In *ICLR*, 2020.

[6] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

[7] Emin Orhan and Xaq Pitkow. Skip connections eliminate singularities. In *ICLR*, 2018.

[8] A. Pourzanjani, Richard M. Jiang, and L. Petzold. Improving the identifiability of neural networks for bayesian inference. 2017.