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

Richard Kurle [*]
AWS AI Labs
kurler@amazon.com

Ralf Herbrich
Hasso-Plattner Institut
ralf.herbrich@hpi.de

Tim Januschowski [†]
Zalando SE
tim.januschowski@zalando.de

Yuyang Wang
AWS AI Labs
yuyawang@amazon.com

Jan Gasthaus
AWS AI Labs
gasthaus@amazon.com

Abstract

Variational Bayesian posterior inference often requires simplifying approximations such as mean-field parametrisation to ensure tractability. However, prior work has associated the variational mean-field approximation for Bayesian neural networks with underfitting 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 contribute to this phenomenon because these invariances complicate the structure of the posterior by introducing discrete and/or continuous modes which cannot be well approximated by Gaussian mean-field distributions. In particular, we show that the mean-field approximation has an additional gap in the evidence lower bound compared to a purpose-built posterior that takes into account the known invariances. Importantly, this invariance gap is not constant; it vanishes as the approximation reverts to the prior. 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 invariance problem.

1 Introduction

Bayesian neural networks (BNNs) have several appealing advantages compared to deterministic neural networks (NN) such as improving generalization [38], capturing epistemic uncertainty [17], and providing a framework for continual learning methods [18, 23]. Unfortunately, reaping these theoretical benefits has so far been impeded [37]. 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 [15, 2]—have been identified to be partially responsible for a performance gap compared to deterministic NNs[16]. The most common variational approximation of the posterior is a product of independent Normal distributions, commonly referred to as the *mean-field* approximation. It has been observed that mean-field variational BNNs (VBNNs) suffer from severe underfitting for large models and small dataset sizes [13]. Recent work [5] 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*.

[*]Correspondence to kurler@amazon.com.
[†]Work done while at AWS AI Labs.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

The goal of this work is to shed light on *why* the mean-field variational approximation collapses and to provide a new angle for future works to address this shortcoming. To this end, we study invariances in the likelihood function of *overparametrised models* and their detrimental effect on variational inference. For instance, it is well known that NNs exhibit several invariances *with respect to their parameters*, including node permutation invariance [19], sign-flip invariance (in the case of odd activation functions) [29, 3], scaling invariance (in the case of piece-wise linear activations) [26], and as we will show in Sec. 5, 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, these invariances are detrimental to the variational Bayesian approach. As a first step to show this, we isolate the impact of the invariances in Sec. 3. To this end, we construct both a *mean-field* as well as an *invariance-abiding* approximation which explicitly models the invariances by integrating over all transformations that leave the likelihood invariant. Notably, both aforementioned posterior approximations are constructed from the same (mean-field) *likelihood* approximation. We then prove that, under the conditions outlined in Sec. 3, the mean-field approximation induces the same posterior predictive distribution as the invariance-abiding approximation. However, we also prove that the ELBO objective corresponding to these two approximations differs by the KL divergence between both posterior approximations; we refer to this difference as *invariance gap*. Importantly, the gap vanishes if both distributions are identical, which is the case if the mean-field approximation reverts to the prior.

We then demonstrate the detrimental effect of invariances in the likelihood function of an overparametrised Bayesian linear regression model in Sec. 4. This model is purposely selected as the canonical model exhibiting (only) *translation invariance*. We provide a detailed analysis of this model, including a tractable solution for the invariance-abiding approximation. It turns out that the optimal parameters w.r.t. the ELBO objective with the invariance-abiding distribution result in the true posterior. In contrast, posterior approximations with parameters that are optimal w.r.t. the mean-field ELBO revert to the prior as the number of dimensions increases.

Finally, we transfer our analysis of the linear model to VBNNs in Sec. 5, showing that subspaces in BNNs exist that exhibit translation invariance. Combined with a previous analysis of the node permutation invariance (cf. Kurle et al. [19] or App. E of the supplementary material), this provides the basis for future work to approximate the (generally intractable) invariance gap and thereby optimise for a tighter and favorable ELBO objective. We start in Sec. 2 by introducing the basic concepts of VBNNs and recent results on which we build our contribution.

2 Background

2.1 Variational Bayesian Neural Networks

Neural network functional model. Deep NNs are layered models that progressively transform their inputs in each layer. More formally, an L–layer NN computes algebraically

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

where the $h_l : \mathbb{R} \to \mathbb{R}$ are monotonic transfer functions that introduce non-linearities. Such a network has $n_1 + n_2 + \cdots + n_L$ many nodes $z_{l,i}$, where i indexes the layer-wise vectors $\mathbf{z}_l$. These nodes are each a weighted linear combination of either the input vector $\mathbf{x}$ (first hidden layer) or the value of the hidden units $\mathbf{z}_l$ (all other hidden layers and the output $f(\mathbf{x})$). We denote all learnable parameters of the model by the stacked weight vectors of each layer and node, $\mathbf{w} = [\mathbf{w}_{1,1}^{\mathrm{T}}, \ldots, \mathbf{w}_{L,n_L}^{\mathrm{T}}]^{\mathrm{T}}$.

Variational Bayesian treatment. If the weights and biases $\mathbf{w}$ are treated as random variables with a prior distribution $p(\mathbf{w})$, then the posterior $p(\mathbf{w} \,|\, \mathcal{D})$—induced by the dataset $\mathcal{D}$ through the likelihood function $\ell(\mathbf{w}; \mathcal{D}) := p(\mathcal{D} \,|\, \mathbf{w})$ that is defined via $f(\mathbf{x})$—is referred to as a Bayesian neural network (BNN). 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 optimization problem

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

where $\mathrm{KL}\left[q_{\boldsymbol{\theta}} \,||\, p(\cdot \,|\, \mathcal{D})\right] := \mathbb{E}_{\mathbf{w}\sim q_{\boldsymbol{\theta}}}\left[\ln q_{\boldsymbol{\theta}}(\mathbf{w}) - \ln p(\mathbf{w}|\mathcal{D})\right]$ and $\mathcal{Q}$ is a family of distributions over $\mathbf{w}$. The optimization of (1) is achieved by maximising a lower bound to the (log) model evidence $\ln p(\mathcal{D})$,

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

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

Definition 1 (Mean-field variational BNN)**.** A *mean-field* variational BNN (VBNN) is the minimizer of (1) where $\mathcal{Q}$ is the family of Gaussians with diagonal covariance matrix.

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 [5] and we recall the result for a Gaussian likelihood with homogeneous noise (for other likelihood functions, we refer to App. F in [5]).

Assume a VBNN with Gaussian observation noise defined through $y = f_{\mathbf{w}}(\mathbf{x}) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$, 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 the NN output variance $\sigma_L^2\left(\mathbf{x}^{(n)}\right) := \mathbb{V}_{\mathbf{w}\sim p}\left[f(\mathbf{x}^{(n)};\, \mathbf{w})\right]$ for some fixed input $\mathbf{x}^{(n)}$, where L indicates the last layer. Then,

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

where q^* is a hypothetical approximation that predicts the data perfectly up to the noise variance σ_y^2 (see App. F for details). Prior work [5] 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 the following section.

3 Invariance-abiding variational approximation

We develop a framework that enables modelling the invariances in the likelihood and understanding their impact on the ELBO objective. 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 q_{mix} which can be related to a *mean-field* approximation q_0 that does not model invariances. We then describe conditions under which both approximations yield an identical posterior predictive, while the KL regularisation term of q_0 is lower bounded by the KL of q_{mix}.

Variational likelihood and posterior approximations. Non-identifiability implies that the posterior does not concentrate on a single set of parameters irrespective of the dataset size, because the same likelihood is assigned to different parameter values [12]. We model this invariance through transformations $t(\cdot, \mathbf{r})$ to which the likelihood $\ell(\mathbf{w};\mathcal{D}) = p(\mathcal{D} \,|\, \mathbf{w})$ is invariant via variables $\mathbf{r}$:

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

From (4) it follows that the likelihood is also invariant w.r.t. the marginalisation

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

We assume that each of the equivalent parametrisations $\mathbf{w}' = t(\mathbf{w}, \mathbf{r})$ of the likelihood has the same probability a priori. For discontinuous transformations such as node permutations (see App. E), $p(\mathbf{r})$ is a uniform distribution over discrete variables indexing these transformations. For the continuous translation invariance (see Sec. 4), we model the uniform distribution as a Gaussian with infinite variance. It is conceivable that the likelihood can be constructed by marginalising over these transformations of a simpler function $\ell_0(\mathbf{w};\mathcal{D})$ such that $\ell(\mathbf{w};\mathcal{D}) = \mathbb{E}_{\mathbf{r}\sim p}\left[\ell_0(t(\mathbf{w}, \mathbf{r}); \mathcal{D})\right]$. For

instance, permutation invariance induces a factorial number of discontinuous modes that are each equivalent (cf. [19], App. E); and as we show in Sec. 4, the full covariance Gaussian likelihood and posterior of an over-parametrised Bayesian linear regression model can be constructed from a product of independent Gaussians. Curiously, it may be sufficient to approximate ℓ_0 while taking into account the known invariances. To this end, we define a Gaussian *variational likelihood approximation* $g_0(\mathbf{w};\boldsymbol{\theta}) \approx \ell_0(\mathbf{w},\mathcal{D})$, with variational parameters $\boldsymbol{\theta}$. From this single mode approximation, we model an *invariance-abiding likelihood* approximation using the same parametrisation through

$$g_{\text{mix}}(\mathbf{w};\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{r}\sim p}\left[g_0(t(\mathbf{w},\mathbf{r});\boldsymbol{\theta})\right] \approx \mathbb{E}_{\mathbf{r}\sim p}\left[\ell_0(t(\mathbf{w},\mathbf{r}),\mathcal{D})\right] = \ell(\mathbf{w};\mathcal{D}). \tag{5}$$

We consider mean-field Gaussians for the prior $p(\mathbf{w})$ and likelihood approximation $g_0(\mathbf{w};\boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\mathbf{m}, \boldsymbol{\lambda}\}$ are means and variances of g_0. We then define a *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:

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

where Z_0 and Z_{mix} are normalisation constants. While $q_0(\mathbf{w};\boldsymbol{\theta})$ is a mean-field approximation, 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 will describe continuous translation invariance in Sec. 4; for the discrete node permutation invariance, see App. E of the supplementary material.

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. We first write the density q_{mix} as an expectation of product densities,

$$q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta}) = Z_{\text{mix}}^{-1}\, p(\mathbf{w}) \cdot \int p(\mathbf{r})\, g_0(t(\mathbf{w},\mathbf{r});\boldsymbol{\theta}) d\mathbf{r} = \int p(\mathbf{r})\, \frac{p(\mathbf{w}) \cdot g_0(t(\mathbf{w},\mathbf{r});\boldsymbol{\theta})}{Z_{\text{mix}}} d\mathbf{r}.$$

Then, we assume that there exists a mapping $\mathbf{r}' = \varphi(\mathbf{r})$ such that

$$\forall \mathbf{r} \sim p(\mathbf{r}) : p(\mathbf{w}) \cdot g_0(t(\mathbf{w},\mathbf{r});\boldsymbol{\theta}) = p(t(\mathbf{w},\varphi(\mathbf{r}))) \cdot g_0(t(\mathbf{w},\varphi(\mathbf{r}));\boldsymbol{\theta}). \tag{7}$$

The condition in (7) allows us to exploit the invariance property of the likelihood (approximation) also for each product density $q_0(t(\mathbf{w},\varphi(\mathbf{r}));\boldsymbol{\theta}) \propto p(t(\mathbf{w},\varphi(\mathbf{r}))) \cdot g_0(t(\mathbf{w},\varphi(\mathbf{r}));\boldsymbol{\theta})$, because then

$$q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta}) = \int \frac{Z_0(\mathbf{r})}{Z_{\text{mix}}} p(\mathbf{r}) \cdot q_0(t(\mathbf{w},\varphi(\mathbf{r}));\boldsymbol{\theta}) d\mathbf{r},$$

where the normalisation constants are $Z_0(\mathbf{r}) = \int p(\mathbf{w}) \cdot g_0(t(\mathbf{w},\varphi(\mathbf{r}));\boldsymbol{\theta}) d\mathbf{w}$, and Z_{mix} is defined in (6b). We show that the condition in (7) holds for translation and permutation invariance in App. C.

The second condition is that all invariance transformations $t(\cdot,\mathbf{r})$ must be volume-preserving, i.e.

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

Although (7) and (8) are fairly restricting, common invariances such as translation and permutation invariance fulfil these conditions for Gaussian priors and likelihood approximations (see App. C). With (7) and (8), we can show the posterior predictive equivalence (see Lemma 1 in App. C)

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

Invariance gap. If the conditions in (7) and (8) are met, we also show (see Lemma 2 in App. C)

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

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 it as the *invariance gap*. We make the following observation about the detrimental effect of the invariances: when maximising the standard ELBO $\mathcal{L}_{\text{ELBO}}\left(q_0,\mathcal{D}\right)$ instead of the tighter objective $\mathcal{L}_{\text{ELBO}}\left(q_{\text{mix}},\mathcal{D}\right)$ w.r.t. the parameters of the variational *likelihood* approximation $g_0(\mathbf{w};\boldsymbol{\theta})$ (used to construct both q_0 and q_{mix}), we see that the former objective favors solutions where q_0 and q_{mix} coincide. This suboptimal solution is obtained if the mean-field posterior $q_0(\mathbf{w})$ and thus also the invariance-abiding posterior $q_{\text{mix}}(\mathbf{w})$ revert to the prior. This is the case if $g_0(\mathbf{w})$ is uniform, because then $g_{\text{mix}}(\mathbf{w})$ is uniform as well, and because both posteriors are constructed as the product of prior and likelihood approximation (cf. (6a)).

Implication of data-related bound. Since $\mathrm{KL}\left[q_0 \,||\, p\right]$ 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)):

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

Consequently, this bound puts a constraint on the achievable solutions to the maximisation of the ELBO objective w.r.t. the variational parameters of the mean-field approximation $q_0(\mathbf{w}; \boldsymbol{\theta})$. As discussed above, one specific parametrisation for which the invariance gap vanishes is the mean-field posterior collapse, $q_0(\mathbf{w}; \boldsymbol{\theta}) \approx p(\mathbf{w})$. However, to show that the invariance gap not only admits posterior collapse as a potential parametrisation but indeed incentivises it, we next tackle the question how the invariance gap behaves for non-collapsed approximations. One particularly relevant variational parametrisation is the optimal solution w.r.t. the ELBO objective with the invariance-abiding approximation. In order to tackle this question, we now consider a simple model for which the relevant distributions and the invariance gap can be computed exactly.

4 Translation invariance in linear models

We study an over-parametrised Bayesian linear regression model as the canonical model that exhibits *translation invariance.* The model 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). We show how to lift results from this canonical model to the more general case of NNs in Sec. 5.

Likelihood model. Consider a linear model with K latent variables $\mathbf{w} = [w_1, \ldots, w_K]^{\mathrm{T}}$ and assume that we have N observations of variables y given dependent inputs $\mathbf{x}$. To simplify the setting, we will assume that $\mathbf{x} = \mathbf{1}$ (see App. D 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, $y = \frac{1}{K}\mathbf{1}^{\mathrm{T}}\mathbf{w} + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma_y^2\right)$. 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

$$\mathbf{1}^{\mathrm{T}}\mathbf{w} = \mathbf{1}^{\mathrm{T}}\left(\mathbf{w} + \mathbf{B}\boldsymbol{\Delta}\right), \quad \mathbf{B} := \begin{bmatrix} \mathbf{I} \\ -\mathbf{1}^{\mathrm{T}} \end{bmatrix}, \tag{12}$$

since $\mathbf{1}^{\mathrm{T}}\mathbf{B}\boldsymbol{\Delta} = \mathbf{0}$. This over-parametrised model exhibits 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 $\boldsymbol{\Sigma} = \mathrm{Diag}\left(\boldsymbol{\sigma}^2\right)$, where we consider $\boldsymbol{\sigma}^2 = K \cdot \sigma_0^2 \cdot \mathbf{1}$ proportional to the number of dimensions K, such that the predictive variance $\mathbb{V}\left[\frac{1}{K}\sum_k w_k + \epsilon\right]$ is constant w.r.t. 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.3)

$$p(\mathbf{w} \,|\, \mathbf{y}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_p^*, \mathbf{V}_p^*\right), \quad \mathbf{V}_p^* = \left(\frac{N}{K^2\sigma_y^2}\mathbf{1}\mathbf{1}^{\mathrm{T}} + \boldsymbol{\Sigma}^{-1}\right)^{-1}, \quad \mathbf{m}_p^* = \mathbf{V}_p^*\frac{\sum_i y_i}{K\sigma_y^2}\mathbf{1}. \tag{13}$$

As can be seen, the resulting posterior has full covariance with a diagonal plus rank-1 matrix. We will now show that we can construct this posterior from the prior and 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

Next, we construct the mean-field and invariance-abiding posterior approximations defined in (6) from the prior defined above and the mean-field likelihood approximation with locations $\mathbf{m}$ and variances $\boldsymbol{\lambda}$. We model that the likelihood is translation invariant by computing the marginal from (5) and considering $p(\Delta) = \mathcal{N}\left(\boldsymbol{\Delta}; \mathbf{0}, \beta^2\mathbf{I}\right)$ with $\beta \to \infty$ as the uniform distribution over all translations:

$$\begin{aligned} g_{\mathrm{mix}}\left(\mathbf{w}; \boldsymbol{\theta}\right) &= \int g_0\left(t(\mathbf{w}, \Delta); \boldsymbol{\theta}\right) \cdot p\left(\boldsymbol{\Delta}\right) \, d\boldsymbol{\Delta} \\ &= \lim_{\beta \to \infty} \int \mathcal{N}\left(\mathbf{w}; \mathbf{m} + \mathbf{B}\boldsymbol{\Delta}, \mathrm{Diag}\left(\boldsymbol{\lambda}\right)\right) \cdot \mathcal{N}\left(\boldsymbol{\Delta}; \mathbf{0}, \beta^2\mathbf{I}\right) \, d\boldsymbol{\Delta} \;=:\; \mathcal{N}\left(\mathbf{w}; \mathbf{m}_{\mathrm{mix}}, \mathbf{V}_{\mathrm{mix}}\right). \end{aligned}$$

Writing the uniform distribution as the limiting case of a Gaussian with infinite variance allows us to compute the integral analytically. As shown in App. D, the resulting function is a multivariate Gaussian with a degenerate rank-1 covariance matrix and the same location parameter as $g_0(\mathbf{w})$:

$$\mathbf{m}_{\text{mix}} = \mathbf{m}, \quad \mathbf{V}_{\text{mix}}^{-1} = \frac{1}{\mathbf{1}^{\text{T}}\boldsymbol{\lambda}}\mathbf{1}\mathbf{1}^{\text{T}}. \tag{14}$$

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

$$q_{\text{mix}}(\mathbf{w};\, \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{w};\, \boldsymbol{\mu} + \frac{\mathbf{1}^{\text{T}}(\mathbf{m}-\boldsymbol{\mu})}{\mathbf{1}^{\text{T}}(\boldsymbol{\lambda}+\boldsymbol{\sigma}^2)}\boldsymbol{\sigma}^2,\, \text{Diag}\left(\boldsymbol{\sigma}^2\right) - \frac{1}{\mathbf{1}^{\text{T}}(\boldsymbol{\lambda}+\boldsymbol{\sigma}^2)} \cdot \boldsymbol{\sigma}^2\left(\boldsymbol{\sigma}^2\right)^{\text{T}}\right). \tag{15}$$

As can be seen, this covariance matrix is not diagonal; it has an additive rank-1 term that vanishes as $\lambda \to \infty$. Interestingly, the location parameter of q_{mix} is translated from the *prior* location $\boldsymbol{\mu}$ along the direction of the *prior* variance vector $\boldsymbol{\sigma}^2$. This is because g_{mix} is uniform along the $K-1$ dimensions hyper-plane determined by its normal vector $\mathbf{1}$. Taking the Gaussian product then translates the location in the direction $\text{Diag}(\boldsymbol{\sigma}^2)\mathbf{1} = \boldsymbol{\sigma}^2$. Note also that the invariance-abiding posterior can be written in the same form as the true posterior in (13) (see App. D.3). The optimal invariance-abiding posterior is thus the true posterior. We visualise the two respective posterior approximations and the corresponding likelihood in Fig. 3 of App. D with two different parametrisations (cf. Sec. 4.3).

4.2 Invariance gap

To quantify the detrimental effect of the translation invariance in the considered linear model we now compute the invariance gap in (11). Note again that we assume a scaled standard normal prior with variance $\boldsymbol{\sigma}^2 = K\sigma_0^2 \cdot \mathbf{1}$. We simplify the form of the KL divergence by assuming that all variances and means take the same value in the likelihood (and thus also in the posterior) approximation, i.e.

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

This choice is motivated by the fact that the posterior approximation is a function of the sum $\mathbf{1}^{\text{T}}\boldsymbol{\lambda}$ only (cf. (15)), and, hence, it makes no difference for q_{mix}. However, since the prior variance is proportional to $\mathbf{1}$, the Gaussian likelihood resulting in the highest ELBO for the approximation q_0 also has a variance vector proportional to $\mathbf{1}$. Using this assumption, the invariance gap is

$$\text{KL}\left[q_0 \,||\, q_{\text{mix}}\right] = \frac{K-1}{2}\left[\ln\left(\frac{\hat{\sigma}^2+\hat{\lambda}}{\hat{\lambda}}\right) + \frac{\hat{\lambda}}{\hat{\sigma}^2+\hat{\lambda}} - 1\right]. \tag{16}$$

This is a convex function in $\phi = \frac{\hat{\sigma}^2+\hat{\lambda}}{\hat{\lambda}}$ with a minimum at $\phi = 1$; it is minimised as $\hat{\lambda} \to \infty$. It is however not evident 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 approximations defined in (17). The corresponding invariance gaps are visualised in Fig. 1 where it can be seen that the invariance gap grows linearly when using the optimal parameters of the invariance-abiding distribution and the unattainable data-related bound is reached quickly, thus preventing this optimum if (17b) is optimised instead of (17a).

4.3 Optimal mean-field and invariance-abiding parametrisations

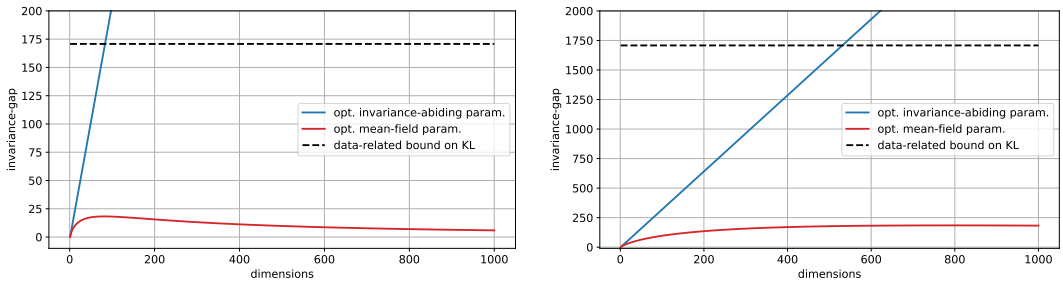
To better understand the detrimental effect of the invariance gap, we now analyse the ELL and KL terms of the ELBO objective for the mean-field and invariance-abiding variational approximations, respectively. We compare these terms for both distributions with the parameters optimized for both posterior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

$$\boldsymbol{\theta}_{\text{mix}}^* = \underset{\boldsymbol{\theta}}{\text{argmax}}\; \mathcal{L}_{\text{ELBO}}\left(q_{\text{mix}}(\cdot\,;\boldsymbol{\theta}), \mathcal{D}\right), \tag{17a}$$

$$\boldsymbol{\theta}_0^* = \underset{\boldsymbol{\theta}}{\text{argmax}}\; \mathcal{L}_{\text{ELBO}}\left(q_0(\cdot\,;\boldsymbol{\theta}), \mathcal{D}\right). \tag{17b}$$

Perhaps the simplest way compute these optimal parameters is to notice that q_{mix} in (15) can be written in the same form as the true posterior and then simply read out the optimal parameters. This is shown in App. D.3; the resulting optimal variational parameters are

$$g_0(\mathbf{w}\,;\boldsymbol{\theta}_{\text{mix}}^*) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_{\text{mix}}^*, \text{Diag}\left(\boldsymbol{\lambda}_{\text{mix}}^*\right)\right), \quad \mathbf{m}_{\text{mix}}^* = \frac{1}{N}\sum_{n=1}^{N} y^{(n)} \cdot \mathbf{1}, \quad \boldsymbol{\lambda}_{\text{mix}}^* = \frac{K\sigma_y^2}{N} \cdot \mathbf{1}. \tag{18}$$



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

Figure 1: Invariance gap evaluated at different optima (cf. (17)), 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 $\sigma^2 = K \cdot \mathbf{1}$, where K are the dimensions; the noise variance is $\sigma_y^2 = (2\pi \mathrm{e})^{-1}$; all observations $y = 1$ and $\mathbf{x} = \mathbf{1}$ are identical. As K increases, the invariance gap vanishes in case of the optimal parameters $\boldsymbol{\theta}^*_0$. In contrast, the gap for $\boldsymbol{\theta}^*_{\mathrm{mix}}$, which induces the true posterior predictive, grows linearly. As the data-related bound (cf. Sec. 2.2) can not be exceeded by the optimal parameters, $\boldsymbol{\theta}^*_0$ can not coincide with the optimal $\boldsymbol{\theta}^*_{\mathrm{mix}}$.

The optimal parameters for the mean-field *posterior* can also be computed analytically, since the Gaussian mean-field distribution that minimises the KL divergence to the multivariate Gaussian true posterior is known [34]. The resulting optimal parameters of the mean-field *likelihood* are

$$g_0(\mathbf{w}\,;\, \boldsymbol{\theta}^*_0) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}^*_0, \mathrm{Diag}\left(\boldsymbol{\lambda}^*_0\right)\right), \quad \mathbf{m}^*_0 = \mathbf{m}^*_{\mathrm{mix}}, \quad \boldsymbol{\lambda}^*_0 = \frac{K^2 \sigma_y^2}{N}\, \mathbf{1}. \tag{19}$$

As can be seen, the two respective optimal parameters differ by the factor K in the likelihood variance. We then construct the two respective posterior approximations q_0 and q_{mix} with these two optimal parameters and compare the 2×2 combinations in Fig. 2, where we visualise the ELBO terms.

Note again that we model prior variances $\boldsymbol{\sigma}^2 = K\sigma_0^2 \cdot \mathbf{1}$ such that prior and posterior over functions are identical for any K. Due to this choice and since $q_{\mathrm{mix}}(\mathbf{w}; \boldsymbol{\theta}^*_{\mathrm{mix}})$ approximates the true posterior exactly, it does not suffer from over-parametrisation. This can be seen by the loss terms in Fig. 2 being constant in K. In contrast, since $\boldsymbol{\lambda}^*_0$ depends quadratically on K and $\boldsymbol{\sigma}^2$ only linearly, $q_{\mathrm{mix}}(\mathbf{w}; \boldsymbol{\theta}^*_0)$ *collapses* to the prior as $K \to \infty$. This can be seen e.g. by the shrinking KL regularisation (Fig. 2a) and ELL (in 2b) term. As a consequence, and in line with Coker et al. [5], the posterior predictive variance of the optimal mean-field approximation reverts to the prior predictive variance (Fig. 2d).

We have shown that—in the simplified setting of a linear model with dependent observations—the consequence of not handling translation invariance is indeed posterior collapse as $K \to \infty$, while modelling the invariance coincides with the true posterior. An important observation and key takeaway is that, while we need to optimise for (17a), the mean-field posterior approximation $q_0(\mathbf{w}; \boldsymbol{\theta}^*_{\mathrm{mix}})$ is sufficient for prediction. Indeed, for the linear model, we could compute the invariance gap exactly and thereby correct the ELBO objective for the invariances of this model. For more complex non-linear models such as BNNs, this section may serve as a direction for approximating the invariance gap. To this end, the next section describes the layer-wise translation invariance exhibited by BNNs.

5 Translation invariance in Bayesian neural networks

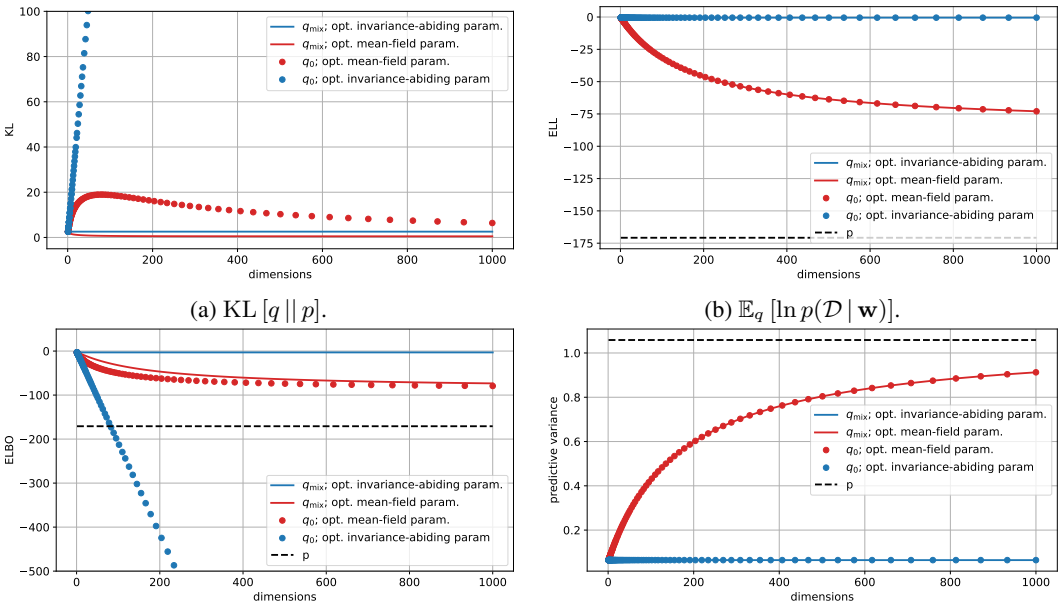
Let us now go back to the general NN model in Sec. 2.1. In order to describe the set of all invariances, note that for each node $z_{l,j}$ in layer l (or the output node y), there is a subspace that keeps the value $z_{l,j}$ (or y) invariant when using the translation-invariance model in (12), because the value itself only depends on the *actual* activation $\mathbf{z}_{l-1}$ (or the input $\mathbf{x}$).

More formally, the following equations model all translation invariances of a NN at a data point $\mathbf{x}$:

$$\forall \boldsymbol{\Delta}_{L,1}: \quad f(\mathbf{x}) = h_L\left(\mathbf{w}_{L,1}^{\mathrm{T}} \mathbf{z}_{L-1}\right) = h_L\left(\left(\mathbf{w}_{L,1} + \mathbf{B}_{\mathbf{z}_{L-1}} \boldsymbol{\Delta}_{L,1}\right)^{\mathrm{T}} \mathbf{z}_{L-1}\right), \tag{20}$$

$$\forall j \in \{1, \ldots, n_l\}, \boldsymbol{\Delta}_{l,j}: \quad z_{l,j} = h_l\left(\mathbf{w}_{l,j}^{\mathrm{T}} \mathbf{z}_{l-1}\right) h_l\left(\left(\mathbf{w}_{l,j} + \mathbf{B}_{\mathbf{z}_{l-1}} \boldsymbol{\Delta}_{l,j}\right)^{\mathrm{T}} \mathbf{z}_{l-1}\right), \tag{21}$$

$$\forall j \in \{1, \ldots, n_1\}, \boldsymbol{\Delta}_{1,j}: \quad z_{1,j} = h_1\left(\mathbf{w}_{1,j}^{\mathrm{T}} \mathbf{x}\right) = h_1\left(\left(\mathbf{w}_{1,j} + \mathbf{B}_{\mathbf{x}} \boldsymbol{\Delta}_{1,j}\right)^{\mathrm{T}} \mathbf{x}\right), \tag{22}$$



(a) KL $[q \,||\, p]$.

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

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

(d) $\mathbb{V}_{\mathbf{w}\sim q}\left[\frac{1}{K}\mathbf{1}^{\mathrm{T}}\mathbf{w}\right] + \sigma_y^2$.

Figure 2: 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. (17)). Prior variance is $\boldsymbol{\sigma}^2 = K \cdot \mathbf{1}$, and $\sigma_y^2 = 1/(2\pi\mathrm{e})$; 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 ELL (b) and predictive variance (d). The variance of the mean-field approximation reverts to the prior variance as $K \to \infty$.

where the layer index l ranges from $2, \ldots, L-1$ and $\mathbf{B}_{\mathbf{z}}$ is 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 invariance-abiding likelihood g_{mix}, note that it also decomposes over the NN layers and the associated parameters as follows: For the n_1 nodes in the first layer we have

$$q_{\text{mix}}(\mathbf{w}_{1,j}) = \mathbb{E}_{\mathbf{x}}\left[\mathcal{N}\left(\mathbf{w}_{1,j}; \boldsymbol{\mu} + \frac{\mathbf{x}^{\mathrm{T}}(\mathbf{m}-\boldsymbol{\mu})}{\mathbf{x}^{\mathrm{T}}(\mathbf{V}+\boldsymbol{\Sigma})\mathbf{x}}(\boldsymbol{\Sigma}\mathbf{x}), \boldsymbol{\Sigma} - \frac{1}{\mathbf{x}^{\mathrm{T}}(\mathbf{V}+\boldsymbol{\Sigma})\mathbf{x}}(\boldsymbol{\Sigma}\mathbf{x})(\boldsymbol{\Sigma}\mathbf{x})^{\mathrm{T}}\right)\right], \tag{23}$$

$$P(z_{1,j}) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{w}\sim q_{\text{mix}}(\mathbf{w}_{1,j})}\left[h_1\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}\right)\right]\right], \tag{24}$$

where the outer expectation is taken over $\mathbf{x} \sim p(\mathcal{D})$ and $p(\mathcal{D})$ is the empirical distribution over the training set $\mathcal{D}$, and where $\mathbf{m}$, $\mathbf{V}$, $\boldsymbol{\mu}$ 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* $\mathbf{z}_{l-1}$ of the hidden units in layer $n-1$ (see (20) and (21)), we have

$$q_{\text{mix}}(\mathbf{w}_{l,j}) = \mathbb{E}_{\mathbf{z}}\left[\mathcal{N}\left(\mathbf{w}_{l,j}; \boldsymbol{\mu} + \frac{\mathbf{z}^{\mathrm{T}}(\mathbf{m}-\boldsymbol{\mu})}{\mathbf{z}^{\mathrm{T}}(\mathbf{V}+\boldsymbol{\Sigma})\mathbf{z}}(\boldsymbol{\Sigma}\mathbf{z}), \boldsymbol{\Sigma} - \frac{1}{\mathbf{z}^{\mathrm{T}}(\mathbf{V}+\boldsymbol{\Sigma})\mathbf{z}}(\boldsymbol{\Sigma}\mathbf{z})(\boldsymbol{\Sigma}\mathbf{z})^{\mathrm{T}}\right)\right], \tag{25}$$

$$P(z_{l,j}) = \mathbb{E}_{\mathbf{z}}\left[\mathbb{E}_{\mathbf{w}\sim q_{\text{mix}}(\mathbf{w}_{l,j})}\left[h_l\left(\mathbf{w}^{\mathrm{T}}\mathbf{z}_{l-1}\right)\right]\right], \tag{26}$$

where again $\mathbf{m}$, $\mathbf{V}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are constrained to the parts of the overall weight vector that correspond to $\mathbf{w}_{l,j}$ and the expectation over $\mathbf{z}$ is taken with respect to $\mathbf{z} \sim P(\mathbf{z}_{l-1})$.

Thus, the invariance-abiding approximation could e.g. be computed with a layer-by-layer iterative optimisation: First, given data $\mathbf{x} \sim p(\mathcal{D})$ and prior $p(\mathbf{w})$, use (23) and point estimates for all weight vectors in later layers to optimise the mean $\mathbf{m}$ and diagonal covariance $\mathbf{V}$ of the weights in the first layer, i.e., $\{\mathbf{w}_{1,j} | j = 1, \ldots, n_1\}$. Then, we can use (24) to generate P samples $\mathbf{z}_{1,n}$ of the activations of the first layer under the (now fitted) optimal $q_{\text{mix}}(\{\mathbf{w}_{1,j}\})$. Next, for each of these samples $\mathbf{z}_{1,n}$, we can use (25) and point estimates for all weight vectors in later layers to optimise the mean $\mathbf{m}$ and diagonal covariance $\mathbf{V}$ of the weights in the second layer and average them for the optimal q_{mix} of the weights in the second layer. Finally, we can use (26) to generate samples $\mathbf{z}_{2,i}$ for the second layer.

This process is repeated until the last layer, and iterated until $q_{\text{mix}}(\mathbf{w})$ converges. The complexity of this procedure is no larger than optimising the weight matrices of a single layer, because all previous layers are fully characterised by the distribution of the latent activations $\mathbf{z}_l$ of the hidden layers.

6 Related work

Poor empirical performance of VBNNs has been reported in several previous works, e.g., [37, 16, 33]. For instance, it has been shown that single-layer mean-field VBNNs with ReLU activations can not have large predictive uncertainty between regions of low uncertainty [10]. In contrast, Farquhar et al. [8] 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 conflict with our work and [5], who show 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. [8] 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. Our framework to model the invariances is similar to and extends previous preliminary works [19, 22].

Surprisingly, restricting the parametrisation of the variational posterior results in competitive or even better performance [30, 21, 31, 7]. For example, Dusenberry et al. [7] 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 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.

Other attempts have proposed to approximate the posterior predictive directly, thereby circumventing the non-identifiability issue [28, 20, 36]. While these approaches have shown promising empirical performance, estimating the KL regularisation term in function space is more complicated and can even be ill-defined due to regions of zero prior probability mass [4]. In a similar vein, previous work also proposed to map the BNN prior to a Gaussian process (functional) prior [9, 32]. Another promising direction to circumvent invariance in layered models are deep kernel processes, in which Gram matrices are progressively transformed by nonlinear kernel functions [1]. The Gram matrices, which are treated as the random variables, are invariant to permutations/rotations of the weights/features.

More broadly, (non-)identifiability of parameters and latent variables in probabilistic models is a widely-studied topic both from frequentist [27] and Bayesian perspectives [6, 25, 11, 24]. In a recent example, Wang et al. [35] attributed posterior collapse in the context of variational auto-encoders to non-identifiability of latent variables.

7 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. We related this to a data-related bound on the KL regularisation, which prevents fits for which the gap is large.

We studied over-parametrised Bayesian linear regression as the canonical model that exhibits *translation invariance* and 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 models or for other invariances. However, our work provides the mathematical tools to address over-regularisation due to invariances. Future work 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] L. Aitchison, A. Yang, and S. W. Ober. Deep kernel processes. 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 130–140. PMLR, 18–24 Jul 2021.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

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

[10] A. Foong, D. Burt, Y. Li, and R. Turner. On the expressiveness of approximate inference in Bayesian neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15897–15908. Curran Associates, Inc., 2020.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] D. A. Moore. Symmetrized variational inference. *NIPS Workshop on Advances in Approximate Bayesian Inference*, December 2016.

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

[24] R. Nishihara, T. P. Minka, and D. Tarlow. Detecting parameter symmetries in probabilistic models. *arXiv: Machine Learning*, 2013.

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

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

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[35] 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.

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

[37] 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.

[38] A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708. Curran Associates, Inc., 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.

A Notation

We denote matrices and vectors with bold upper- and lower-case letters, e.g. $\mathbf{a}$ and $\mathbf{A}$. In order to address an element of a vector or matrix, we will use non-bold letters, e.g. $\mathbf{x} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}}$. We will use the superscript $^{\mathrm{T}}$ to denote a transpose of a vector and $|\mathbf{A}|$ to denote the determinant of the matrix $\mathbf{A}$. We use $\mathrm{Diag}(\mathbf{a})$ to denote the diagonal matrix constructed from a vector, and $\mathrm{diag}(\mathbf{A})$ to denote the diagonal vector of a matrix. We use the notation $\mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{V})$ to denote the density of the Gaussian distribution given by

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

An alternative representation of the Gaussian distribution is in terms of its canonical parameters $\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$:

$$\mathcal{G}(\mathbf{x}; \boldsymbol{\eta}, \boldsymbol{\Lambda}) := \exp\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\eta} - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{x} - \frac{1}{2}\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta} - \frac{1}{2}\ln\left(|2\pi\boldsymbol{\Lambda}^{-1}|\right)\right).$$

Note that $\mathcal{G}(\mathbf{x}; \boldsymbol{\eta}_1, \boldsymbol{\Lambda}_1) \cdot \mathcal{G}(\mathbf{x}; \boldsymbol{\eta}_2, \boldsymbol{\Lambda}_2) \propto \mathcal{G}(\mathbf{x}; \boldsymbol{\eta}_1 + \boldsymbol{\eta}_2, \boldsymbol{\Lambda}_1 + \boldsymbol{\Lambda}_2)$ which renders this canonical parameterisation very useful when considering the multiplication of Gaussian densities. Furthermore, we use $\mathbb{E}_{\mathbf{w}\sim p}[f(\mathbf{w})]$ to denote the expectation of the function $f(\cdot)$ when $\mathbf{w}$ is drawn from the

distribution p. Also, we use $\mathbb{V}_{\mathbf{w}\sim p}\left[f(\mathbf{w})\right]$ to denote the variance of the function $f(\cdot)$ when $\mathbf{w}$ is drawn from the distribution p defined by

$$\mathbb{V}_{\mathbf{w}\sim p}\left[f(\mathbf{w})\right] := \mathbb{E}_{\mathbf{w}\sim p}\left[\left(f(\mathbf{w}) - \mathbb{E}_{\mathbf{w}'\sim p}\left[f(\mathbf{w}')\right]\right)^2\right] .$$

B Convolution of Gaussian Measures

Theorem 1 (Convolution of Gaussian Measures)**.** *For any $\mathbf{A} \in \mathbb{R}^{n\times k}$ and $\mathbf{b} \in \mathbb{R}^n$ it holds that*

$$p\left(\mathbf{x}|\boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{x}; \mathbf{A}\boldsymbol{\theta} + \mathbf{b}, \mathbf{V}\right) \ , \ p\left(\boldsymbol{\theta}\right) = \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) ,$$

implies that

$$p\left(\boldsymbol{\theta}|\mathbf{x}\right) = \mathcal{N}\left(\boldsymbol{\theta}; \mathbf{C}^{-1}\left(\mathbf{A}^T\mathbf{V}^{-1}\left(\mathbf{x}-\mathbf{b}\right) + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \mathbf{C}^{-1}\right) , \tag{27}$$

$$p\left(\mathbf{x}\right) = \mathcal{N}\left(\mathbf{x}; \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{V} + \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\right) , \tag{28}$$

$$\mathbf{C} = \mathbf{A}^T\mathbf{V}^{-1}\mathbf{A} + \boldsymbol{\Sigma}^{-1} .$$

Proof. Let's start by proving (27). First, we note that

$$p\left(\boldsymbol{\theta}|\mathbf{x}\right) = \frac{p\left(\mathbf{x}|\boldsymbol{\theta}\right)\cdot p\left(\boldsymbol{\theta}\right)}{\int p\left(\mathbf{x}|\widetilde{\boldsymbol{\theta}}\right)\cdot p\left(\widetilde{\boldsymbol{\theta}}\right)\, d\widetilde{\boldsymbol{\theta}}} = \frac{\mathcal{N}\left(\mathbf{x}; \mathbf{A}\boldsymbol{\theta}+\mathbf{b}, \mathbf{V}\right)\cdot\mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)}{\int \mathcal{N}\left(\mathbf{x}; \mathbf{A}\widetilde{\boldsymbol{\theta}}+\mathbf{b}, \mathbf{V}\right)\cdot\mathcal{N}\left(\widetilde{\boldsymbol{\theta}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\, d\widetilde{\boldsymbol{\theta}}} ,$$

where only the numerator depends on $\boldsymbol{\theta}$. Thus, the numerator is given by

$$c\cdot\exp\left(-\frac{1}{2}\left[\left(\left(\mathbf{x}-\mathbf{b}\right)-\mathbf{A}\boldsymbol{\theta}\right)^{\mathrm{T}}\mathbf{V}^{-1}\left(\left(\mathbf{x}-\mathbf{b}\right)-\mathbf{A}\boldsymbol{\theta}\right) + \left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}-\boldsymbol{\mu}\right)\right]\right) ,$$

where $c = \left|2\pi\mathbf{V}\right|^{-\frac{1}{2}}\cdot\left|2\pi\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}$ is independent of $\mathbf{x}$ and $\boldsymbol{\theta}$. Using Theorem A.86 in [14], this quadratic form in $\boldsymbol{\theta}$ can be rewritten as

$$c\cdot\exp\left(-\frac{1}{2}\left[\left(\boldsymbol{\theta}-\mathbf{c}\right)^{\mathrm{T}}\mathbf{C}\left(\boldsymbol{\theta}-\mathbf{c}\right) + d\left(\mathbf{x}\right)\right]\right) ,$$

where

$$\begin{aligned}
\mathbf{C} &= \mathbf{A}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{A} + \boldsymbol{\Sigma}^{-1} ,\\
\mathbf{C}\mathbf{c} &= \mathbf{A}^{\mathrm{T}}\mathbf{V}^{-1}\left(\mathbf{x}-\mathbf{b}\right) + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} ,\\
d\left(\mathbf{x}\right) &= \left(\mathbf{x}-\mathbf{b}-\mathbf{A}\boldsymbol{\mu}\right)^{\mathrm{T}}\left(\mathbf{V}+\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right)^{-1}\left(\mathbf{x}-\mathbf{b}-\mathbf{A}\boldsymbol{\mu}\right) .
\end{aligned} \tag{29}$$

Since $d(\mathbf{x})$ does not depend on $\boldsymbol{\theta}$, it get incorporated into the normalization constant which proves (27).

In order to prove (28), note that

$$\begin{aligned}
p\left(\mathbf{x}\right) &= \int p\left(\mathbf{x}|\widetilde{\boldsymbol{\theta}}\right)\cdot p\left(\widetilde{\boldsymbol{\theta}}\right)\, d\widetilde{\boldsymbol{\theta}}\\
&= \int \mathcal{N}\left(\mathbf{x}; \mathbf{A}\widetilde{\boldsymbol{\theta}}+\mathbf{b}, \mathbf{V}\right)\cdot\mathcal{N}\left(\widetilde{\boldsymbol{\theta}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\, d\widetilde{\boldsymbol{\theta}}\\
&= \int c\cdot\exp\left(-\frac{1}{2}\left[\left(\widetilde{\boldsymbol{\theta}}-\mathbf{c}\right)^{\mathrm{T}}\mathbf{C}\left(\widetilde{\boldsymbol{\theta}}-\mathbf{c}\right) + d\left(\mathbf{x}\right)\right]\right)\, d\widetilde{\boldsymbol{\theta}}\\
&= c\cdot\exp\left(-\frac{1}{2}d\left(\mathbf{x}\right)\right)\cdot\int\exp\left(-\frac{1}{2}\left[\left(\widetilde{\boldsymbol{\theta}}-\mathbf{c}\right)^{\mathrm{T}}\mathbf{C}\left(\widetilde{\boldsymbol{\theta}}-\mathbf{c}\right)\right]\right)\, d\widetilde{\boldsymbol{\theta}}\\
&= c\cdot\exp\left(-\frac{1}{2}d\left(\mathbf{x}\right)\right)\cdot\left|2\pi\mathbf{C}^{-1}\right|^{\frac{1}{2}}\\
&= \tilde{c}\cdot\exp\left(-\frac{1}{2}\left(\left(\mathbf{x}-\left(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}\right)\right)^{\mathrm{T}}\left(\mathbf{V}+\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right)^{-1}\left(\mathbf{x}-\left(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}\right)\right)\right)\right)\\
&= \mathcal{N}\left(\mathbf{x}; \mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{V}+\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\right) ,
\end{aligned}$$

where the penultimate line follows again from Theorem A.86 in [14] with $d(\mathbf{x})$ defined in (29). □

Theorem 2 (Woodbury formula)**.** *Let $\mathbf{C}$ be an invertible $n\times n$ matrix. Then, for any $\mathbf{A}\in\mathbb{R}^{n\times k}$ and $\mathbf{B}\in\mathbb{R}^{k\times n}$,*

$$\left(\mathbf{C}+\mathbf{A}\mathbf{B}\right)^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{A}\left(\mathbf{I}+\mathbf{B}\mathbf{C}^{-1}\mathbf{A}\right)^{-1}\mathbf{B}\mathbf{C}^{-1} .$$

C Invariance gap and posterior-predictive equivalence

In this section, we show that the posterior predictive distribution of the mean-field and invariance-abiding distribution are identical (for identical parametrisation of the likelihood approximation) as stated in (9), and that the difference in the respective KL regularisation terms can be quantified by the KL defined in (10).

We first discuss the conditions from (7) and (8). The goal is to use the invariance property of the likelihood function (4). Unfortunately, the prior does not generally have the same invariance. Yet, the mean-field approximate posterior defined as the product between prior and likelihood approximation can still have the invariance property, as we show for permutation and translation invariance with mean-field Gaussian likelihood approximation and prior. This condition is defined in (7) for each mean-field posterior in the integral over $\mathbf{r}$:

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

where Z_{mix} and $Z_0(\mathbf{r})$ are normalisation constants. The second line introduced the assumption that there exists a surjective mapping $\mathbf{r}' = \varphi(\mathbf{r})$ such that the product of the untransformed prior $p(\mathbf{w})$ and the transformed likelihood approximation $g_0(t(\mathbf{w},\varphi(\mathbf{r}));\boldsymbol{\theta})$ are identical to the product of the transformed prior and likelihood approximation with $\mathbf{r}'$, i.e. as stated in the main text,

$$\forall \mathbf{r} \sim p(\mathbf{r}) : p(\mathbf{w}) \cdot g_0(t(\mathbf{w},\mathbf{r});\boldsymbol{\theta}) = p(t(\mathbf{w},\varphi(\mathbf{r}))) \cdot g_0(t(\mathbf{w},\varphi(\mathbf{r}));\boldsymbol{\theta}).$$

This condition may seem quite limiting, however, we show that this condition holds for *permutation* invariance with the isotropic Gaussian prior and for *translation* invariance with Gaussian priors and Gaussian likelihood approximation.

Condition (7) **and** (8) **for permutation invariance.** Consider first permutation invariance (cf. App. E). It is easy to see that for the isotropic Gaussian prior: $p(\mathbf{w}) = p(\mathbf{P_r w})$. Also, note that $\mathcal{G}(\mathbf{P}\mathbf{w};\, \boldsymbol{\eta}, \boldsymbol{\Lambda}) = \mathcal{G}(\mathbf{w}; \mathbf{P}^{\mathrm{T}}\boldsymbol{\eta}, \mathbf{P}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{P})$. Thus, the product between the untransformed Gaussian prior and the permuted Gaussian likelihood approximation is

$$\begin{aligned} p(\mathbf{w}) \cdot g_0(\mathbf{P}\mathbf{w}) = p(\mathbf{P}\mathbf{w}) \cdot g_0(\mathbf{P}\mathbf{w}) &= \mathcal{G}(\mathbf{P}\mathbf{w}; \boldsymbol{\eta}_{\text{p}}, \boldsymbol{\Lambda}_{\text{p}}) \cdot \mathcal{G}(\mathbf{P}\mathbf{w}; \boldsymbol{\eta}_g, \boldsymbol{\Lambda}_g) && (30) \\ &= \mathcal{G}(\mathbf{w}; \mathbf{P}^{\mathrm{T}}\boldsymbol{\eta}_{\text{p}}, \mathbf{P}^{\mathrm{T}}\boldsymbol{\Lambda}_{\text{p}}\mathbf{P}) \cdot \mathcal{G}(\mathbf{w}; \mathbf{P}^{\mathrm{T}}\boldsymbol{\eta}_g, \mathbf{P}^{\mathrm{T}}\boldsymbol{\Lambda}_g\mathbf{P}) && (31) \\ &\propto \mathcal{G}(\mathbf{w}; \mathbf{P}^{\mathrm{T}}(\boldsymbol{\eta}_{\text{p}} + \boldsymbol{\eta}_g), \mathbf{P}^{\mathrm{T}}(\boldsymbol{\Lambda}_{\text{p}} + \boldsymbol{\Lambda}_g)\mathbf{P}) && (32) \\ &= \mathcal{G}(\mathbf{P}\mathbf{w}; \boldsymbol{\eta}_{\text{p}} + \boldsymbol{\eta}_g, \boldsymbol{\Lambda}_{\text{p}} + \boldsymbol{\Lambda}_g) = q_0(\mathbf{P}\mathbf{w}). && (33) \end{aligned}$$

Hence, for permutation invariance with the isotropic Gaussian prior and Gaussian likelihood approximation, we have

$$\forall \mathbf{r} \sim p(\mathbf{r}) : p(\mathbf{w}) \cdot g_0(\mathbf{P_r w}) = p(\mathbf{P_r w}) \cdot g_0(\mathbf{P_r w}) \propto q_0(\mathbf{P_r w}). \tag{34}$$

The invariance transformation is thus simply $t(\mathbf{w},\varphi(\mathbf{r})) = t(\mathbf{w},\mathbf{r}) = \mathbf{P_r w}$ (i.e. we do not need the mapping φ). This is because the prior has no preference over the permutation-induced modes of the likelihood. Thus, we have

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

Condition (7) **and** (8) **for translation invariance.** Next, consider translation invariance, and note that $\mathcal{N}(\mathbf{w} - \mathbf{v}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu} + \mathbf{v}, \boldsymbol{\Sigma})$, and, consequently, $\mathcal{G}(\mathbf{w} - \mathbf{v}; \boldsymbol{\eta}, \boldsymbol{\Lambda}) = \mathcal{G}(\mathbf{w}; \boldsymbol{\eta} + \boldsymbol{\Lambda}\mathbf{v}, \boldsymbol{\Lambda})$. It is easier to show directly that the product between the untransformed Gaussian prior and translated

Gaussian likelihood approximation can be written as a Gaussian posterior translated, because:

$$
\begin{aligned}
p(\mathbf{w}) \cdot g_0(\mathbf{w}-\mathbf{v}) &= \mathcal{G}(\mathbf{w};\boldsymbol{\eta}_\mathrm{p},\boldsymbol{\Lambda}_\mathrm{p}) \cdot \mathcal{G}(\mathbf{w}-\mathbf{v};\boldsymbol{\eta}_g,\boldsymbol{\Lambda}_g) && (35)\\
&= \mathcal{G}(\mathbf{w};\boldsymbol{\eta}_\mathrm{p},\boldsymbol{\Lambda}_\mathrm{p}) \cdot \mathcal{G}(\mathbf{w};\boldsymbol{\eta}_g+\boldsymbol{\Lambda}_g\mathbf{v},\boldsymbol{\Lambda}_g) && (36)\\
&\propto \mathcal{G}(\mathbf{w};\boldsymbol{\eta}_\mathrm{p}+\boldsymbol{\eta}_g+\boldsymbol{\Lambda}_g\mathbf{v},\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g) && (37)\\
&= \mathcal{G}\Big(\mathbf{w};\boldsymbol{\eta}_\mathrm{p}+\boldsymbol{\eta}_g+\overbrace{(\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g)\,(\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g)^{-1}}^{\mathbf{I}}\boldsymbol{\Lambda}_g\mathbf{v},\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g\Big) && (38)\\
&= \mathcal{G}\left(\mathbf{w}-(\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g)^{-1}\boldsymbol{\Lambda}_g\mathbf{v};\boldsymbol{\eta}_\mathrm{p}+\boldsymbol{\eta}_g,\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g\right) && (39)\\
&= q_0\left(\mathbf{w}-(\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g)^{-1}\boldsymbol{\Lambda}_g\mathbf{v}\right). && (40)
\end{aligned}
$$

Since the precision matrices are diagonal, $((\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g)^{-1}\boldsymbol{\Lambda}_g)\mathbf{B}\mathbf{r} = \mathbf{B}((\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g)^{-1}\boldsymbol{\Lambda}_g)\mathbf{r}$. Consequently, for the translation invariance $g_0(\mathbf{w}) = g_0(t(\mathbf{w},\mathbf{r})) = g_0(\mathbf{w}-\mathbf{B}\mathbf{r})$, we have

$$
\forall \mathbf{r}\sim p(\mathbf{r}) : p(\mathbf{w})\cdot g_0(\mathbf{w}-\mathbf{B}\mathbf{r}) = p\left(\mathbf{w}-\mathbf{B}\mathbf{r}'\right)\cdot g_0\left(\mathbf{w}-\mathbf{B}\mathbf{r}'\right) \propto q_0\left(\mathbf{w}-\mathbf{B}\mathbf{r}'\right), \tag{41}
$$

where $\mathbf{r}' = \varphi(\mathbf{r}) = \left((\boldsymbol{\Lambda}_\mathrm{p}+\boldsymbol{\Lambda}_g)^{-1}\boldsymbol{\Lambda}_g\right)\mathbf{r}$. Thus, we have

$$
\forall \mathbf{r}\sim p(\mathbf{r}) : \left|\det\frac{\partial t(\mathbf{w},\mathbf{r})}{\partial \mathbf{w}}\right|^{-1} = \left|\det\frac{\partial(\mathbf{w}-\mathbf{B}\mathbf{r})}{\partial \mathbf{w}}\right|^{-1} = |\det\mathbf{I}|^{-1} = 1\,.
$$

Lemma 1. *For any distribution $p(\mathbf{w})$, likelihood approximation $g_0(\mathbf{w};\boldsymbol{\theta})$, $q_0(\mathbf{w};\boldsymbol{\theta})$ as defined in* (6)*, $q_\mathrm{mix}(\mathbf{w};\boldsymbol{\theta})$ as defined in* (6b) *and $p(\mathbf{r})$, assume there exists a mapping $\varphi:\mathbf{r}\mapsto\mathbf{r}'$ such that*

$$
\forall \mathbf{r}\sim p(\mathbf{r}) : p(\mathbf{w})\cdot g_0(t(\mathbf{w},\mathbf{r});\boldsymbol{\theta}) = p(t(\mathbf{w},\varphi(\mathbf{r})))\cdot g_0(t(\mathbf{w},\varphi(\mathbf{r}));\boldsymbol{\theta}), \tag{42}
$$

as well as

$$
\forall \mathbf{r}\sim p(\mathbf{r}) : \left|\det\frac{\partial t(\mathbf{w},\mathbf{r})}{\partial \mathbf{w}}\right|^{-1} = 1\,. \tag{43}
$$

Then,

$$
\mathbb{E}_{\mathbf{w}\sim q_\mathrm{mix}(\mathbf{w};\boldsymbol{\theta})}\left[\ln p(\mathcal{D}\,|\,\mathbf{w})\right] = \mathbb{E}_{\mathbf{w}\sim q_0(\mathbf{w};\boldsymbol{\theta})}\left[\ln p(\mathcal{D}\,|\,\mathbf{w})\right]\,. \tag{44}
$$

Proof. The lemma can be proven by applying the change-of-variables formula and using the invariance property (42) of $g(\cdot;\boldsymbol{\theta})$:

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{w}\sim q_\mathrm{mix}(\mathbf{w};\boldsymbol{\theta})}\left[\ln p(\mathcal{D}\,|\,\mathbf{w})\right] && (45)\\
&= \int \overbrace{\int p(\mathbf{r})\frac{Z_0(\mathbf{r})}{Z_\mathrm{mix}} q_0(t(\mathbf{w},\varphi(\mathbf{r}));\,\boldsymbol{\theta})\,d\mathbf{r}}^{q_\mathrm{mix}(\mathbf{w})} \ln\left[p(\mathcal{D}\,|\,\mathbf{w})\right]\,d\mathbf{w} && (46)\\
&= \int p(\mathbf{r})\frac{Z_0(\mathbf{r})}{Z_\mathrm{mix}}\int q_0(t(\mathbf{w},\varphi(\mathbf{r}));\,\boldsymbol{\theta})\,\ln\left[p(\mathcal{D}\,|\,\mathbf{w})\right]\,d\mathbf{w}\,d\mathbf{r} && (47)\\
&= \int p(\mathbf{r})\frac{Z_0(\mathbf{r})}{Z_\mathrm{mix}}\int q_0(t(\mathbf{w},\varphi(\mathbf{r}));\,\boldsymbol{\theta})\,\ln\left[p(\mathcal{D}\,|\,t(\mathbf{w},\varphi(\mathbf{r})))\right]\,d\mathbf{w}\,d\mathbf{r} && (48)\\
&= \int p(\mathbf{r})\frac{Z_0(\mathbf{r})}{Z_\mathrm{mix}}\int q_0(\mathbf{w};\,\boldsymbol{\theta})\overbrace{\left|\det\frac{\partial t(\mathbf{w},\varphi(\mathbf{r}))}{\partial \mathbf{w}}\right|^{-1}}^{1}\ln\left[p(\mathcal{D}\,|\,\mathbf{w})\right]\,d\mathbf{w}\,d\mathbf{r} && (49)\\
&= \overbrace{\int p(\mathbf{r})\frac{Z_0(\mathbf{r})}{Z_\mathrm{mix}}d\mathbf{r}}^{1}\int q_0(\mathbf{w};\,\boldsymbol{\theta})\ln\left[p(\mathcal{D}\,|\,\mathbf{w})\right]\,d\mathbf{w} && (50)\\
&= \mathbb{E}_{\mathbf{w}\sim q_0}\left[\ln p(\mathcal{D}\,|\,\mathbf{w})\right]. && (51)
\end{aligned}
$$

where the second line uses (42), the third line changes the order of integration (Fubini's theorem), the fourth line uses the invariance property $\ln p(\mathcal{D}\,|\,\mathbf{w}) = \ln p(\mathcal{D}\,|\,t(\mathbf{w},\varphi(\mathbf{r})))$, the fifth line then applies

the change of variables theorem together with (43), the sixth line then uses again the invariance property of the log likelihood, and the seventh line notes that the integral over the normalisation constants $Z_0(\mathbf{r})$ cancels with the normalisation constant Z_{mix} of the invariance-abiding posterior, since

$$\int p(\mathbf{r}) Z_0(\mathbf{r}) d\mathbf{r} = \int\int p(\mathbf{r})\, p(\mathbf{w}) \cdot g_0(t(\mathbf{w}, \mathbf{r}))\, d\mathbf{w} d\mathbf{r} \tag{52}$$

$$= \int p(\mathbf{w}) \cdot \int p(\mathbf{r})\, g_0(t(\mathbf{w}, \mathbf{r}))\, d\mathbf{r} d\mathbf{w} \tag{53}$$

$$= \int p(\mathbf{w}) \cdot g_{\text{mix}}(\mathbf{w})\, d\mathbf{w} = Z_{\text{mix}}, \tag{54}$$

where we changed the integration order, resulting in the normalisation constant of the invariance-abiding likelihood approximation. □

Lemma 2. *For any distribution $p(\mathbf{w})$, likelihood approximation $g_0(\mathbf{w}; \boldsymbol{\theta})$, $q_0(\mathbf{w}; \boldsymbol{\theta})$ as defined in* (6)*, $q_{\text{mix}}(\mathbf{w}; \boldsymbol{\theta})$ as defined in* (6b) *and $p(\mathbf{r})$, assume there exist two mappings $t : \mathbf{w} \times \mathbf{r} \mapsto \mathbf{w}'$ and $\varphi : \mathbf{r} \mapsto \mathbf{r}'$ such that* (42) *and* (43) *hold. Then,*

$$\text{KL}\left[q_0 \,||\, p\right] - \text{KL}\left[q_{\text{mix}} \,||\, p\right] = \text{KL}\left[q_0 \,||\, q_{\text{mix}}\right] . \tag{55}$$

Proof. First, note that

$$\text{KL}\left[q_0 \,||\, p\right] = \mathbb{E}_{\mathbf{w}\sim q_0}\left[\ln\left[Z_0\, g_0(\mathbf{w})\right]\right] , \tag{56}$$

$$\text{KL}\left[q_{\text{mix}} \,||\, p\right] = \mathbb{E}_{\mathbf{w}\sim q_{\text{mix}}}\left[\ln\left[Z_{\text{mix}}\, g_{\text{mix}}(\mathbf{w})\right]\right] . \tag{57}$$

Now, in the latter KL, we expand the distribution $q_{\text{mix}}(\mathbf{w})$ as in (42),

$$\text{KL}\left[q_{\text{mix}} \,||\, p\right] = \overbrace{\int\int p(\mathbf{r}) \frac{Z_0(\mathbf{r})}{Z_{\text{mix}}} q_0(t(\mathbf{w}, \varphi(\mathbf{r})); \boldsymbol{\theta}) d\mathbf{r}}^{q_{\text{mix}}(\mathbf{w})} \ln\left[Z_{\text{mix}}\, g_{\text{mix}}(\mathbf{w})\right]\, d\mathbf{w} \tag{58}$$

$$= \int\int p(\mathbf{r}) \frac{Z_0(\mathbf{r})}{Z_{\text{mix}}} q_0(t(\mathbf{w}, \varphi(\mathbf{r})); \boldsymbol{\theta})\, \ln\left[Z_{\text{mix}}\, g_{\text{mix}}(\mathbf{w})\right]\, d\mathbf{r} d\mathbf{w} \tag{59}$$

$$= \int\int p(\mathbf{r}) \frac{Z_0(\mathbf{r})}{Z_{\text{mix}}} q_0(t(\mathbf{w}, \varphi(\mathbf{r})); \boldsymbol{\theta})\, \ln\left[Z_{\text{mix}}\, g_{\text{mix}}(t(\mathbf{w}, \varphi(\mathbf{r})))\right]\, d\mathbf{r} d\mathbf{w} \tag{60}$$

$$= \int\int p(\mathbf{r}) \frac{Z_0(\mathbf{r})}{Z_{\text{mix}}} q_0(\mathbf{w}; \boldsymbol{\theta}) \overbrace{\left|\det \frac{\partial t(\mathbf{w}, \varphi(\mathbf{r}))}{\partial \mathbf{w}}\right|^{-1}}^{1} \ln\left[Z_{\text{mix}}\, g_{\text{mix}}(\mathbf{w})\right]\, d\mathbf{r} d\mathbf{w} \tag{61}$$

$$= \int q_0(\mathbf{w}; \boldsymbol{\theta}) \overbrace{\int p(\mathbf{r}) \frac{Z_0(\mathbf{r})}{Z_{\text{mix}}}\, d\mathbf{r}}^{1} \ln\left[Z_{\text{mix}}\, g_{\text{mix}}(\mathbf{w})\right] d\mathbf{w} \tag{62}$$

$$= \int q_0(\mathbf{w}; \boldsymbol{\theta})\, \ln\left[Z_{\text{mix}}\, g_{\text{mix}}(\mathbf{w})\right]\, d\mathbf{w}' , \tag{63}$$

where the third line uses the invariance property of the invariance-abiding likelihood approximation, $\forall \mathbf{r} : g_{\text{mix}}(t(\mathbf{w}, \varphi(\mathbf{r}))) = g_{\text{mix}}(\mathbf{w})$, the change of variables formula is applied to the fourth line for the volume-preserving transformation, and the fifth line re-arranges the integration, noting that the integral over normalisation constants equals one.

The proposition follows by taking the difference between the regularisation terms corresponding to the respective ELBO objectives:

$$\text{KL}\left[q_0 \,||\, p\right] - \text{KL}\left[q_{\text{mix}} \,||\, p\right] = \mathbb{E}_{\mathbf{w}\sim q_0}\left[\ln \frac{Z_0\, g_0(\mathbf{w})}{Z_{\text{mix}} g_{\text{mix}}(\mathbf{w})}\right] \tag{64}$$

$$= \mathbb{E}_{\mathbf{w}\sim q_0}\left[\ln \frac{Z_0\, p(\mathbf{w}) g_0(\mathbf{w})}{Z_{\text{mix}} p(\mathbf{w}) g_{\text{mix}}(\mathbf{w})}\right] = \text{KL}\left[q_0 \,||\, q_{\text{mix}}\right] . \tag{65}$$

□

D Translation invariance in linear models

In this section, we derive the results from Sec. 4 for a Bayesian linear regression with a single input vector $\mathbf{x}$ and corresponding target observation y. The result stated in Sec. 4 for $\mathbf{x} = \mathbf{1}$ follows as a special case.

We assume we have K latent variables $\mathbf{w} = [w_1, \ldots, w_K]^{\mathrm{T}}$ and one observation $\mathbf{x} = [x_1, \ldots, x_K]^{\mathrm{T}}$. We further assume that the likelihood $p\left(y, |\mathbf{w}, \mathbf{x}\right)$ only depend on their inner product, that is, $\mathbf{x}^{\mathrm{T}}\mathbf{w}$. Then we know that any change in $\mathbf{w}$, which leaves the sum of the elements weighted by $\mathbf{x}$ unaffected, does not change the likelihood. For example $\mathbf{w}' = [w_1 + \Delta, w_2 - \Delta \cdot \frac{x_1}{x_2}, w_3, \ldots, w_K]^{\mathrm{T}}$ has the exact same likelihood than $\mathbf{w} = [w_1, w_2, w_3, \ldots, w_K]^{\mathrm{T}}$. In general, we can model this translation invariance using an $K-1$ dimensional vector $\boldsymbol{\Delta} \in \mathbb{R}^{K-1}$ and noting that

$$\mathbf{x}^{\mathrm{T}}\mathbf{w} = \mathbf{x}^{\mathrm{T}}\left(\mathbf{w} + \underbrace{\begin{bmatrix} \mathbf{I} \\ -x_K^{-1}\mathbf{x}_{K-1}^{\mathrm{T}} \end{bmatrix}}_{\mathbf{B}} \boldsymbol{\Delta}\right),$$

where we used $\mathbf{x}_{K-1} := [x_1, \ldots, x_{K-1}]^{\mathrm{T}}$ because

$$\mathbf{x}^{\mathrm{T}}\mathbf{B}\boldsymbol{\Delta} = \begin{bmatrix} \mathbf{x}_{K-1}^{\mathrm{T}} & x_K \end{bmatrix} \begin{bmatrix} \boldsymbol{\Delta} \\ -x_K^{-1}\mathbf{x}_{K-1}^{\mathrm{T}}\boldsymbol{\Delta} \end{bmatrix} = 0\,.$$

D.1 Likelihood Model

Now let us assume that we approximate the likelihood by a function that has Gaussian shape, that is $p\left(y, |\mathbf{w}, \mathbf{x}\right) \approx q\left(\mathbf{w}\right) \propto \mathcal{N}\left(\mathbf{w}; \mathbf{m}, \mathbf{V}\right)$. In order to model that the likelihood is translation invariant, we compute the marginal $q\left(\mathbf{w} - \mathbf{B}\boldsymbol{\Delta}\right)$ over $p\left(\boldsymbol{\Delta}\right) = \mathcal{N}\left(\boldsymbol{\Delta}; \mathbf{0}, \beta^2\mathbf{I}\right)$ and considering the case of $\beta \to \infty$, that is

$$\begin{aligned} q_\beta\left(\mathbf{w}\right) &:= \int q\left(\mathbf{w} - \mathbf{B}\boldsymbol{\Delta}\right) \cdot p\left(\boldsymbol{\Delta}\right)\, d\boldsymbol{\Delta} \\ &= \int \mathcal{N}\left(\mathbf{w}; \mathbf{m} + \mathbf{B}\boldsymbol{\Delta}, \mathbf{V}\right) \cdot \mathcal{N}\left(\boldsymbol{\Delta}; \mathbf{0}, \beta^2\mathbf{I}\right)\, d\boldsymbol{\Delta}\,. \end{aligned}$$

According to Theorem 1, for any $\beta \in \mathbb{R}^+$ this is another Gaussian given by

$$\begin{aligned} q_\beta\left(\mathbf{w}\right) &= \mathcal{N}\left(\mathbf{w}; \mathbf{m} + \mathbf{B}\mathbf{0}, \underbrace{\mathbf{V} + \beta\mathbf{B} \cdot \beta\mathbf{B}^{\mathrm{T}}}_{\mathbf{V}_\beta}\right) \\ &= \mathcal{N}\left(\mathbf{w}; \mathbf{m}, \mathbf{V} + \beta^2 \cdot \begin{bmatrix} \mathbf{I} & -x_K^{-1}\mathbf{x}_{K-1} \\ -x_K^{-1}\mathbf{x}_{K-1}^{\mathrm{T}} & x_K^{-2}\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{x}_{K-1} \end{bmatrix}\right). \end{aligned}$$

Note that $\lim_{\beta\to\infty} q_\beta\left(\mathbf{w}\right) = g_{\mathrm{mix}}(\mathbf{w})$ as defined in Subsection 4.1. Using the Woodbury formula in Theorem 2, the inverse of the covariance can be re-written as

$$\begin{aligned} \mathbf{V}_\beta^{-1} &:= \left(\mathbf{V} + \beta\mathbf{B} \cdot \beta\mathbf{B}^{\mathrm{T}}\right)^{-1} \\ &= \mathbf{V}^{-1} - \beta^2 \cdot \mathbf{V}^{-1}\mathbf{B}\left(\mathbf{I} + \beta^2\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1} \\ &= \mathbf{V}^{-1} - \beta^2 \cdot \mathbf{V}^{-1}\mathbf{B}\left(\beta^2\left(\beta^{-2}\mathbf{I} + \mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{B}\right)\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1} \\ &= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{B}\left(\beta^{-2}\mathbf{I} + \mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}\,. \end{aligned}$$

D.2 Posterior Model

If we assume a prior $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the posterior for the Gaussian approximation is given by

$$\begin{aligned} p(\mathbf{w}|\mathbf{x}) &\propto q(\mathbf{w}) \cdot p(\mathbf{w}) \\ &\propto \mathcal{G}\left(\mathbf{w}; \mathbf{V}^{-1}\mathbf{m}, \mathbf{V}^{-1}\right) \cdot \mathcal{G}\left(\mathbf{w}; \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\right) \\ &= \mathcal{G}\left(\mathbf{w}; \mathbf{V}^{-1}\mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \mathbf{V}^{-1} + \boldsymbol{\Sigma}^{-1}\right) \\ &= \mathcal{N}\left(\mathbf{w}; \boldsymbol{\Sigma}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{m} + \mathbf{V}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}, \mathbf{V}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}\right), \end{aligned}$$

where we used the identity $\left(\boldsymbol{\Sigma}^{-1} + \mathbf{V}^{-1}\right)^{-1} = \mathbf{V}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\left(\mathbf{V} + \boldsymbol{\Sigma}\right)^{-1}\mathbf{V}$ in the last line. Similarly, if we use the likelihood approximation which has incorporated the translation invariance, we get

$$\begin{aligned} p_\beta(\mathbf{w}|\mathbf{x}) &\propto q_\beta(\mathbf{w}) \cdot p(\mathbf{w}) \\ &\propto \mathcal{G}\left(\mathbf{w}; \mathbf{V}_\beta^{-1}\mathbf{m}, \mathbf{V}_\beta^{-1}\right) \cdot \mathcal{G}\left(\mathbf{w}; \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \boldsymbol{\Sigma}^{-1}\right) \\ &= \mathcal{G}\left(\mathbf{w}; \mathbf{V}_\beta^{-1}\mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \mathbf{V}_\beta^{-1} + \boldsymbol{\Sigma}^{-1}\right) \\ &= \mathcal{N}\left(\mathbf{w}; \left(\mathbf{V}_\beta^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\mathbf{V}_\beta^{-1}\mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \left(\mathbf{V}_\beta^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right) \end{aligned}$$

Observing that $\lim_{\beta\to\infty} \beta^{-2}\mathbf{I} = \mathbf{0}$ we thus see that

$$p_\infty(\mathbf{w}|\mathbf{x}) = \mathcal{N}\left(\mathbf{w}; \left(\mathbf{V}_\mathbf{B}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\mathbf{V}_\mathbf{B}^{-1}\mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \left(\mathbf{V}_\mathbf{B}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right), \tag{66}$$

$$\mathbf{V}_\mathbf{B}^{-1} := \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{B}\left(\mathbf{B}^\mathsf{T}\mathbf{V}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^\mathsf{T}\mathbf{V}^{-1}. \tag{67}$$

Note again that $p_\infty(\mathbf{w}) = q_{\text{mix}}(\mathbf{w}; \boldsymbol{\theta})$ as defined in (15).

D.2.1 Special Case of Diagonal Likelihood Covariance

Now let us consider the special case where the covariance matrix of the Gaussian likelihood approximation is a diagonal matrix, that is $\mathbf{V} = \text{Diag}(\boldsymbol{\lambda})$. Then, observe that

$$\mathbf{V}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{V}_{K-1}^{-1} & \mathbf{0} \\ \mathbf{0} & \lambda_K^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{I} \\ -x_K^{-1}\mathbf{x}_{K-1}^\mathsf{T} \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{K-1}^{-1} \\ -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}^\mathsf{T} \end{bmatrix}, \tag{68}$$

$$\mathbf{V}^{-1}\mathbf{B}\mathbf{V}_{K-1}\mathbf{x}_{K-1} = \left(\mathbf{V}^{-1}\mathbf{B}\right)\cdot\mathbf{V}_{K-1}\mathbf{x}_{K-1} = \begin{bmatrix} \mathbf{x}_{K-1} \\ -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}^\mathsf{T}\mathbf{V}_{K-1}\mathbf{x}_{K-1} \end{bmatrix}, \tag{69}$$

$$\mathbf{B}^\mathsf{T}\mathbf{V}^{-1}\mathbf{B} = \begin{bmatrix} \mathbf{I} & -x_K^{-1}\mathbf{x}_{K-1} \end{bmatrix}\begin{bmatrix} \mathbf{V}_{K-1}^{-1} \\ -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}^\mathsf{T} \end{bmatrix} = \mathbf{V}_{K-1}^{-1} + \lambda_K^{-1}x_K^{-2}\mathbf{x}_{K-1}\mathbf{x}_{K-1}^\mathsf{T},$$

where we used the notation $\mathbf{V}_{K-1}^{-1}$ to denote the diagonal $(K-1)\times(K-1)$ matrix with all

$$\boldsymbol{\lambda}_{K-1} := \left[\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_{K-1}^{-1}\right]^\mathsf{T}$$

on the diagonal. Thus, using Theorem 2 with $\mathbf{C} = \mathbf{V}_{K-1}^{-1}$, $\mathbf{A} = x_K^{-2}\lambda_K^{-1}\mathbf{x}_{K-1}$ and $\mathbf{B} = \mathbf{x}_{K-1}^\mathsf{T}$, we see that

$$\begin{aligned} \left(\mathbf{B}^\mathsf{T}\mathbf{V}^{-1}\mathbf{B}\right)^{-1} &= \mathbf{V}_{K-1} - \frac{1}{\lambda_K x_K^2 + \mathbf{x}_{K-1}^\mathsf{T}\mathbf{V}_{K-1}\mathbf{x}_{K-1}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\mathbf{x}_{K-1}^\mathsf{T}\mathbf{V}_{K-1} \\ &= \mathbf{V}_{K-1} - \frac{1}{\mathbf{x}^\mathsf{T}\mathbf{V}\mathbf{x}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\mathbf{x}_{K-1}^\mathsf{T}\mathbf{V}_{K-1}, \end{aligned}$$

Covariance $\left(\mathbf{V}_\mathbf{B}^{-1} + \boldsymbol{\Sigma}^{-1}\right)$. Thus, for the covariance (67) of the posterior we have

$$
\begin{aligned}
&\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1} \\
&\quad=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{B}\left[\mathbf{V}_{K-1}-\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\right]\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1} \\
&\quad=\mathbf{V}^{-1}-\underbrace{\mathbf{V}^{-1}\mathbf{B}\mathbf{V}_{K-1}\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}}_{S}+\underbrace{\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\mathbf{V}^{-1}\mathbf{B}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{B}^{\mathrm{T}}\mathbf{V}^{-1}}_{T}\,.
\end{aligned}
$$

Let's focus on the expression S first. Using (68) we have

$$
\begin{aligned}
S&=\begin{bmatrix}\mathbf{V}_{K-1}^{-1}\\ -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}^{\mathrm{T}}\end{bmatrix}\mathbf{V}_{K-1}\begin{bmatrix}\mathbf{V}_{K-1}^{-1} & -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}\end{bmatrix}\\
&=\begin{bmatrix}\mathbf{V}_{K-1}^{-1} & -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}\\ -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}^{\mathrm{T}} & \lambda_K^{-2}x_K^{-2}\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\end{bmatrix}. \qquad (70)
\end{aligned}
$$

Similarly, for the expression Tusing (69) we have

$$
\begin{aligned}
T&=\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\begin{bmatrix}\mathbf{x}_{K-1}\\ -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\end{bmatrix}\begin{bmatrix}\mathbf{x}_{K-1} & -\lambda_K^{-1}x_K^{-1}\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\end{bmatrix}\\
&=\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\begin{bmatrix}\mathbf{x}_{K-1}\mathbf{x}_{K-1}^{\mathrm{T}} & -\frac{\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}}{\lambda_K x_K}\mathbf{x}_{K-1}\\ -\frac{\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}}{\lambda_K x_K}\mathbf{x}_{K-1}^{\mathrm{T}} & \frac{\left(\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\right)^2}{\lambda_K^2 x_K^2}\end{bmatrix}. \qquad (71)
\end{aligned}
$$

Putting (70) and (71) together, we get

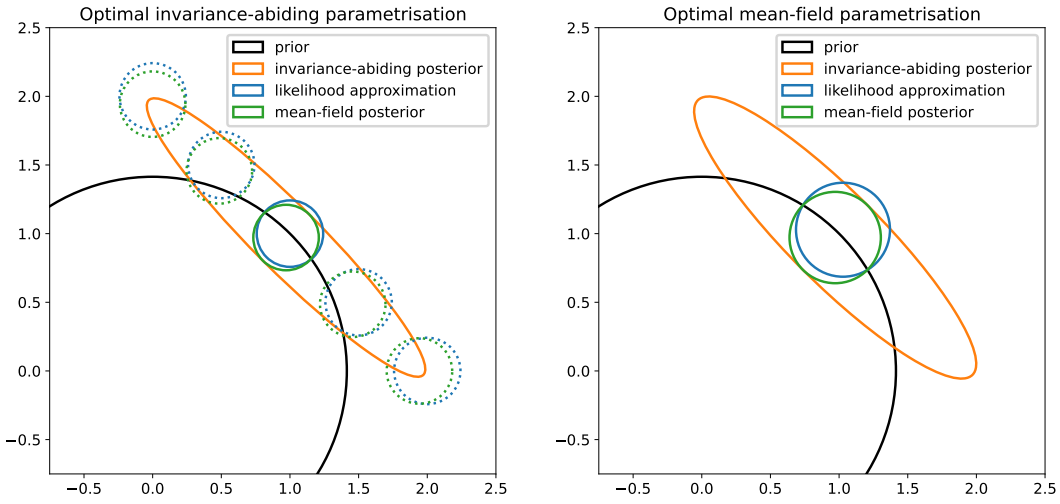
$$
\begin{aligned}
\mathbf{V}_{\mathbf{B}}^{-1}&=\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\begin{bmatrix}\mathbf{x}_{K-1}\mathbf{x}_{K-1}^{\mathrm{T}} & \left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}{\lambda_K x_K}-\frac{\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}}{\lambda_K x_K}\right)\mathbf{x}_{K-1}\\ \left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}{\lambda_K x_K}-\frac{\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}}{\lambda_K x_K}\right)\mathbf{x}_{K-1}^{\mathrm{T}} & \frac{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}{\lambda_K}-\frac{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}\cdot\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}}{\lambda_K^2 x_K^2}+\frac{\left(\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}\right)^2}{\lambda_K^2 x_K^2}\end{bmatrix}\\
&=\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\begin{bmatrix}\mathbf{x}_{K-1}\mathbf{x}_{K-1}^{\mathrm{T}} & x_K\mathbf{x}_{K-1}\\ x_K\mathbf{x}_{K-1}^{\mathrm{T}} & x_K^2\end{bmatrix}\\
&=\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\mathbf{x}\mathbf{x}^{\mathrm{T}}, \qquad (72)
\end{aligned}
$$

where we repeatedly used that $\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}-\mathbf{x}_{K-1}^{\mathrm{T}}\mathbf{V}_{K-1}\mathbf{x}_{K-1}=\lambda_K x_K^2$. Thus, the covariance in (66) can be written as

$$
\begin{aligned}
\left(\mathbf{V}_{\mathbf{B}}^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}&=\left(\boldsymbol{\Sigma}^{-1}+\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)^{-1}\\
&=\boldsymbol{\Sigma}-\frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\cdot\frac{1}{1+\left(\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}\right)^{-1}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{x}}\cdot\boldsymbol{\Sigma}\mathbf{x}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}\\
&=\boldsymbol{\Sigma}-\frac{1}{\mathbf{x}^{\mathrm{T}}\left(\mathbf{V}+\boldsymbol{\Sigma}\right)\mathbf{x}}\cdot\left(\boldsymbol{\Sigma}\mathbf{x}\right)\left(\boldsymbol{\Sigma}\mathbf{x}\right)^{\mathrm{T}}, \qquad (73)
\end{aligned}
$$

where we used Theorem 2 in the third step. Note that (14) is a special case of (72) when using $\mathbf{x}=\mathbf{1}$ and observing that $\mathbf{V}\mathbf{1}=\boldsymbol{\lambda}$. Similarly, the covariance in (15) is a special case of (73) when further noticing that $\boldsymbol{\Sigma}\mathbf{1}=\boldsymbol{\sigma}^2$.

Mean $\left(\mathbf{V}_{\mathbf{B}}^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\mathbf{V}_{\mathbf{B}}^{-1}\mathbf{m}+\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$. In order to derive an efficient update for the mean of the posterior, please note that by virtue of (72), $\mathbf{V}_{\mathbf{B}}^{-1}$ can be written as $\mathbf{d}\mathbf{d}^{\mathrm{T}}$with $\mathbf{d}=\left(\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}\right)^{-\frac{1}{2}}\cdot\mathbf{x}$.



(a) Optimal parameters $\boldsymbol{\theta}^*_{\text{mix}}$, cf. (17a)

(b) Optimal parameters $\boldsymbol{\theta}^*_0$, cf. (17b)

Figure 3: Gaussian likelihood and posterior approximations q_0 and q_{mix} (see (15)) with different parameter optima (cf. (17)). The dotted circles show alternative parameter values that induce the same predictive distribution but do not correspond to one of the optima in (17).

Thus, using (73) we have

$$
\begin{aligned}
&\left(\mathbf{V}_\mathbf{B}^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\mathbf{V}_\mathbf{B}^{-1}\mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= \left(\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)\left(\mathbf{d}\mathbf{d}^\mathrm{T}\mathbf{m} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= \boldsymbol{\Sigma}\mathbf{d}\mathbf{d}^\mathrm{T}\mathbf{m} + \boldsymbol{\mu} - \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}\mathbf{d}\mathbf{d}^\mathrm{T}\mathbf{m} - \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}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
&= \boldsymbol{\mu} + \left(\frac{\mathbf{x}^\mathrm{T}\mathbf{m}}{\mathbf{x}^\mathrm{T}\mathbf{V}\mathbf{x}}\right) \cdot \left(\boldsymbol{\Sigma}\mathbf{x}\right) - \frac{\left(\boldsymbol{\Sigma}\mathbf{x}\right)^\mathrm{T}\mathbf{d}\mathbf{d}^\mathrm{T}\mathbf{m}}{\mathbf{x}^\mathrm{T}\left(\mathbf{V}+\boldsymbol{\Sigma}\right)\mathbf{x}} \cdot \left(\boldsymbol{\Sigma}\mathbf{x}\right) - \frac{\left(\boldsymbol{\Sigma}\mathbf{x}\right)^\mathrm{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\mathbf{x}^\mathrm{T}\left(\mathbf{V}+\boldsymbol{\Sigma}\right)\mathbf{x}} \cdot \left(\boldsymbol{\Sigma}\mathbf{x}\right) \\
&= \boldsymbol{\mu} + \left[\frac{\mathbf{x}^\mathrm{T}\mathbf{m}}{\mathbf{x}^\mathrm{T}\mathbf{V}\mathbf{x}} - \frac{\mathbf{x}^\mathrm{T}\mathbf{m}}{\mathbf{x}^\mathrm{T}\mathbf{V}\mathbf{x}} \cdot \frac{\mathbf{x}^\mathrm{T}\boldsymbol{\Sigma}\mathbf{x}}{\mathbf{x}^\mathrm{T}\left(\mathbf{V}+\boldsymbol{\Sigma}\right)\mathbf{x}} - \frac{\mathbf{x}^\mathrm{T}\boldsymbol{\mu}}{\mathbf{x}^\mathrm{T}\left(\mathbf{V}+\boldsymbol{\Sigma}\right)\mathbf{x}}\right] \cdot \left(\boldsymbol{\Sigma}\mathbf{x}\right) \\
&= \boldsymbol{\mu} + \left(\frac{\mathbf{x}^\mathrm{T}\left(\mathbf{m}-\boldsymbol{\mu}\right)}{\mathbf{x}^\mathrm{T}\left(\mathbf{V}+\boldsymbol{\Sigma}\right)\mathbf{x}}\right) \cdot \left(\boldsymbol{\Sigma}\mathbf{x}\right) .
\end{aligned}
$$

Thus, the location parameter in (15) is a special case of this more general result when using $\mathbf{x} = \mathbf{1}$ and noticing again that $\boldsymbol{\Sigma}\mathbf{1} = \boldsymbol{\sigma}^2$ and $\mathbf{V}\mathbf{1} = \boldsymbol{\lambda}$, respectively. It can be seen that the Gaussian product updates the prior location in the direction $\boldsymbol{\Sigma}\mathbf{x}$. This is because the likelihood is translation invariant wrt. all directions perpendicular to $\mathbf{x}$ (i.e. the hyper plane determined by the normal vector $\mathbf{x}$).

The two posterior approximations q_0 and q_{mix} as well as the corresponding likelihood approximation is visualised in Fig. 3 for two different parametrisations (cf. Sec. 4.3).

D.3 True posterior and optimal invariance-abiding parameters

For the linear model with a single observation y and inputs $\mathbf{x}$, the true posterior $p(\mathbf{w} \mid \mathbf{x}, y)$ follows from the standard Bayesian update equation for Gaussian linear models (cf. (27) in App. B with $\mathbf{A} = \frac{1}{K} \cdot \mathbf{x}^\mathrm{T}$):

$$
p(\mathbf{w} \mid y, \mathbf{x}) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_p^*, \mathbf{V}_p^*\right), \quad \mathbf{V}_p^* = \left(\frac{1}{K^2\sigma_y^2}\mathbf{x}\mathbf{x}^\mathrm{T} + \boldsymbol{\Sigma}^{-1}\right)^{-1}, \quad \mathbf{m}_p^* = \mathbf{V}_p^*\frac{y}{K\sigma_y^2}\mathbf{x} . \tag{74}
$$

For N observations $Y := \{y^{(n)}\}_{n=1}^N$ with identical input $\mathbf{x}$, the posterior is

$$
p\left(\mathbf{w} \mid Y, \mathbf{x}\right) = \mathcal{N}\left(\mathbf{w}; \mathbf{m}_p^*, \mathbf{V}_p^*\right), \mathbf{V}_p^* = \left(\frac{N}{K^2\sigma_y^2}\mathbf{x}\mathbf{x}^\mathrm{T} + \boldsymbol{\Sigma}^{-1}\right)^{-1}, \mathbf{m}_p^* = \mathbf{V}_p^*\frac{\sum_{n=1}^N y^{(n)}}{K\sigma_y^2}\mathbf{x} . \tag{75}
$$

We want to relate the true posterior (75) to the parameters of the invariance-abiding posterior $q_{\text{mix}}(\mathbf{w};\boldsymbol{\theta})$. It suffices to consider the special case $\mathbf{x}=\mathbf{1}$ from the main text. We will use the form

$$q_{\text{mix}}(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}; \left(\boldsymbol{\Sigma}^{-1}+\mathbf{V}_{\text{mix}}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\mathbf{V}_{\text{mix}}^{-1}\mathbf{m}_{\text{mix}}\right), \boldsymbol{\Sigma}^{-1}+\mathbf{V}_{\text{mix}}^{-1}\right) \tag{76}$$

For the precision matrix $\boldsymbol{\Sigma}^{-1}+\mathbf{V}_{\text{mix}}^{-1}$, using the form in (72) with $\mathbf{x}=\mathbf{1}$ gives

$$\mathbf{V}_{\text{mix}}^{-1} = \frac{1}{\mathbf{x}^{\mathrm{T}}\mathbf{V}\mathbf{x}}\mathbf{x}\mathbf{x}^{\mathrm{T}} = \frac{1}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda}}\mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{77}$$

since $\mathbf{V}\mathbf{1}=\boldsymbol{\lambda}$. Comparing this to (75), we see that $\boldsymbol{\Sigma}^{-1}+\mathbf{V}_{\text{mix}}^{-1}$ has the same structure as the precision matrix of the true posterior. By setting the optimal variances as $\mathbf{1}^{\mathrm{T}}\boldsymbol{\lambda}^{*}=\frac{K^2\sigma_y^2}{N}$, and using $\frac{K^2\sigma_y^2}{N}=\frac{K\sigma_y^2}{N}\cdot\mathbf{1}^{\mathrm{T}}\mathbf{1}$, we see that one possible choice for the optimal variance parameters is

$$\boldsymbol{\lambda}^{*} = \frac{K\sigma_y^2}{N}\cdot\mathbf{1}. \tag{78}$$

We note that other vectors that are not proportional to $\mathbf{1}$ and also sum to $\frac{K^2\sigma_y^2}{N}$ are also valid optima.

Similarly, with $\boldsymbol{\mu}=\mathbf{0}$, and by noting that $\boldsymbol{\Sigma}^{-1}+\mathbf{V}_{\text{mix}}^{-1}=\mathbf{V}_p^{-1}$, we see that

$$\mathbf{m}^{*} = \frac{1}{N}\sum_{n=1}^{N} y^{(n)}\cdot\mathbf{1}. \tag{79}$$

E Permutation invariance in Bayesian neural networks

Here we analyse the *permutation* invariance in BNNs. This invariance is independent of the data, persisting for any dataset size.

We first describe the set of permutation matrices corresponding to the transformations that leave the likelihood invariant, i.e. $t(\mathbf{w},\mathbf{r})=\mathbf{P}_{\mathbf{r}}\mathbf{w}$. We ignore the biases for simplicity and describe these permutations first on a node/neuron level, then layer-wise for the weight matrices, and finally on the weight vector that is obtained by stacking all weight matrices. We will denote layers with indices l and the number of layers by L. Layer-wise permutation matrices are then denoted as $\tilde{\mathbf{P}}_l$ and the corresponding weight matrices are denoted as $\mathbf{W}_l$. The permutation matrix that results when stacking all weights matrices into a vector $\mathbf{w}$ is denoted as $\mathbf{P}$. When necessary, one particular permutation matrix from the set of all possible permutation matrices for a given architecture is indexed with superscript $^{(i)}$, i.e. $\mathcal{P}=\{\mathbf{P}^{(i)}\}_i$ and $|\mathcal{P}|=\prod_{l=1}^{L-1}k_l!$, where k_l is the number of nodes in layer l.

Node permutations. Each hidden layer $\mathbf{z}_l\in\mathbf{z}_1,\ldots,\mathbf{z}_{L-1}$ is a set of nodes that can be permuted in $k_l!$ possible ways, relabelling the corresponding parameters attached to these nodes. Each of these $k_l!$ permutations per layer can be combined with any of the permutations of another layer. Each permutation matrix $\tilde{\mathbf{P}}_l$ for a layer l corresponds to one of the unique orderings of nodes $\mathbf{z}_l$. For instance, the permutation matrix that reorders the first 3 nodes as $(z_{l,3},z_{l,1},z_{l,2})$ is

$$\tilde{\mathbf{P}}_l = \begin{pmatrix} 0 & 0 & 1 & \ldots \\ 1 & 0 & 0 & \ldots \\ 0 & 1 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}. \tag{80}$$

The remaining entries in the matrix are ones on the diagonal and zeros elsewhere.

Layer-wise permutation of weight matrices. Next, we describe how node permutations correspond to permutations of weight *matrices* in terms of permutations to the in- and outgoing weights. For one particular instance of permutations (omitting superscript $^{(i)}$) to each of the hidden layers, the corresponding weight matrices can be permuted as follows:

$$\forall l\in 1,\ldots,L: \quad \mathbf{w}_l' = \tilde{\mathbf{P}}_l\mathbf{W}_l\tilde{\mathbf{P}}_{l-1}^{\mathrm{T}}, \quad \tilde{\mathbf{P}}_0=\tilde{\mathbf{P}}_L=\mathbf{I}. \tag{81}$$

The identities corresponding to the first and last layers is because only hidden layer nodes can be permuted but not the data itself. Note also that each permutation matrix is applied to two weight matrices, since permutations to nodes of a particular layer correspond to the simultaneous permutation of the weight matrices from the preceding and subsequent layer.

Permutation of stacked weight vectors. The layer-wise formulation can be written in terms of the stacked weight vector $\mathbf{w} = [\mathrm{vec}(\mathbf{W}_1), \ldots, \mathrm{vec}(\mathbf{W}_{L+1})]^{\mathrm{T}}$, using $\mathrm{vec}(ABC) = \left(C^{\mathrm{T}} \otimes A\right) \mathrm{vec}(B)$:

$$\mathrm{vec}(\mathbf{W}_l') = \left(\tilde{\mathbf{P}}_{l-1} \otimes \tilde{\mathbf{P}}_l\right) \mathrm{vec}(\mathbf{W}_l) =: \mathbf{P}_{l,l-1} \mathrm{vec}(\mathbf{W}_l), \tag{82}$$

where we denote the new permutation matrix that is applied to the vectorised weights with a bar and the subscripts correspond to the two successive layers. The overall permutation matrix corresponding to the entire weight vector $\mathbf{w}$ is given by forming the block-diagonal matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{1,0} & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{P}_{2,1} & \mathbf{0} & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_{3,2} & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{pmatrix}, \tag{83}$$

where each $\mathbf{0}$ denotes block-matrices of zeros with the respective dimensions and the remaining parts of the matrix are the identity on the diagonal and zeros elsewhere.

Invariance gap. Note again that the permutation invariance for BNNs is independent of the data. The invariance gap $\mathrm{KL}\left[q_0 \,||\, q_{\mathrm{mix}}\right]$ takes values in the range $\left[\, 0,\, \ln |\mathcal{P}| \,\right]$, where $|\mathcal{P}|$ is the factorial number of modes. The gap takes the maximal value when each mode is completely separated from the other modes, and the gap is zero when both $q_0(\mathbf{w})$ and $q_{\mathrm{mix}}(\mathbf{w})$ are identical. This is the case e.g. if $q_0(\mathbf{w})$ reverts to the prior $p(\mathbf{w})$ since all modes are then identical, i.e. $q_0(\mathbf{P}\mathbf{w}) = q(\mathbf{w})$.

F Data-related bound on the mean-field KL divergence

Assume the regression setting described in Sec. 2.2, where we assume a finite dataset $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$ and a regression setting with fixed homogeneous noise variance σ_y^2. We can further assume that the prior is chosen such that it induces a reasonably bounded output variance of the neural network, $\sigma_L^2\left(\mathbf{x}^{(n)}\right) := \mathbb{V}_{\mathbf{w}\sim p}\left[f(\mathbf{x}^{(n)};\, \mathbf{w})\right]$, for some fixed input $\mathbf{x}^{(n)}$. We will then have a finite expected log-likelihood and the ELBO for a mean-field variational approximation $q_{\boldsymbol{\theta}}$ is

$$\mathcal{L}_{\mathrm{ELBO}}\left(q_{\boldsymbol{\theta}}, \mathcal{D}\right) = \underbrace{\sum_{n=1}^{N} \mathbb{E}_{\mathbf{w}\sim q_{\boldsymbol{\theta}}}\left[\ln p(y^{(n)} \,|\, \mathbf{x}^{(n)}, w)\right]}_{\mathrm{ELL}(q_{\boldsymbol{\theta}}, \mathcal{D})} - \mathrm{KL}\left[q_{\boldsymbol{\theta}} \,||\, p\right]. \tag{84}$$

Let us now consider a hypothetical *worst-case* fit in terms of the ELBO objective. This is when the data is completely ignored with $q_{\boldsymbol{\theta}}(\mathbf{w}) = p(\mathbf{w})$, as any worse fit could be trivially improved by setting the approximation to the prior. This gives then the inequality (using (84))

$$\forall q_{\boldsymbol{\theta}}: \quad \mathcal{L}_{\mathrm{ELBO}}\left(q_{\boldsymbol{\theta}}, \mathcal{D}\right) \geq \mathcal{L}_{\mathrm{ELBO}}\left(p, \mathcal{D}\right) \;\Leftrightarrow\; \mathrm{ELL}\left(q_{\boldsymbol{\theta}}, \mathcal{D}\right) - \mathrm{KL}\left[q_{\boldsymbol{\theta}} \,||\, p\right] \geq \mathrm{ELL}\left(p, \mathcal{D}\right). \tag{85}$$

Although we can not compute the expected log-likelihood for an arbitrary fit $q_{\boldsymbol{\theta}}$, we can quantify the upper bound given above, by considering the *best-case* fit q^*, i.e. a hypothetical optimum where the data is perfectly predicted up to the known observation noise variance σ_y^2. The resulting bound is then

$$\mathrm{KL}\left[q_{\boldsymbol{\theta}} \,||\, p\right] \leq \mathrm{ELL}\left(q^*, \mathcal{D}\right) - \mathrm{ELL}\left(p, \mathcal{D}\right). \tag{86}$$

Next, we compute the two expected log-likelihood terms in (86). We first consider the worst-case, where we have

$$\mathrm{ELL}(p, \mathcal{D}) = \sum_{n=1}^{N} \mathbb{E}_{\mathbf{w}\sim p}\left[\ln p(y^{(n)} | \mathbf{x}^{(n)}, \mathbf{w})\right] \tag{87}$$

$$= \sum_{n=1}^{N} -\frac{1}{2\sigma_y^2} \mathbb{E}_{\mathbf{w}\sim p}\left[\left(y^{(n)} - z_L\right)^2\right] - \frac{N}{2} \ln\left[2\pi\sigma_y^2\right], \tag{88}$$

where $\mathbf{z}_L$ is the noisy output of the neural network given by

$$z_L = f(\mathbf{x}^{(n)};\, \mathbf{w}) + \epsilon, \quad \mathbf{w} \sim p(\mathbf{w}), \quad \epsilon \sim \mathcal{N}\left(0,\, \sigma_y^2\right). \tag{89}$$

Splitting the expectation of the quadratic term into variance and square of the expectation, we have

$$\mathbb{E}_{\mathbf{w}\sim p}\left[\left(y^{(n)}-z_L\right)^2\right]=\mathbb{E}_{\mathbf{w}\sim p}\left[y^{(n)}-z_L\right]^2+\mathbb{V}_{\mathbf{w}\sim p}\left[y^{(n)}-z_L\right] \tag{90}$$

Since the last layer computes $f(\mathbf{x}^{(n)};\,\mathbf{w})=\mathbf{w}_{L,1}^{\mathrm{T}}\mathbf{z}_{L-1}$ in the regression setting and the prior weights are independent of $\mathbf{z}_{L-1}$ and centred at zero, i.e. $\mathbb{E}_{\mathbf{w}\sim p}\left[\mathbf{w}_{L,1}\right]=\mathbf{0}$, it follows that

$$\mathbb{E}_{\mathbf{w}\sim p}\left[y^{(n)}-z_L\right]=y^{(n)}. \tag{91}$$

$$\mathbb{V}_{\mathbf{w}\sim p}\left[y^{(n)}-z_L\right]=\mathbb{E}_{\mathbf{w}\sim p}\left[z_L^2\right]=\sigma_L^2\left(\mathbf{x}^{(n)}\right)+\sigma_y^2, \tag{92}$$

where we denote the variance of the neural network outputs by $\sigma_L^2\left(\mathbf{x}^{(n)}\right):=\mathbb{V}_{\mathbf{w}\sim p}\left[f(\mathbf{x}^{(n)};\,\mathbf{w})\right]$. Hence, the expected log likelihood under the prior is

$$\mathrm{ELL}(p,\mathcal{D})=\sum_{n=1}^{N}-\frac{1}{2\sigma_y^2}\left[\sigma_L^2\left(\mathbf{x}^{(n)}\right)+\sigma_y^2+\left(y^{(n)}\right)^2\right]-\frac{N}{2}\ln\left[2\pi\sigma_y^2\right] \tag{93}$$

$$=-\frac{1}{2}\sum_{n=1}^{N}\frac{\sigma_L^2\left(\mathbf{x}^{(n)}\right)+\left(y^{(n)}\right)^2}{\sigma_y^2}-\frac{N}{2}\left(1+\ln\left[2\pi\sigma_y^2\right]\right). \tag{94}$$

For the best-case fit in terms of the ELBO, we assume that we completely overfit and perfectly predict the data by putting the mean of the Gaussian exactly on the data, i.e.

$$p(y^{(n)}|\mathbf{x}^{(n)},\mathbf{w})=\mathcal{N}(y;y^{(n)},\sigma_y^2).$$

Then, we have

$$\mathbb{E}_{\mathbf{w}\sim q^*}\left[y^{(n)}-z_L\right]=0, \tag{95}$$

$$\mathbb{V}_{\mathbf{w}\sim q^*}\left[y^{(n)}-z_L\right]=\sigma_y^2. \tag{96}$$

Hence, the expected log likelihood of the best case fit is

$$\mathrm{ELL}(q^*,\mathcal{D})=\sum_{n=1}^{N}\mathbb{E}_{\mathbf{w}\sim q^*}\left[\ln p(y^{(n)}\,|\,\mathbf{x}^{(n)},\mathbf{w})\right] \tag{97}$$

$$=\sum_{n=1}^{N}-\frac{1}{2\sigma_y^2}\left[\sigma_y^2\right]-\frac{N}{2}\ln\left[2\pi\sigma_y^2\right] \tag{98}$$

$$=-\frac{N}{2}\left(1+\ln\left[2\pi\sigma_y^2\right]\right). \tag{99}$$

Taking the difference between (93) and (97), we obtain the result in (3),

$$\mathrm{KL}\left[q_{\boldsymbol{\theta}}\,||\,p\right]\leq\sum_{n=1}^{N}\frac{\sigma_L^2\left(\mathbf{x}^{(n)}\right)+\left(y^{(n)}\right)^2}{2\sigma_y^2}.$$