On the detrimental effect of invariances in the likelihood for variational inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

Variational Bayesian inference of the posterior requires simplifying approximations such as mean-field parametrisation to ensure tractability for practically relevant models with many parameters. However, prior work has associated the variational mean-field approximation for Bayesian neural networks with collapses to the prior in the case of small datasets or large model sizes. In this work, we show that invariances in the likelihood function of over-parametrised models such as permutation of neurons in the same layer in the case of Bayesian neural networks or other cases such as translation invariances contribute to this phenomenon because these invariances complicate the structure of the posterior. They introduce discrete finite and/or infinite continuous modes which cannot be well approximated by mean-field distributions. In particular, we show that the mean-field approximation has a gap in the evidence lower bound objective compared to a purpose-built posterior that takes into account the considered invariances. Importantly, this invariance gap is not constant; it vanishes as the approximation collapses to the prior. In our approach, we proceed by first considering translation invariances in a linear model with a single data point in detail. We show that, while the true posterior can be constructed from a mean-field parametrisation, this is achieved only if the objective function takes into account the invariance gap. Then, we transfer our analysis of the linear model to neural networks. Our analysis provides a framework for future work to explore solutions to the posterior collapse problem.

1 Introduction

Bayesian neural networks (BNNs) have several appealing advantages compared to deterministic neural networks (NN) such as improving generalization [33], capturing epistemic uncertainty [15], and providing a framework for continual learning methods [16, 20]. Unfortunately, repeating this theoretical benefits has thus far been impeded [32]. In particular, several practical issues with variational Bayesian inference methods—which are the de-facto standard technique for scaling inference in BNNs to large datasets [13, 1]—have been identified to be partially responsible for this performance gap [14]. Indeed, mean-field variational BNNs (VBNN) suffer from severe underfitting for large models and small dataset sizes [11]. Recent work [4] showed that under certain assumptions such as an odd Lipschitz activation function and a finite dataset with bounded likelihood, the optimal mean-field approximation *collapses* to the prior as the NN width increases, *ignoring the data*.

In this work, we study invariants of the likelihood function and the resulting non-identifyability in overparametrised models. We identify it as the primary reason why mean-field variational distributions cannot approximate the posterior, ultimately resulting in the reversion to the prior analysed in [4]. It is well known that NNs exhibit several invariances *with respect to their parameters*, including node permutation invariance and sign-flip invariance (in case of odd activation functions) [25, 2], and scaling invariance in case of piece-wise linear activations [22]. Recent work showed that the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

permutation invariance introduces an additional gap in the evidence lower bound (ELBO) for a mean-field approximation compared to an (intractable) mixture that includes all modes [17]. We show that the parameter space of BNNs additionally has subspaces with *translation invariance*.

Invariance in the likelihood function does not necessarily pose a challenge if maximum likelihood point estimation through stochastic gradient descent is used, because convergence to any of the equivalent optima (resulting from the invariance) suffices for prediction. However, we show that the variational Bayesian approach is particularly impeded by invariances in the likelihood, since the resulting posterior has a complicated structure that can not be well approximated by the mean-field approximation. In order to isolate the impact of the invariances, we define a posterior approximation in Sec. 3 that explicitly models the invariances by integrating over all transformations that leave the likelihood invariant. We *construct* this *invariance-abiding* distribution from a *mean-field likelihood* approximation such that it induces an identical posterior predictive distribution compared to an alternative *mean-field posterior* approximation that is agnostic to the invariances. Both approximations are parametrised with the same variational parameters and result in the same expected log-likelihood, but differ in their modelling of the invariances, and, hence, the corresponding ELBO objective differs only in the KL regularisation term. We refer to the difference between these two KL terms as the *invariance-gap*. While our constructed approximation is intractable in general, it serves as a useful tool for studying the invariance in simple models where it is indeed tractable, and it provides a framework for future work to devise practical remedies.

The main contribution of this work is a detailed analysis (Sec. 4) of a linear model that implements *translation invariance* and explains the reversion of the posterior to the prior as the result of optimising the ELBO that exhibits a non-trivial invariance-gap. This gap vanishes for approximations that collapse to the prior and grows linearly in the number of dimensions with translation invariance for mean-field approximations that have the same posterior predictive distribution as the true posterior. We transfer our analysis of the linear model to VBNNs in Sec. 4.3, showing that subspaces in BNNs exist that exhibit translation invariance. We start by introducing the basic concepts and recent results on which we build our contribution in Sec. 2.

2 Background

2.1 Variational Bayesian Neural Networks

Neural network functional model: Deep neural networks are layered models that progressively transform their inputs in each layer. More formally, an L–layer neural network with $n_1, \ldots, n_L$ many hidden units is algebraically computing

$$f(\mathbf{x}) = h_L\left(\mathbf{w}_{L,1}^{\mathrm{T}}\mathbf{z}_{L-1}\right) , \quad z_{1,i} = h_1\left(\mathbf{w}_{1,i}^{\mathrm{T}}\mathbf{x}\right) , \quad z_{l,i} = h_l\left(\mathbf{w}_{l,i}^{\mathrm{T}}\mathbf{z}_{l-1}\right) ,$$

where the $h_l : \mathbb{R} \to \mathbb{R}$ are monotonic transfer functions which introduce non-linearities. Such a network has $n_1 + n_2 + \cdots + n_L$ many nodes $z_{l,i}$ and y which are each a weighted linear combination of either the input vector $\mathbf{x}$ (for the first hidden layer) or the value of the hidden units $\mathbf{z}_l$ (for all the other hidden layers and the output $f(\mathbf{x})$). We use $\mathbf{w} = [\mathbf{w}_1^{\mathrm{T}}, \ldots, \mathbf{w}_{n_1+\cdots+n_L}^{\mathrm{T}}]^{\mathrm{T}}$ to denote all learnable parameters of the model.

Variational Bayesian treatment: If the weights and biases $\mathbf{w}$ are treated as random variables with a prior distribution $p(\mathbf{w})$ and a posterior $p(\mathbf{w} \,|\, \mathcal{D})$ induced by the dataset $\mathcal{D}$ through the likelihood function $l(\mathbf{w}) := p(\mathcal{D} \,|\, \mathbf{w})$ (defined via $f(\mathbf{x})$), these models are referred to as Bayesian neural networks (BNNs). The variational Bayesian method approximates the posterior by a distribution $q_{\boldsymbol{\theta}}(\mathbf{w}) \approx p(\mathbf{w} \,|\, \mathcal{D})$ with variational parameters $\boldsymbol{\theta}$, casting inference as an optimisation problem

$$q_{\boldsymbol{\theta}}^*(\mathbf{w}) = \underset{q_{\boldsymbol{\theta}} \,\in\, \mathcal{Q}}{\operatorname{argmin}} \; \mathrm{KL}\left[q_{\boldsymbol{\theta}} \,||\, p(\cdot \,|\, \mathcal{D})\right] , \tag{1}$$

where $\mathrm{KL}\left[q_{\boldsymbol{\theta}} \,||\, p(\cdot \,|\, \mathcal{D})\right] := \mathbb{E}_{\mathbf{w}\sim q_{\boldsymbol{\theta}}}\left[\ln(q_{\boldsymbol{\theta}}(\mathbf{w}))/(p(\mathbf{w} \,|\, \mathcal{D}))\right]$ and $\mathcal{Q}$ is a family of distributions over $\mathbf{w}$. In case of Variational Bayesian neural networks (VBNNs) this is typically a Gaussian distribution with diagonal covariance matrix. The optimisation of (1) is achieved by maximising a lower bound to the model evidence $\ln p(\mathcal{D})$,

$$\begin{aligned} \mathcal{L}_{\mathrm{ELBO}}\left(q_{\boldsymbol{\theta}}, \mathcal{D}\right) = \ln p(\mathcal{D}) - \mathrm{KL}\left[q_{\boldsymbol{\theta}} \,||\, p(\cdot \,|\, \mathcal{D})\right] &= \mathbb{E}_{\mathbf{w}\sim q_{\boldsymbol{\theta}}}\left[\ln \frac{p(\mathcal{D} \,|\, \mathbf{w})\, p(\mathbf{w})}{q_{\boldsymbol{\theta}}(\mathbf{w})}\right] \\ &= \mathrm{ELL}\left(q_{\boldsymbol{\theta}},\, \mathcal{D}\right) - \mathrm{KL}\left[q_{\boldsymbol{\theta}} \,||\, p\right] , \end{aligned} \tag{2}$$

where $\mathrm{ELL}\left(q_{\boldsymbol{\theta}},\, \mathcal{D}\right) := \mathbb{E}_{\mathbf{w}\sim q_{\boldsymbol{\theta}}}\left[\ln p(\mathcal{D}\,|\,\mathbf{w})\right]$ is the expected log-likelihood.

2.2 Data-related bound on the KL divergence

The term $\mathrm{KL}\left[q_{\boldsymbol{\theta}}\,||\,p\right]$ in (2) admits a data-dependent upper bound that naturally occurs as a consequence of the finite information provided by a finite dataset in the presence of noise. This result has been shown in [4] and we recall the result for a Gaussian likelihood with homogeneous noise (for other likelihood functions, we refer to App. F in [4]).

Assume a VBNN with Gaussian observation noise defined through $y = f_{\mathbf{w}}(\mathbf{x}) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma_y)$; an isotropic Gaussian prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{0}, \mathbf{I})$; and a mean-field variational approximation $q_{\boldsymbol{\theta}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}, \mathrm{diag}(\mathbf{v}))$. Define $\sigma_{k+1}\big(\mathbf{x}^{(n)}\big) := \mathbb{V}_{\mathbf{w}\sim p}\left[f_{\mathbf{w}}(\mathbf{x}^{(n)})\right]$, for some fixed input $\mathbf{x}^{(n)}$. Then

$$\mathrm{KL}\left[q_{\boldsymbol{\theta}}\,||\,p\right] \leq \mathrm{ELL}\left(q^*, \mathcal{D}\right) - \mathrm{ELL}\left(p, \mathcal{D}\right) = \sum_{n=1}^{N} \frac{\sigma_{k+1}\big(\mathbf{x}^{(n)}\big) + \big(y^{(n)}\big)^2}{2\sigma_y}. \tag{3}$$

See App. F of the supplementary material for details and the derivation. Prior work [4] has used the data-related bound to prove *that* the predictive distribution of mean-field BNNs converges to the prior predictive distribution if the network width is large and the activation function is odd, ultimately resulting in posterior collapse to the prior. In contrast, we address the question *why* the mean-field approximation cannot approximate the posterior by relating this data-related bound to invariances in neural network functions in Sec. 3.

3 Invariance-abiding variational approximation

We develop a framework to model the invariances in the likelihood and helps understanding their impact on the ELBO objective. In order to achieve this, we approximate the likelihood by a Gaussian function with variational parameters and marginalise over all transformations to which the true likelihood is invariant. By taking the product with the prior, we construct an *invariance-abiding* posterior approximation that can be related to a standard *mean-field* approximation that does not explicitly model the considered invariances. We then describe conditions under which both approximations yield an identical posterior predictive, while the KL regularisation term of the mean-field approximation is lower bounded by the KL of the invariance-abiding approximation.

Likelihood and Posterior approximations: Non-identifiability implies that the posterior does not concentrate on a single set of parameters irrespective of the dataset size [10], because the same likelihood is assigned to different parameters irrespective of the data. We model this invariance through transformations $t(\cdot, \mathbf{r})$ to which the likelihood is invariant with D-dimensional variables $\mathbf{r}$:

$$\forall \mathbf{r} \sim p(\mathbf{r}): \quad p(\mathcal{D}\,|\,t(\mathbf{w}, \mathbf{r})) = p(\mathcal{D}\,|\,\mathbf{w}). \tag{4}$$

We assume that each of the equivalent parametrisations of the likelihood $t(\mathbf{w}, \mathbf{r})$ has the same probability a priori. For a finite set of discontinuous transformations such as permutation invariance (see App. E in the supplementary material), $p(\mathbf{r})$ is a uniform distribution over discrete variables that index the resepctive transformations; for continuous transformations such as translation invariance (see 4, we model the uniform distribution as a Gaussian with infinite variance. From (4) it follows that the likelihood is also invariant wrt. the expectation over all transformations,

$$p(\mathcal{D}\,|\,\mathbf{w}) = \mathbb{E}_{\mathbf{r}\sim p}\left[p(\mathcal{D}\,|\,t(\mathbf{w}, \mathbf{r})\right].$$

It is conceivable that the likelihood can be constructed by marginalising over these transformations with a simpler function $l(t(\mathbf{w}, \mathbf{r}))$ with $p(\mathcal{D}\,|\,\mathbf{w}) = \mathbb{E}_{\mathbf{r}\sim p}\left[l(t(\mathbf{w}, \mathbf{r}), \mathcal{D})\right]$ that is easier to approximate. We define a function with variational parameters $\boldsymbol{\theta}$ to approximate a single mode $g_0(\mathbf{w}; \boldsymbol{\theta}) \approx l(\mathbf{w}, \mathcal{D})$, and model a transformation-invariant likelihood approximation from the same parametrisation through

$$g_{\mathrm{mix}}(\mathbf{w}; \boldsymbol{\theta}) := \mathbb{E}_{\mathbf{r}\sim p}\left[g_0(t(\mathbf{w}, \mathbf{r}); \boldsymbol{\theta})\right] \approx \mathbb{E}_{\mathbf{r}\sim p}\left[l(t(\mathbf{w}, \mathbf{r}), \mathcal{D})\right] = p(\mathcal{D}\,|\,\mathbf{w}). \tag{5}$$

We consider factorising Gaussians for the prior $p(\mathbf{w})$ and likelihood approximation $g(\mathbf{w}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\mathbf{m}, \boldsymbol{\lambda}\}$ are means and variances. We then define a standard *mean-field* posterior as the product

$$q_0(\mathbf{w}; \boldsymbol{\theta}) = Z_0^{-1}\, p(\mathbf{w}) \cdot g_0(\mathbf{w}; \boldsymbol{\theta}), \qquad Z_0 = \int p(\mathbf{w}) \cdot g_0(\mathbf{w}; \boldsymbol{\theta}) d\mathbf{w}. \tag{6a}$$

Similarly, we define an *invariance-abiding* posterior as the product of prior and *invariant* likelihood:

$$q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta}) = Z_{\text{mix}}^{-1}\, p(\mathbf{w}) \cdot g_{\text{mix}}(\mathbf{w};\boldsymbol{\theta}), \qquad Z_{\text{mix}} = \int p(\mathbf{w}) \cdot g_{\text{mix}}(\mathbf{w};\boldsymbol{\theta}) d\mathbf{w}. \tag{6b}$$

Consequently, $q_0(\mathbf{w};\boldsymbol{\theta})$ is also a mean-field approximation, while the invariance-abiding approximation $q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta})$ is a mixture with infinite continuous or finite discrete modes depending on the type of invariance. We describe the discrete node permutation invariance in BNNs in App. E of the supplementary material, while we consider continuous translation invariance in Sec. 4.

Posterior predictive equivalence: Next, we discuss conditions under which we can construct approximations $q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta})$ and $q_0(\mathbf{w};\boldsymbol{\theta})$ that yield the same predictive distribution. First, we must be able to write each product density $p(\mathbf{w}) \cdot g_0(t(\mathbf{w},\mathbf{r}))$ in the expectation over $p(\mathbf{r})$ as a function of $t(\cdot,\cdot)$ s.t. each component in this mixture has the same invariance as the likelihood function:

$$\begin{aligned} q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta}) = Z_{\text{mix}}^{-1}\, p(\mathbf{w}) \cdot \int p(\mathbf{r})\, g_0(t(\mathbf{w},\mathbf{r}),\boldsymbol{\theta}) d\mathbf{r} &= Z_{\text{mix}}^{-1} \int p(\mathbf{r})\, p(\mathbf{w}) \cdot g_0(t(\mathbf{w},\mathbf{r}),\boldsymbol{\theta}) d\mathbf{r} \\ &\overset{!}{=} \int p(\mathbf{r})\, \frac{Z_0(\mathbf{r})}{Z_{\text{mix}}}\, q_0(t(\mathbf{w}, f(\mathbf{r}));\boldsymbol{\theta}) d\mathbf{r}, \end{aligned} \tag{7}$$

where $Z_{\text{mix}} = \int p(\mathbf{r}) Z_0(\mathbf{r}) d\mathbf{r}$, and $f : \mathbb{R}^D \to \mathbb{R}^D$. Since $p(\mathbf{r})$ is uniform, $\mathbf{r}' \sim p(f(\mathbf{r}))$ is contained in the set of possible values that $\mathbf{r}$ can take. Then, we can use $\ln p(\mathcal{D} \,|\, \mathbf{w}) = \ln p(\mathcal{D} \,|\, t(\mathbf{w}, f(\mathbf{r}))$. Second, all invariance transformations $t(\cdot,\mathbf{r})$ must be volume-preserving, i.e.

$$\forall \mathbf{r} \sim p(\mathbf{r}) : \; \left| \det \frac{\partial t(\mathbf{w},\mathbf{r})}{\partial \mathbf{w}} \right|^{-1} \overset{!}{=} 1. \tag{8}$$

Although (7) and (8) are fairly restricting, common invariances such as translation and permutation invariance fulfil this condition for Gaussian priors and likelihood approximations. We provide further discussion in App. C of the supplementary material. With (7) and (8), we can also show that

$$\mathbb{E}_{\mathbf{w}\sim q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta})} \left[\ln p(\mathcal{D} \,|\, \mathbf{w})\right] = \mathbb{E}_{\mathbf{w}\sim q_0(\mathbf{w};\boldsymbol{\theta})} \left[\ln p(\mathcal{D} \,|\, \mathbf{w})\right]. \tag{9}$$

Invariance gap: If the conditions in (7) and (8) are met, we can also show that

$$\mathcal{L}_{\text{ELBO}}\left(q_0, \mathcal{D}\right) - \mathcal{L}_{\text{ELBO}}\left(q_{\text{mix}}, \mathcal{D}\right) = \text{KL}\left[q_0 \,||\, p\right] - \text{KL}\left[q_{\text{mix}} \,||\, p\right] = \text{KL}\left[q_0 \,||\, q_{\text{mix}}\right]. \tag{10}$$

See App. C in the supplementary material for the derivations. Interestingly, the gap between the two respective ELBO objectives is given exactly by the relative entropy between the mean-field and invariance-abiding approximation; we therefore refer to this difference as the *invariance gap*. Now consider the case when the likelihood (approximation) is uninformative, i.e. as $g_0(\mathbf{w};\, \boldsymbol{\theta})$ and consequently $g_{\text{mix}}(\mathbf{w};\, \boldsymbol{\theta})$ is uniform. Then it follows that

$$\lim_{\boldsymbol{\lambda}\to\infty} q_0(\mathbf{w};\boldsymbol{\theta}) = p(\mathbf{w}), \qquad \lim_{\boldsymbol{\lambda}\to\infty} q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta}) = p(\mathbf{w}),$$

and the KL regularisation term for both approximations converges to zero. We thus know that, if the *mean-field posterior reverts* to the prior, the *invariance gap vanishes*, i.e.

$$\lim_{\boldsymbol{\lambda}\to\infty} \text{KL}\left[q_0 \,||\, q_{\text{mix}}\right] = 0.$$

Implication of data-related bound: Since $\text{KL}\left[q_0 \,||\, p\right]$ is upper bounded by the best- and worst-case ELL as described in Sec. 2.2, the invariance gap is also bounded (using (10) and (3)):

$$\text{KL}\left[q_0(\cdot;\boldsymbol{\theta}) \,||\, q_{\text{mix}}(\cdot;\boldsymbol{\theta})\right] \leq \text{KL}\left[q_0(\cdot;\boldsymbol{\theta}) \,||\, p\right] \leq \text{ELL}\left(q^*, \mathcal{D}\right) - \text{ELL}\left(p, \mathcal{D}\right). \tag{11}$$

Consequently, this bound puts a constraint on the achievable solutions to the maximisation of the ELBO objective wrt. the variational parameters of the mean-field approximation $q_0(\mathbf{w};\boldsymbol{\theta})$ One particular parametrization that guarantees this behaviour is the mean-field posterior collapse, $q_0(\mathbf{w};\boldsymbol{\theta}) \approx p(\mathbf{w})$ introduced above. However, in order to show that the invariance-gap not only admits posterior collapse as a potential parametrization but indeed incentivises posterior collapse, we next tackle the question how the invariance-gap grows for non-collapsed approximations.

We observe in the subsequent sections that the invariance-gap increases with the model size for non-collapsed approximations and the invariances we consider. Examples of the invariance

include permutation invariance (interchangeability of neurons in the same layer of a BNN) in the supplementary material. We show that $\mathrm{KL}\left[q_0 \,||\, q_{\mathrm{mix}}\right]$ is between zero and the logarithm of the number of discrete permutation-induced modes, which grows factorial with the number of neurons per layer. Another example and the focus of the main body of this work is continuous translation invariance. In Sec. 4), we analyse translation invariance in over-parametrised linear models, show that the gap grows linear with the number of dimensions (weights) for parameter values θ^*_{mix} that are optimal in terms of the posterior predictive and the ELBO optimised wrt. $q_{\mathrm{mix}}(\mathbf{w},\, \boldsymbol{\theta})$.

4 Translation invariance in linear models

We study an over-parametrised Bayesian linear regression model as the canonical model that implements *translation invariance.* This models serves as a useful tool to understand the detrimental effect of translation invariance, since all interesting quantities can be computed analytically, incl. the mean-field and invariance-abiding distribution defined in (6), and the invariance gap from (10).

Likelihood model: Consider a linear model with K latent variables $\mathbf{w} = [w_1, \ldots, w_K]^{\mathrm{T}}$ and assume that we only have observations of variables y given dependent inputs $\mathbf{x}$. To simplify the setting, we will assume that $\mathbf{x} = \mathbf{1}$ (see App. D in the supplementary material for the more general case). We further assume that the observation depends only on the inner product with the inputs and additive Gaussian noise. That is,

$$y = \frac{1}{K}\mathbf{1}^{\mathrm{T}}\mathbf{w} + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma_y\right). \tag{12}$$

Note that any change in $\mathbf{w}$ which leaves the sum of the elements unaffected does not change the likelihood. In general, we can model this translation invariance using a $K-1$ dimensional vector $\boldsymbol{\Delta} \in \mathbb{R}^{K-1}$ and observing that

$$\mathbf{1}^{\mathrm{T}}\mathbf{w} = \mathbf{1}^{\mathrm{T}}\left(\mathbf{w} + \mathbf{B}\boldsymbol{\Delta}\right), \quad \mathbf{B} := \begin{bmatrix} \mathbf{I} \\ -\mathbf{1}^{\mathrm{T}} \end{bmatrix}, \tag{13}$$

because $\mathbf{1}^{\mathrm{T}}\mathbf{B}\boldsymbol{\Delta} = \mathbf{0}$. This over-parametrised model implements the translation invariance $t(\mathbf{w}, \Delta) := \mathbf{w} - \mathbf{B}\Delta$.

Prior: We assume a Gaussian prior $p\left(\mathbf{w}\right) = \mathcal{N}\left(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ with diagonal covariance $\boldsymbol{\Sigma} = \mathrm{Diag}\left(\boldsymbol{\sigma}\right)$, where $\boldsymbol{\sigma} = K \cdot \sigma_0 \cdot \mathbf{1}$ as proportional to the number of dimension K s.t. the predictive variance $\mathbb{V}\left[\frac{1}{K}\sum_k w_k + \epsilon\right]$ is constant for different K. Another way to view this is to model a prior over parameters that induces the same prior over functions for different K.

Posterior: The posterior of this Gaussian linear model can be computed as (see App. D)

$$p(\mathbf{w} \,|\, \mathbf{y}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}^*, \mathbf{V}^*\right), \quad \mathbf{V}^* = \left(\frac{1}{K^2\sigma_y}\mathbf{1}\mathbf{1}^{\mathrm{T}} + \boldsymbol{\Sigma}^{-1}\right)^{-1}, \quad \mathbf{m}^* = \mathbf{V}^*\frac{\sum_i y_i}{K\sigma_y}\mathbf{1}. \tag{14}$$

As can be seen, the resulting posterior has full covariance with a diagonal plus rank-1. We will now show that we can construct this posterior from a mean-field Gaussian approximation of the likelihood by marginalising over all translations $\Delta \sim p(\Delta)$ to which the likelihood is invariant.

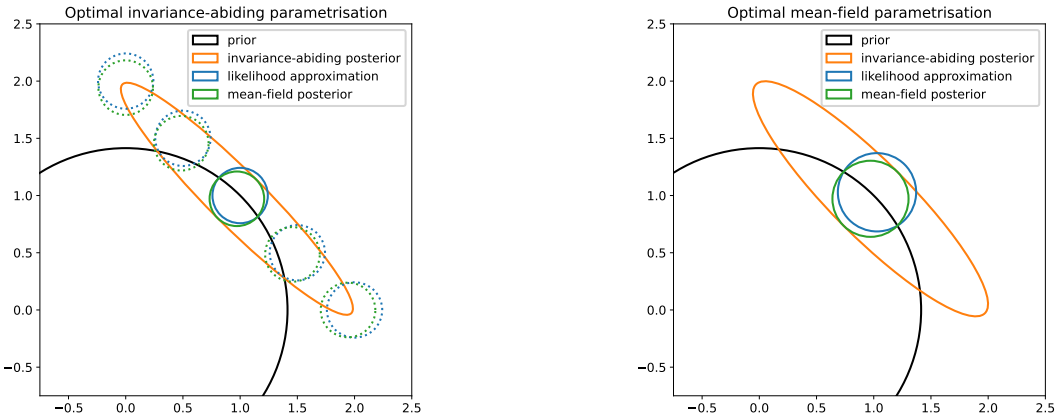
4.1 Mean-field parametrisation of the likelihood function

We now construct the mean-field and invariance-abiding posterior approximations defined in (6) with $\theta = \{\mathbf{m}, \boldsymbol{\lambda}\}$ and denote likelihood and prior covariances as $\mathbf{V} = \mathrm{diag}\left(\boldsymbol{\lambda}\right)$ and $\boldsymbol{\Sigma} = \mathrm{Diag}\left(\boldsymbol{\sigma}\right)$, and we denote the parameters of the standard mean-field posterior from (6a) with

$$q_0\left(\mathbf{w};\, \boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{w};\, \mathbf{m}_{q_0}, \mathbf{V}_{q_0}\right). \tag{15}$$

Next, we model that the likelihood is translation invariant by computing the marginal from (5):

$$\begin{aligned} g_{\mathrm{mix}}\left(\mathbf{w}\right) &= \int g\left(t(\mathbf{w}, \Delta)\right) \cdot p\left(\boldsymbol{\Delta}\right)\, d\boldsymbol{\Delta} \\ &= \lim_{\beta \to \infty} \int \mathcal{N}\left(\mathbf{w}; \mathbf{m} + \mathbf{B}\boldsymbol{\Delta}, \mathbf{V}\right) \cdot \mathcal{N}\left(\boldsymbol{\Delta}; \mathbf{0}, \beta^2\mathbf{I}\right)\, d\boldsymbol{\Delta} \;=:\; \mathcal{N}\left(\mathbf{w}; \mathbf{m}_{\mathrm{mix}}, \mathbf{V}_{\mathrm{mix}}\right). \end{aligned}$$



(a) Optimal parameters $\boldsymbol{\theta}^*_{\text{mix}}$, cf. (19a)

(b) Optimal parameters $\boldsymbol{\theta}^*_0$, cf. (19b)

Figure 1: Visualisation of the Gaussian likelihood approximation and posteriors defined in (15) and (17) with different parameter optima (cf. (19)). Dotted circles show the mean-field likelihood/posterior with alternative parameter values, each inducing the same predictive distribution.

As shown in App. D in the supplementary material, the resulting function is a multivariate Gaussian with a degenerate rank-1 covariance matrix and the location parameter

$$\mathbf{m}_{\text{mix}} = \mathbf{m}, \quad \mathbf{V}_{\text{mix}}^{-1} = \mathbf{V}^{-1}\left(\mathbf{I} - \mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}\right) = \frac{1}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda}}\mathbf{1}\mathbf{1}^{\mathrm{T}}. \tag{16}$$

However, the invariance-abiding posterior is non-degenerate and can be computed as

$$\begin{aligned} q_{\text{mix}}(\mathbf{w};\, \boldsymbol{\theta}) &= \mathcal{N}\left(\mathbf{w};\, \left(\mathbf{V}_{\text{mix}}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\mathbf{V}_{\text{mix}}^{-1}\mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \left(\mathbf{V}_{\text{mix}}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right) \\ &= \mathcal{N}\left(\mathbf{w};\, \boldsymbol{\mu} + \frac{\mathbf{1}^{\mathrm{T}}(\mathbf{m} - \boldsymbol{\mu})}{\mathbf{1}^{\mathrm{T}}(\boldsymbol{\lambda} + \boldsymbol{\sigma})}\boldsymbol{\sigma},\, \boldsymbol{\Sigma} - \frac{1}{\mathbf{1}^{\mathrm{T}}(\boldsymbol{\lambda} + \boldsymbol{\sigma})} \cdot \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}\right) =: \mathcal{N}\left(\mathbf{w};\, \mathbf{m}_{q_{\text{mix}}}, \mathbf{V}_{q_{\text{mix}}}\right). \end{aligned} \tag{17}$$

The standard Gaussian product form in the second line is useful as it shows that this approximation has the covariance matrix structure as the true posterior (cf. (14)). See Fig. 1 for a visualisation of the two respective posterior approximations and the corresponding likelihood with two different parametrisations (cf. Sec. 4.2).

Next, we compute the invariance-gap for our linear model. To simplify the form, we can assume that all variances and means take the same value for the approximation and the prior,

$$\forall k : \lambda_k = \hat{\lambda},\, m_k = \hat{m},\, \sigma_k = \hat{\sigma},\, \mu_k = \hat{\mu},$$

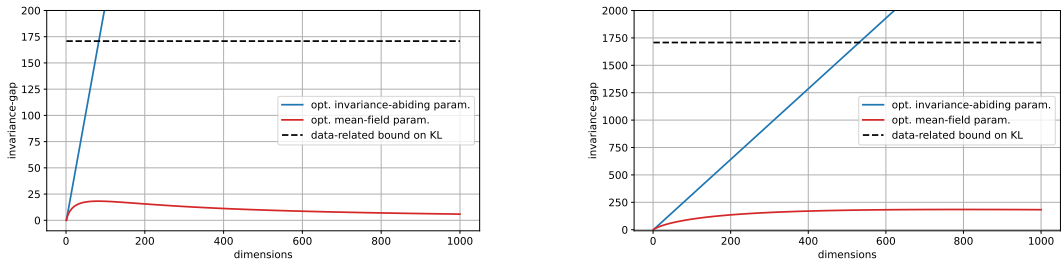
since the posterior approximation is a function of the sum $\mathbf{1}^T\boldsymbol{\lambda}$ only. The invariance gap is then

$$\mathrm{KL}\left[q_0 \,||\, q_{\text{mix}}\right] = \frac{K-1}{2}\left[\ln\left(\frac{\hat{\sigma} + \hat{\lambda}}{\hat{\lambda}}\right) + \frac{\hat{\lambda}}{\hat{\sigma} + \hat{\lambda}} - 1\right]. \tag{18}$$

This is a convex function in $\phi = \frac{\hat{\sigma}+\hat{\lambda}}{\hat{\lambda}}$ with a minimum at $\phi = 1$; it is minimised as $\hat{\lambda} \to \infty$. It is however not evident from the above whether the gap has a large magnitude compared to the rest of the regularisation term and over-regularises in practice. In the next section, we therefore analyse this term at the optima for the mean-field and the invariance-abiding approximation defined in (19), respectively. The corresponding invariance gaps are visualised in Fig. 2.

4.2 Comparison of optimal mean-field and invariance-abiding parametrisation

To better understand the detrimental effect of the invariance gap, we now analyse the ELL and KL terms of the ELBO objective for the mean-field and invariance-abiding variational approximations, respectively. We compare these terms for both distributions with the parameters optimised for both posterior distributions, giving 2×2 combinations. We denote the optimal parameters wrt. the



(a) $N = 10$ observations. (b) $N = 100$ observations.

Figure 2: Invariance-gap evaluated at different optima (cf. (19)), i.e. $\mathrm{KL}\left[q_0(\cdot;\, \boldsymbol{\theta}^*_{\mathrm{mix}}) \,||\, q_{\mathrm{mix}}(\cdot;\, \boldsymbol{\theta}^*_{\mathrm{mix}})\right]$ and $\mathrm{KL}\left[q_0(\cdot;\, \boldsymbol{\theta}^*_0) \,||\, q_{\mathrm{mix}}(\cdot;\, \boldsymbol{\theta}^*_0)\right]$. Prior variances are $\sigma = K \cdot \mathbf{1}$, where K is the number of dimensions; the noise variance is $\sigma_y = (2\pi \mathrm{e})^{-1}$; all observations $y = 1$ and $\mathbf{x} = \mathbf{1}$ are identical. As the K increases, the gap for the mean-field approximation vanishes (variances collapse to prior variances). In contrast, the gap for $\boldsymbol{\theta}^*_{\mathrm{mix}}$, which induces the true posterior predictive, grows linearly and is quickly above the data-related bound (cf. Sec. 2.2).

invariance-abiding approximation and the mean-field approximation, respectively, as

$$\boldsymbol{\theta}^*_{\mathrm{mix}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathcal{L}_{\mathrm{ELBO}}\left(q_{\mathrm{mix}}\left(\cdot\,;\, \boldsymbol{\theta}\right), \mathcal{D}\right), \tag{19a}$$

$$\boldsymbol{\theta}^*_0 = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathcal{L}_{\mathrm{ELBO}}\left(q_0\left(\cdot\,;\, \boldsymbol{\theta}\right), \mathcal{D}\right). \tag{19b}$$

Since q_{mix} is of the same form as the true posterior, the corresponding optimal parameters are

$$g_0(\cdot\,;\, \boldsymbol{\theta}^*_{\mathrm{mix}}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}^*_{\mathrm{mix}}, \mathrm{Diag}\left(\boldsymbol{\lambda}^*_{\mathrm{mix}}\right)\right), \quad \mathbf{m}^*_{\mathrm{mix}} = \frac{1}{N}\sum_{n=1}^{N} y^{(n)} \cdot \mathbf{1}, \quad \boldsymbol{\lambda}^*_{\mathrm{mix}} = \frac{K\sigma_y}{N} \cdot \mathbf{1}. \tag{20}$$

The optimal parameters for the mean-field posterior approximation can also be computed analytically (cf. App. D in the supplementary material) since the true posterior is a multivariate Gaussian:

$$g_0(\cdot\,;\, \boldsymbol{\theta}^*_0) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}^*_0, \mathrm{Diag}\left(\boldsymbol{\lambda}^*_0\right)\right), \quad \mathbf{m}^*_{q_0} = \mathbf{m}^*_{q_{\mathrm{mix}}}, \quad \boldsymbol{\lambda}^*_0 = \frac{K^2\sigma_y}{N}\,\mathbf{1}. \tag{21}$$

We plot the ELL and KL terms and the ELBO for the 2×2 combinations above in Fig. 3. The optimal parametrisation differs by the dependence of the likelihood variance $\boldsymbol{\lambda}$ on the dimension K. Note again that we model the prior variance $\boldsymbol{\sigma} = K\sigma_0\mathbf{1}$ proportional to the number of dimensions s.t. the prior over functions $\mathbb{E}_{p(\mathbf{w})}\left[\ln p(\mathcal{D} \,|\, \mathbf{w})\right]$ is identical for any dimension K. Consequently, also the true posterior over functions is identical for all K. See e.g. 3b, where $\mathbb{E}_{\mathbf{w}\sim q_{\mathrm{mix}}(\cdot\,;\,\boldsymbol{\theta}^*_{\mathrm{mix}})}\left[\ln p(\mathcal{D} \,|\, \mathbf{w})\right]$ is constant in K and $q_{\mathrm{mix}}(\mathbf{w}\,;\, \boldsymbol{\theta}^*_{\mathrm{mix}})$ is identical to the true posterior. In contrast, $\boldsymbol{\lambda}^*_0$ depends quadratically on K; therefore the variance of the posterior approximation $q_0(\cdot\,;\, \boldsymbol{\theta}^*_0)$ *collapses* to the prior variance as $K \to \infty$. Consequently, the predictive variance reverts to the prior predictive variance (cf. Fig. 3d). We have thus shown that—in this simplified setting of a linear model with a single data point for $K \to \infty$—the consequence of not handling translation invariance is indeed posterior collapse, while the handling of it coincides with the true posterior.

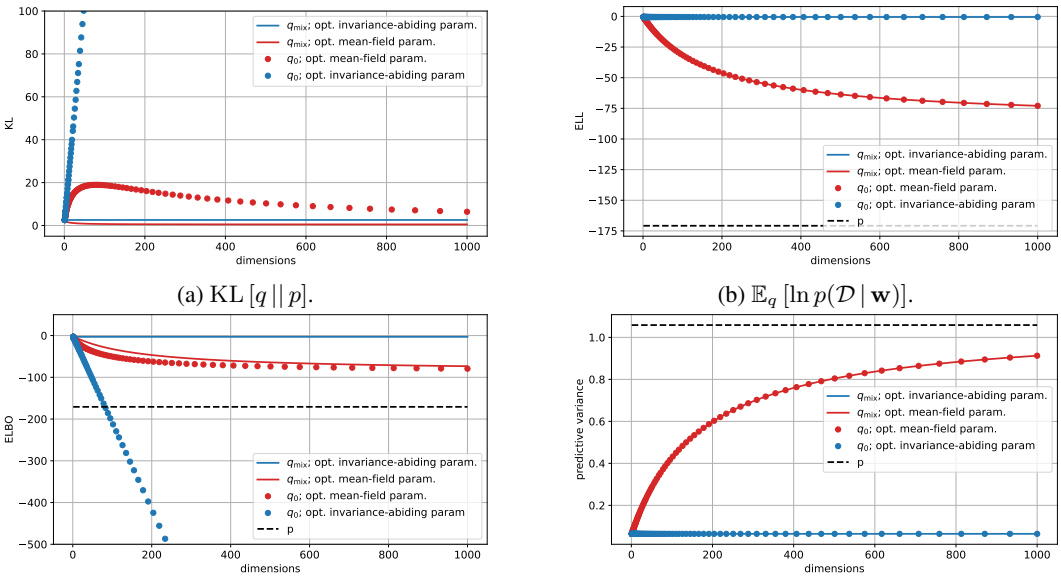
4.3 Translation invariance in Bayesian neural networks

Let us now go back to the general neural network model in Sec. 2.1. In order to describe the set of all invariances, note that for each node $z_{i,l}$ in layer l (or the output node $f(\mathbf{x})$) with a D–dimensional input (either $\mathbf{x}$ at the input layer or $\mathbf{z}_{l-1}$ at all other layers), there is a $D-1$ dimensional subspace that keeps the value $z_{i,l}$ (or the output node y) invariant when using the translation-invariance model in (13). More formally, it is easy to see that the following set of equations models all translation invariances of a general neural network at a data point $\mathbf{x}$:

$$\forall \boldsymbol{\Delta}_{L,1} \in \mathbb{R}^{n_{L-1}-1}: \quad f(\mathbf{x}) = h_L\left(\mathbf{w}^{\mathrm{T}}_{L,1}\mathbf{z}_{L-1}\right) = h_L\left(\left(\mathbf{w}_{L,1} + \mathbf{B}_{\mathbf{z}_{L-1}}\boldsymbol{\Delta}_{L,1}\right)^{\mathrm{T}}\mathbf{z}_{L-1}\right), \tag{22}$$

$$\forall \boldsymbol{\Delta}_{l,i} \in \mathbb{R}^{n_{l-1}-1}: \quad z_{l,i} = h_l\left(\mathbf{w}^{\mathrm{T}}_{l,i}\mathbf{z}_{l-1}\right) h_l\left(\left(\mathbf{w}_{l,i} + \mathbf{B}_{\mathbf{z}_{l-1}}\boldsymbol{\Delta}_{l,i}\right)^{\mathrm{T}}\mathbf{z}_{l-1}\right), \tag{23}$$

$$\forall \boldsymbol{\Delta}_{1,j} \in \mathbb{R}^{\|\mathbf{x}\|_0-1}: \quad z_{1,j} = h_1\left(\mathbf{w}^{\mathrm{T}}_{1,j}\mathbf{x}\right) = h_1\left(\left(\mathbf{w}_{1,j} + \mathbf{B}_{\mathbf{x}}\boldsymbol{\Delta}_{1,j}\right)^{\mathrm{T}}\mathbf{x}\right), \tag{24}$$



(a) $\mathrm{KL}\left[q \,||\, p\right]$.

(b) $\mathbb{E}_q\left[\ln p(\mathcal{D} \,|\, \mathbf{w})\right]$.

(c) $\mathcal{L}_{\mathrm{ELBO}}(q,\, \mathcal{D}) = \mathbb{E}_q\left[\ln p(\mathcal{D} \,|\, \mathbf{w})\right] - \mathrm{KL}\left[q \,||\, p\right]$.

(d) $\mathbb{V}_{\mathbf{w}\sim q}\left[\frac{1}{K}\mathbf{1}^{\mathrm{T}}\mathbf{w}\right] + \sigma_y$.

Figure 3: ELBO loss terms and predictive variance for different posterior approximations. Lines/dots indicate the invariance-abiding/mean-field posterior, respectively. Red/blue indicates the optimal invariance-abiding/mean-field parameters (cf. (19)). Prior variance is $\boldsymbol{\sigma} = K \cdot \mathbf{1}$, and $\sigma_y = (2\pi \mathrm{e})^{-1}$; we used $N = 10$ identical observations with value $y = 1$ and inputs $\mathbf{x} = \mathbf{1}$. Notably, both posterior approximations yield the same predictive distribution as can be seen by the and ELL (b) and predictive variance (d), and the variance of the mean-field approximation reverts to the prior variance.

where $l \in \{2, \ldots, L-1\}$, $i \in \{1, \ldots, n_l\}$, $j \in \{1, \ldots, n_1\}$ and $\mathbf{B}_{\mathbf{z}}$ given by

$$\mathbf{B}_{\mathbf{z}} := \begin{bmatrix} & \mathbf{I} & \\ -\frac{z_1}{z_k} & \cdots & -\frac{z_{k-1}}{z_k} \end{bmatrix} .$$

In order to efficiently compute the invariant-abiding likelihood g_{mix}, we note that it decomposes over the layers of the neural network as follows: For all the n_1 nodes in the first layer we have

$$\begin{aligned} q_{\mathrm{mix}}(\mathbf{w}_{1,j}) &= \mathcal{N}\left(\mathbf{w}_{1,j}; \boldsymbol{\mu} + \frac{\mathbf{x}^{\mathrm{T}}\left(\mathbf{m} - \boldsymbol{\mu}\right)}{\mathbf{x}^{\mathrm{T}}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)\mathbf{x}} \cdot \left(\boldsymbol{\Sigma}\mathbf{x}\right), \boldsymbol{\Sigma} - \frac{1}{\mathbf{x}^{\mathrm{T}}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)\mathbf{x}} \cdot \left(\boldsymbol{\Sigma}\mathbf{x}\right)\left(\boldsymbol{\Sigma}\mathbf{x}\right)^{\mathrm{T}}\right) , \\ P(z_{1,j}) &= \mathbb{E}_{\mathbf{w}\sim q_{\mathrm{mix}}(\mathbf{w}_{1,j})}\left[h_1\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)\right] , \end{aligned}$$

where $j \in \{1, \ldots, n_1\}$ and $\mathbf{m}$, $\mathbf{V}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are constrained to the parts of the overall weight vector that correspond to $\mathbf{w}_{1,j}$. Now, since the invariance-abiding mechanism for nodes in layer l only depends on the *value* of the hidden units in layer $n-1$ (see (22) amd (23)), we have

$$\begin{aligned} q_{\mathrm{mix}}(\mathbf{w}_{l,i}) &= \mathbb{E}_{\mathbf{z}\sim P(\mathbf{z}_{l-1})}\left[\mathcal{N}\left(\mathbf{w}_{l,i}; \boldsymbol{\mu} + \frac{\mathbf{z}^{\mathrm{T}}\left(\mathbf{m} - \boldsymbol{\mu}\right)}{\mathbf{z}^{\mathrm{T}}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)\mathbf{z}} \cdot \left(\boldsymbol{\Sigma}\mathbf{z}\right), \boldsymbol{\Sigma} - \frac{1}{\mathbf{z}^{\mathrm{T}}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)\mathbf{z}} \cdot \left(\boldsymbol{\Sigma}\mathbf{z}\right)\left(\boldsymbol{\Sigma}\mathbf{z}\right)^{\mathrm{T}}\right)\right] , \\ P(z_{l,i}) &= \mathbb{E}_{\mathbf{z}\sim P(\mathbf{z}_{l-1})}\left[\mathbb{E}_{\mathbf{w}\sim q_{\mathrm{mix}}(\mathbf{w}_{l,i})}\left[h_l\left(\mathbf{w}^{\mathrm{T}}\mathbf{z}_{l-1}\right)\right]\right] , \end{aligned} \tag{25}$$

where again $\mathbf{m}$, $\mathbf{V}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are constrained to the parts of the overall weight vector that correspond to $\mathbf{w}_{l,i}$. Even though the non-linear transformations lead to a non-Gaussian distribution of $P(z_{l,i})$, we note that the mean and variance of each component $z_{1,j}$ in the first layer is identical and equal to $\mathbf{x}^{\mathrm{T}}\left(\boldsymbol{\mu}\frac{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}(\mathbf{V}+\boldsymbol{\Sigma})\mathbf{x}} + \mathbf{m}\frac{\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}(\mathbf{V}+\boldsymbol{\Sigma})\mathbf{x}}\right)$ and $\frac{(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x})(\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x})}{(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x})+(\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x})}$, respectively. A similar argument holds for the distributions of the unit values $z_{l,i}$ in all other layers (by swapping the expectation in (25)) which implies that the the translation invariance in deeper layers leads to factorizing posteriors q_{mix} for all weights in deeper layers. Bounding the KL divergence between q_0 and q_{mix} in a computationally efficient way is still an open problem and not addressed here.

5 Related work

Poor empirical performance of BNNs in general and mean-field VBNNs in particular has been reported in several previous works, e.g., [32, 14, 29]. However, in Farquhar et al. [7], the authors argue that mean-field approximations are expressive enough for deep BNNs; they prove a universality result that the predictive distribution of mean-field approximations of BNNs with at least 2 layers of hidden units *can* approximate any true posterior distribution over function values arbitrarily closely. This result seems in apparent conflict with our line of reasoning and [4] who shows that the predictive distribution of mean-field VBNNs reverts to the prior predictive distribution as the network width increases. The discrepancy between these two results is resolved by noting that Farquhar et al. [7] only shows that the expressive power of the model is large enough but not that the approximate inference algorithms will converge to this solution. Our work sheds light on why the mean-field approximation fails to approximate expressive posteriors via the invariance-gap in the ELBO.

Other attempts have proposed to approximate the posterior predictive directly, thereby circumventing the non-identifiability issue [24, 18, 31]; while these approaches have shown promising empirical performance, estimating the KL regularisation term in function space is significantly more complicated and can even be ill-defined due to regions of zero prior probability mass [3]. In a similar vein, previous work also proposed to map the BNN prior to a Gaussian process (functional) prior [8, 28].

Interestingly, it has been shown that restricting the parametrisation of the variational posterior is competitive or even leads to better performance [26, 19, 27, 6]. For example, Dusenberry et al. [6] proposes a variational approximation only for rank-1 factors, resulting in inference of a lower-dimensional subspace; the outer product of these low-rank factors are then used to perturb a weight matrix that is treated deterministically through maximum a posteriori (MAP) estimation. This approach avoids the invariance problem associated with variational Bayesian inference since the MAP estimate is not impeded by this problem and the subspace of the variational weights do not possess the same invariances.

More broadly, the (non-)identifiability of parameters and latent variables in probabilistic models is a widely-studied topic both from frequentist [23] and Bayesian perspectives [5, 21, 9]. In a recent example, Wang et al. [30] attibute posterior collapse in the context of variational auto-encoders to non-identifiability of latent variables.

6 Conclusion

We have associated the posterior collapse phenomenon of mean-field VBNNs with invariances in the likelihood function, in particular translation invariance. While the invariance does not affect the predictive distribution, the approximations of the posterior which abide and do not abide the invariance differ in the KL regularisation term and consequently in the tightness of the ELBO objective. We proved that the objectives of the two approximations differ by the relative entropy (KL) between the standard mean-field approximation and the invariance-abiding distribution which we related to a data-related bound on the KL regularisation, preventing fits that for which the gap is large.

We studied over-parametrised Bayesian linear regression as the canonical model with *translation invariance* for which the relevant terms can be computed exactly. A detailed analysis of this model confirms our hypothesis that the invariance leads to a significant additional gap in the ELBO objective compared to approximations that model the invariance. It is this very gap which prevents mean-field approximation to achieve the same fit as an invariance-abiding approximation and instead leads to a collapse of the posterior variances to the prior variance.

While we can compute the invariance gap for the over-parametrised linear model, we have not yet identified a computationally efficient procedure for layered model and other invariances (e.g., permutation invariance). However, our work provides the mathematical tools to address over-regularisation due to invariances. In our future work we will focus on approximations of the invariance gap in order to correct the ELBO objective for translation and permutation invariances in general neural network functions.

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*, ICML'15, page 1613–1622. JMLR.org, 2015.

[2] J. Brea, B. Simsek, B. Illing, and W. Gerstner. Weight-space symmetry in deep networks gives rise to permutation saddles, connected by equal-loss valleys across the loss landscape. *ArXiv*, abs/1907.02911, 2019.

[3] D. R. Burt, S. W. Ober, A. Garriga-Alonso, and M. van der Wilk. Understanding variational inference in function-space. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021.

[4] B. Coker, W. P. Bruinsma, D. R. Burt, W. Pan, and F. Doshi-Velez. Wide mean-field variational bayesian neural networks ignore the data. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[5] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. ISSN 00359246.

[6] M. W. Dusenberry, G. Jerfel, Y. Wen, Y.-A. Ma, J. Snoek, K. Heller, B. Lakshminarayanan, and D. Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[7] S. Farquhar, L. Smith, and Y. Gal. Liberty or depth: Deep bayesian neural nets do not need complex weight posterior approximations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4346–4357. Curran Associates, Inc., 2020.

[8] D. Flam-Shepherd. Mapping gaussian process priors to bayesian neural networks. In *Bayesian Deep Learning NeurIPS workshop*, 2017.

[9] A. E. Gelfand and S. K. Sahu. Identifiability, improper priors, and gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253, 1999.

[10] A. Gelman, D. Simpson, and M. Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10), 2017. ISSN 1099-4300. doi: 10.3390/e19100555.

[11] 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 of *Proceedings of Machine Learning Research*, pages 1744–1753. PMLR, 10–15 Jul 2018.

[12] R. Herbrich. *Learning Kernel Classifiers: Theory and Algorithms*. The MIT Press, 2nd edition edition, 2002.

[13] G. E. Hinton and D. van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, page 5–13. Association for Computing Machinery, 1993. ISBN 0897916115. doi: 10.1145/168304.168306.

[14] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. G. Wilson. What are bayesian neural network posteriors really like? In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4629–4640. PMLR, 18–24 Jul 2021.

[15] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[16] R. Kurle, B. Cseke, A. Klushyn, P. van der Smagt, and S. Günnemann. Continual learning with bayesian neural networks for non-stationary data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[17] R. Kurle, T. Januschowski, J. Gasthaus, and B. Wang. On symmetries in variational bayesian neural nets. In *Bayesian Deep Learning NeurIPS workshop*, 2021.

[18] C. Ma, Y. Li, and J. M. Hernandez-Lobato. Variational implicit processes. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4222–4233. PMLR, 09–15 Jun 2019.

[19] A. Mishkin, F. Kunstner, D. Nielsen, M. W. Schmidt, and M. E. Khan. Slang: Fast structured covariance approximations for bayesian deep learning with natural gradient. In *NeurIPS*, pages 6248–6258, 2018.

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

[21] D. J. Poirier. Revising beliefs in nonidentified models. *Econometric Theory*, 14(4):483–509, 1998.

[22] A. Pourzanjani, R. M. Jiang, and L. Petzold. Improving the identifiability of neural networks for bayesian inference. In *Bayesian Deep Learning NeurIPS workshop*, 2017.

[23] B. L. S. Prakasa Rao. *Identifiability in Stochastic Models*. Academic Press, Oxford, 1992.

[24] S. Sun, G. Zhang, J. Shi, and R. B. Grosse. Functional variational bayesian neural networks. *ArXiv*, abs/1903.05779, 2019.

[25] H. J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, 1992. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(05)80037-1.

[26] J. Swiatkowski, K. Roth, B. S. Veeling, L. Tran, J. V. Dillon, J. Snoek, S. Mandt, T. Salimans, R. Jenatton, and S. Nowozin. The k-tied normal distribution: A compact parameterization of gaussian mean field posteriors in bayesian neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[27] M. Tomczak, S. Swaroop, and R. Turner. Efficient low rank gaussian variational inference for neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4610–4622. Curran Associates, Inc., 2020.

[28] B.-H. Tran, S. Rossi, D. Milios, and M. Filippone. All you need is a good functional prior for bayesian deep learning. *Journal of Machine Learning Research*, 23(74):1–56, 2022.

[29] B. Trippe and R. Turner. Overpruning in variational bayesian neural networks. *arXiv preprint arXiv:1801.06230*, 2018.

[30] Y. Wang, D. Blei, and J. P. Cunningham. Posterior collapse and latent variable non-identifiability. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5443–5455. Curran Associates, Inc., 2021.

[31] Z. Wang, T. Ren, J. Zhu, and B. Zhang. Function space particle optimization for bayesian neural networks. In *International Conference on Learning Representations*, 2019.

[32] F. Wenzel, K. Roth, B. Veeling, J. Swiatkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. How good is the Bayes posterior in deep neural networks really? In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10248–10259. PMLR, 13–18 Jul 2020.

[33] A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *CoRR*, abs/2002.08791, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We point out limitations and future work throughout the paper.

(c) Did you discuss any potential negative societal impacts of your work? [No] This is purely theoretical work without direct negative societal impacts.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes]

   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...[No] No experiments

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets... [No] No existing assets used.

5. If you used crowdsourcing or conducted research with human subjects... [No] No crowdsourcing/human subjects.