Variational Bayesian Bow tie Neural Networks with Shrinkage

Alisa Sheinkman[*] and Sara Wade[†]

November 20, 2024

Abstract

Despite the dominant role of deep models in machine learning, limitations persist, including overconfident predictions, susceptibility to adversarial attacks, and underestimation of variability in predictions. The Bayesian paradigm provides a natural framework to overcome such issues and has become the gold standard for uncertainty estimation with deep models, also providing improved accuracy and a framework for tuning critical hyperparameters. However, exact Bayesian inference is challenging, typically involving variational algorithms that impose strong independence and distributional assumptions. Moreover, existing methods are sensitive to the architectural choice of the network. We address these issues by constructing a relaxed version of the standard feed-forward rectified neural network, and employing Polya-Gamma data augmentation tricks to render a conditionally linear and Gaussian model. Additionally, we use sparsity-promoting priors on the weights of the neural network for data-driven architectural design. To approximate the posterior, we derive a variational inference algorithm that avoids distributional assumptions and independence across layers and is a faster alternative to the usual Markov Chain Monte Carlo schemes.

Keywords: Bayesian neural networks, variational inference, uncertainty quantification, shrinkage priors.

1   Introduction

Neural networks (NNs) are effective deep models that play a dominant role in machine learning and have achieved remarkable success across various domains including medicine and biological sciences Jumper et al. (2021); Yu et al. (2021), natural language processing Mikolov et al. (2013); Touvron et al. (2023), computer vision and image analysis Dosovitskiy et al. (2020), data privacy and security Yang et al. (2019) and beyond. However, modern machine learning applications often lack reliable, if not any, uncertainty estimates Guo et al. (2017); Gal (2016); Ashukha et al. (2020). Classical deep models are easily fooled and are susceptible to adversarial attacks Szegedy et al. (2014); Nguyen et al. (2015); Zong et al. (2024), and even when the adversarial attacks fail, the saliency interpretations of deep neural networks (DNNs) are rather brittle Carbone et al. (2022). When data variations leading to out-of-distribution (OOD) shifts occur neural networks often fail to generalize well Hein et al. (2019); Zhang et al. (2024); Ashukha et al. (2020). Moreover, standard neural networks usually lack intuitive interpretation and explainability and so are regarded as black boxes Lipton (2018). To address these challenges, Bayesian neural networks (BNNs) have emerged as a compelling extension of conventional neural networks (for a review, see e.g. Jospin et al. (2022); Arbel et al. (2023)). While finite (non-Bayesian) deep ensembles of independent neural networks have been shown to improve prediction and uncertainty estimates Lakshminarayanan et al. (2017), the Bayesian approach creates infinite ensembles of deep neural networks. The advantage of this approach is that it controls the model complexity and builds regularization into the model by marginalizing out the parameters. Indeed, Bayesian neural networks have become the gold standard for uncertainty estimation in the context of data-driven decision-making and in safety-critical applications, where robustness and calibration are crucial McAllister et al. (2017); Carbone et al. (2020); Gruver et al. (2023); Yang et al. (2024); Klarner et al. (2023).

A core problem of Bayesian machine learning lies in performing the inference; in practice, the posterior distribution of the model's parameters given observations is not available in the closed form and direct sampling from the posterior is computationally expensive, meaning one has to employ approximate Bayesian

[*]School of Mathematics, University of Edinburgh, Edinburgh, UK, a.sheinkman@sms.ed.ac.uk
[†]School of Mathematics, University of Edinburgh, Edinburgh, UK, sara.wade@ed.ac.uk

inference. Markov chain Monte Carlo (MCMC) is a gold standard solution since it produces draws, which are asymptotically exact samples from the posterior but for large data sets or complex models with multi-modal posteriors it can be prohibitively slow. Variational inference Hinton and van Camp (1993); Jordan et al. (1999) (VI) instead utilizes optimization rather than sampling making it a more computationally effective method suitable for high-dimensional problems. VI approximates the posterior with the closest (most commonly, in terms of the Kullback–Leibler divergence) member of some tractable variational family of distributions taken as close as possible to the true posterior Blei et al. (2017). Recently, several variational algorithms and methods have been proposed and proven to achieve desirable consistency and predictive performance Bai et al. (2020); Zhang et al. (2018); Castillo and Egels (2024); Yang et al. (2020); Chérief-Abdellatif (2020).

In this paper, we follow the setup of Smith et al. (2021) and introduce a bow tie neural network, where a stochastic relaxation of the rectified linear unit (ReLU) activation function leads to a model amendable to the Polya-Gamma (PG) data augmentation trick Polson et al. (2013) and results in conditionally linear and Gaussian stochastic activations. Additionally, on the weights of the network, we place sparsity-inducing priors, which are known for their ability to provide improvement in the predictive performance of Bayesian deep models; not only do sparse models ease the storage and computational burden, they also improve the calibration and may recover the potential sparse structure of the target function Polson and Ročková (2018); Bai et al. (2020); Griffin and Brown (2021); Law and Zankin (2022); Ray and Szabó (2022). Specifically, we consider sparsity-inducing global-local normal-generalized inverse Gaussians (N-GIG) priors Polson and Scott (2012). Section 2 describes the bow tie model with shrinkage priors and implementation of PG data augmentation. Having constructed the bow tie neural network, we propose a (block) structured mean-field family for the approximate variational posterior which is flexible enough and doesn't require assumptions on the distributional form of each component as well as on independence across layers. For the chosen family, coordinate ascent variation inference (CAVI) Bishop (2016) can be performed, with all variational updates available in the closed form. Whilst continuous shrinkage priors result in more tractable computations, they do not incur exact zeros on the neural network's weights. We address this issue by implementing a simple post-process node selection algorithm controlled by the empirical Bayesian false discovery rate (FDR). Finally, we derive the predictive distribution and propose improving the accuracy and uncertainty estimation by considering ensembles of variational approximations obtained by running several parallel variational algorithms with different random starting points. In this way, our approach accounts for the multimodality of the posterior distributions arising in Bayesian deep models. Section 3 derives the inference algorithm, variable selection procedure and predictive distribution. We evaluate our method on a range of classical regression tasks as well as synthetic regression tasks and demonstrate its competitiveness compared to alternative well-known Bayesian algorithms in Section 4 [1].

2 Bayesian Augmented Bow Tie Neural Network with Shrinkage

2.1 Bow tie neural networks

We begin by describing the class of recently proposed *bow tie networks* Smith et al. (2021), which are deep generative models that generalize feed-forward rectified linear neural networks with stochastic activations. Let $\mathbf{x}_n \in \mathbb{R}^{D_0}$ be the inputs, $\mathbf{y}_n \in \mathbb{R}^{D_{L+1}}$ be the outputs and $\mathbf{a}_n = \{\mathbf{a}_{n,l}\}_{l=1}^L$ with $\mathbf{a}_{n,l} \in \mathbb{R}^{D_l}$ be the latent activations at each of the L intermediate layers. For notational purposes, assume $\mathbf{a}_{n,0} = \mathbf{x}_n$. The model assumes:

$$\mathbf{y}_n \mid \mathbf{a}_n, \mathbf{x}_n, \boldsymbol{\theta} \sim \mathrm{N}\left(\mathbf{y}_n | \mathbf{z}_{n,L+1}, \boldsymbol{\Sigma}_{L+1}\right) \quad \text{for } n = 1, \dots, N,$$

where

$$\mathbf{a}_{n,l} | \mathbf{z}_{n,l}, \boldsymbol{\theta} \sim \mathrm{N}\left(f(\mathbf{z}_{n,l}), \boldsymbol{\Sigma}_l\right), \quad \text{with} \quad \mathbf{z}_{n,l} = \mathbf{W}_l \mathbf{a}_{n,l-1} + \mathbf{b}_l \quad \text{for} \quad l = 1, \dots, L+1. \tag{1}$$

Here $f(\mathbf{z})$ is a nonlinear activation function applied elementwise and the parameters $\boldsymbol{\theta} = (\mathbf{W}_l, \mathbf{b}_l, \boldsymbol{\Sigma}_l)_{l=1}^{L+1}$ consist of the weights $\mathbf{W}_l \in \mathbb{R}^{D_l \times D_{l-1}}$, biases $\mathbf{b}_l \in \mathbb{R}^{D_l}$ and covariance matrices $\boldsymbol{\Sigma}_l \in \mathbb{R}^{D_l \times D_l}$.

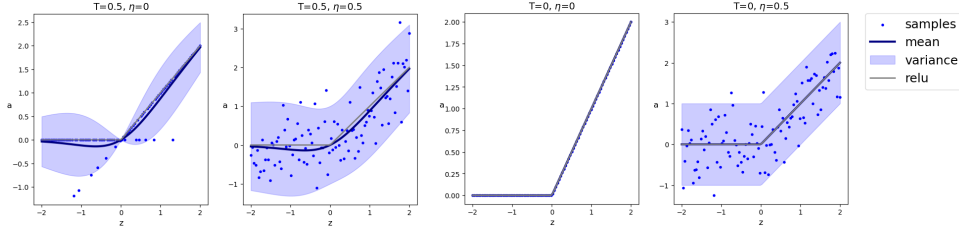[1]We provide a Python implementation of our model at `https://github.com/sheinkmana/V_bowtie_NN`.

Figure 1: Conditional distribution of a given the input z for various settings of the temperature T and noise η, with the conditional mean in Equation (3) (solid line), conditional variance in Equation (4) (shaded region) and samples from the conditional distribution in Equation (2) (points).

Note that Equation (1) is a stochastic relaxation of the standard feed-forward NN, which is recovered in the limiting case when $\mathbf{\Sigma}_l \to \mathbf{0}$ for $l = 1, \ldots, L$. Instead of relying on local gradient-based algorithms, Smith et al. (2021) introduces another relaxation of the model and employs a *Polya-Gamma data augmentation trick* Polson et al. (2013) to render the model conditionally linear with Gaussian activations. Specifically, consider the ReLU activation function $f(z) = \max(0, z)$. It can alternatively be written as a product of z and a binary function γ, i.e. $f(z) = \gamma z$ where $\gamma = \mathbf{1}(z > 0)$. In this way, γ determines whether the node is activated ($\gamma = 1$) or not ($\gamma = 0$). In a similar fashion, the additional stochastic relaxation replaces $f(\mathbf{z}_{n,l})$ with $\boldsymbol{\gamma}_{n,l} \odot \mathbf{z}_{n,l}$:

$$\mathbf{a}_{n,l} | \mathbf{z}_{n,l}, \boldsymbol{\gamma}_{n,l}, \boldsymbol{\theta} \sim \mathrm{N}\left(\boldsymbol{\gamma}_{n,l} \odot \mathbf{z}_{n,l}, \mathbf{\Sigma}_l\right),$$

$$\gamma_{n,l,d} \overset{ind}{\sim} \mathrm{Bern}\left(\sigma\left(z_{n,l,d}/\mathrm{T}\right)\right),$$

where $\mathrm{T} \geq 0$ is the temperature parameter, $\sigma(x) = \exp(x)/(1 + \exp(x))$ is the logistic function and $\odot$ represents the elementwise product. Thus, the nodes are turned off or on stochastically depending on their input. Note that after marginalizing over the binary activations, the latent variables $\mathbf{a}_n$ are distributed as a mixture of two normals:

$$a_{n,l,d} | z_{n,l,d}, \theta \sim \sigma(z_{n,l,d}/\mathrm{T})\mathrm{N}\left(z_{n,l,d}, \eta_{l,d}^2\right) + (1 - \sigma(z_{n,l,d}/\mathrm{T}))\,\mathrm{N}\left(0, \eta_{l,d}^2\right), \tag{2}$$

where the variance $\eta_{l,d}^2$ is the (d, d)th element of $\mathbf{\Sigma}_l$, and

$$\begin{aligned} \mathbb{E}[a_{n,l,d} | z_{n,l,d}, \theta] &= \mathbb{E}[\mathbb{E}[a_{n,l,d} | z_{n,l,d}, \gamma_{n,l,d}, \theta]] = \mathbb{E}[\gamma_{n,l,d} z_{n,l,d}] \\ &= \sigma(z_{n,l,d}/\mathrm{T}) z_{n,l,d}, \end{aligned} \tag{3}$$

$$\begin{aligned} \mathbb{V}(a_{n,l,d} | z_{n,l,d}, \theta) &= \mathbb{E}[\mathbb{V}(a_{n,l,d} | z_{n,l,d}, \gamma_{n,l,d}, \theta)] + \mathbb{V}(\mathbb{E}[a_{n,l,d} | z_{n,l,d}, \gamma_{n,l,d}, \theta]) \\ &= \mathbb{E}[\eta_{l,d}^2] + \mathbb{V}(\gamma_{n,l,d} z_{n,l,d}) \\ &= \eta_{l,d}^2 + z_{n,l,d}^2 \sigma(z_{n,l,d}/\mathrm{T}) \left(1 - \sigma(z_{n,l,d}/\mathrm{T})\right). \end{aligned} \tag{4}$$

We display the conditional distribution of $a_{n,l,d}$ in Figure 1, for different combinations of the temperature parameter T and variance $\eta_{l,d}^2$. The ReLU activation is recoverd in the case of $T = 0$ and $\eta_{l,d}^2 = 0$, while other choices of T and $\eta_{l,d}^2$ generalize the ReLU. The density resembles a bow tie, hence the name of the model.

2.2 Shrinkage Priors

Prior elicitation in Bayesian neural networks is challenging, as understanding how the high-dimensional weights map to the functions implemented by the network is not trivial. Standard Gaussian priors are often a default choice, also due to their link with ℓ_2 regularization in maximum a posteriori (MAP) inference; indeed, such priors were used in Smith et al. (2021). For an overview and discusson on priors in Bayesian neural networks, see Fortuin (2022).

3

We take an alternative approach to the Gaussian priors of Smith et al. (2021) in order to sparsify our model. Sparsity-inducing priors ease the problem of storage and computational costs, have been shown to provide improvement in the predictive performance of deep models, and can provide a data-driven approach to selecting the width and depth, easing the difficult task of specifying the network architecture. For these reasons, such priors have been considered in a large number of works to produce sparse BNNs. For classical proposals on the two-group discrete mixture priors with a point mass at zero (referred to as spike-and-slab priors) in high-dimensional regression, see George and McCulloch (1993); Mitchell and Beauchamp (1988). More recently, many consider spike-and-slab priors on the neural network weights and provide both theoretical guarantees and demonstrate empirical improvements Blundell et al. (2015); Polson and Ročková (2018); Bai et al. (2020); Sun et al. (2022); Lee and Lee (2022). As an alternative to spike-and-slab priors, shrinkage priors employ a single distribution to approximate the spike-and-slab shape, yet are more computationally attractive, as they avoid exploring the space of all possible models (corresponding to selecting subsets of inputs/nodes). Moreover, (nearly) optimal theoretical guarantees of spike-and-slab priors can still be obtained with shrinkage priors in the linear regression setting Song and Liang (2023). A popular choice of shrinkage prior is the horseshoe Carvalho et al. (2009); Piironen and Vehtari (2017), and imposing regularized horseshoe priors on the BNN weights combined with a structured variational approximation provides competitive empirical results Ghosh et al. (2018). Another example arising in neural networks is a Gaussian scale mixture prior with automatic relevance determination, which when combined with suitable approximate Bayesian inference is equivalent to introducing dropout regularization in NN Nalisnick et al. (2019). Recently, Castillo and Egels (2024) shows that a suitably rescaled heavy-tailed prior on the neural network weights achieves automatic adaptation, simultaneously to both the intrinsic dimension and smoothness of the underlying function, and near-optimal minimax contraction rates of the fractional posterior distribution and its mean-field variational approximation.

In this work, we focus on a class of continuous shrinkage priors, namely, global-local normal scale-mixtures with generalized inverse Gaussian shrinkage priors on the scale parameters, referred to as *global-local normal-generalized inverse Gaussian priors* Griffin and Brown (2021). Global-local scale-mixtures aim to shrink less important weights whilst leaving large ones, which is achieved through a global parameter controlling the overall shrinkage, with the local parameters allowing deviations at the level of individual nodes Polson and Scott (2012); Bhadra et al. (2019). This choice of priors is also motivated by the theoretical guarantees for high-dimensional regression Song and Liang (2023); Griffin and Brown (2010); Polson et al. (2013), for a survey on global-local shrinkage methods we refer to Griffin and Brown (2021).

The N-GIG priors on the weights have the following hierarchical structure:

$$\mathbb{P}(\mathbf{W}_l|\boldsymbol{\psi}_l, \tau_l) = \prod_{d=1}^{D_l} \prod_{d'=1}^{D_{l-1}} \mathrm{N}\left(W_{l,d,d'}|0, \tau_l \psi_{l,d,d'}\right), \tag{5}$$

$$\mathbb{P}(\boldsymbol{\psi}_l) = \prod_{d=1}^{D_l} \prod_{d'=1}^{D_{l-1}} \mathrm{GIG}\left(\psi_{l,d,d'} \mid \nu_{\mathrm{loc},l}, \delta_{\mathrm{loc},l}, \lambda_{\mathrm{loc},l}\right), \tag{6}$$

$$\mathbb{P}(\tau_l) = \mathrm{GIG}\left(\tau_l \mid \nu_{\mathrm{glob}}, \delta_{\mathrm{glob}}, \lambda_{\mathrm{glob}}\right), \tag{7}$$

where τ_l is the global shrinkage parameter for layer l and $\psi_{l,d,d'}$ is the local shrinkage parameter for each weight. The generalized inverse Gaussian (GIG) prior is given by:

$$\mathrm{GIG}\left(\psi \mid \nu, \delta, \lambda\right) \propto \psi^{\nu-1} \exp\left(-\frac{1}{2}(\delta^2/\psi + \lambda^2 \psi)\right),$$

with parameters ν, δ, and λ; for a proper prior, $\nu > 0$ if $\delta = 0$ or $\nu < 0$ if $\lambda = 0$. In Equation (6), we allow the GIG parameters for the local scale parameters $\psi_{l,d,d'}$ to vary across layers to adjust local shrinkage for wider layers. Furthermore, to encourage more shrinkage for larger depth and width, we scale the global parameters τ_l with respect to L and the local parameters $\psi_{l,d,d'}$ with respect to D_l (details of our approach are provided Appendix C.1).
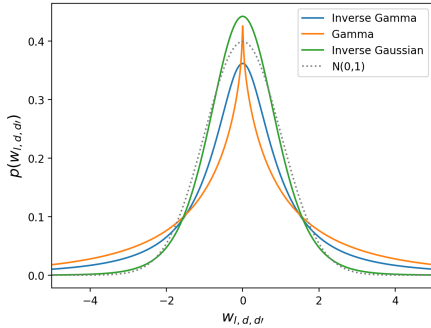
When the global shrinkage parameter τ_l is fixed, examples of the marginal distribution for $w_{l,d,d'}$ include Laplace Park and Casella (2008), Student-t (ST) Tipping (2001), Normal-Gamma (NG) Caron and Doucet (2008); Griffin and Brown (2010), Normal inverse Gaussian (NIG) Caron and Doucet (2008). Each example

has a different tail behavior, inducing different forms of shrinkage (see Table 1 for a overview and Figure 2 (a) for a visualization). Note that if the prior is polynomial-tailed, then for large signals the amount of shrinkage is mitigated even given small τ_l Polson and Scott (2012). The global shrinkage parameter τ_l leads
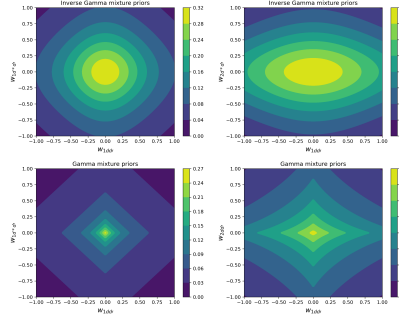
Table 1: Examples with the class of N-GIG priors.

Marginal for $w_{l,d,d'}$ when τ_l is fixed				
	Student-T	Laplace	NG	NIG
Mixing distribution	IG	Gamma	Gamma	IGauss
Parameters	$\nu < 0, \delta > 0, \lambda = 0$	$\nu = 1, \delta = 0, \lambda$	$\nu, \delta = 0, \lambda$	$\nu = \frac{1}{2}, \delta, \lambda$
Tail behavior	polynomial-tailed	exponential-tailed	exponential-tailed	exponential-tailed

to a non-separable penalty for the weights within the same layer, i.e. after integrating out τ_l, the weights within the same layer are dependent. This is illustrated in Figure 2 (b), which provides contour plots for the joint marginal distribution of two weights, within the same layer (dependence, left column) and across different layers (independence, right column), for the two choice of IG and Gamma mixing priors. Figure 2 (b) also highlights how the variance depends on the width of the layer, with more hidden units and smaller variance for the second layer compared to the first.

(a) Marginal prior for the weights.   (b) Joint prior for the weights.

Figure 2: Illustration of the prior on the weights. (a) the marginal density of the weights for different choices within the GIG family. (b) the joint prior on two weights within the same layer (left) and across layers (right) for the two choices of IG (top) and Gamma (bottom) mixing priors.

The opposite effects of varying width and depth in deep neural networks are studied in Vladimirova et al. (2021); while depth accentuates a model's non-Gaussianity, the width makes models increasingly Gaussian. Indeed, infinitely wide BNNs are closely related to Gaussian processes (GPs), typically relying on appropriately scaled i.i.d. Gaussian weights Neal (2012); Lee et al. (2018); Matthews et al. (2018) and relaxing these assumptions, e.g. through ordering, constraints, heavy tails, or bottlenecks, results in non-Gaussian limits, such as stable processes Peluchetti et al. (2020), deep GPs Agrawal et al. (2020) or more exotic processes Sell and Singh (2023); Chada et al. (2022). The sparsity promoting priors in Equations (5) to (7) provide a framework for the data to inform on the width and depth of the network.

2.3 Polya-Gamma Data Augmentation

As in Smith et al. (2021), we employ *Polya-Gamma data augmentation* Polson et al. (2013) to render the model conditionally linear and Gaussian. First, recall the definition of the Polya-Gamma distribution with parameters $b > 0$ and $c \in \mathbb{R}$, denoted PG(b, c). The random variable $X \sim$ PG(b, c) if

$$X \stackrel{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/4\pi^2}, \quad \text{where } g_k \stackrel{iid}{\sim} \text{Gam}(b, 1).$$

The key identity that we use is:

$$\frac{\exp(z)^a}{(1+\exp(z))^b} = 2^{-b}\exp(\kappa z)\int_0^\infty \exp(-\frac{\omega z^2}{2})p(\omega)d\omega, \tag{8}$$

where $\kappa = a - b/2$ and $p(\omega) = PG(\omega|b,0)$. The integral is a Gaussian kernel, thus if $z = \mathbf{w}^T\mathbf{x}$, conditioned on the latent variable ω, $\mathbf{w}$ has a Gaussian distribution and conditioned on $\mathbf{w}$, ω has a PG distribution. While to sample from the PG distribution, one can use the alternating series method of Devroye (2006), all finite moments of the PG random variables are available in closed form and that becomes useful for expectation-maximization or variational Bayes algorithms. Specifically, for $c > 0$

$$\mathbb{E}[\omega] = \frac{b}{2c}\frac{\exp(c)-1}{1+\exp(c)}. \tag{9}$$

Moreover, the PG distribution is closed under convolution with the same scale parameter; if $\omega_1 \sim PG(b_1,c)$ and $\omega_2 \sim PG(b_2,c)$, then $\omega_1 + \omega_2 \sim PG(b_1+b_2,c)$.

2.4 Augmented Model

The model described in Section 2.1 augmented with stochastic activations $\mathbf{a} = (\mathbf{a}_{n,l})$ and binary activations $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{n,l})$ is:

$$p(\mathbf{y},\mathbf{a},\boldsymbol{\gamma}|\boldsymbol{\theta}) = \prod_{n=1}^N \mathrm{N}\left(\mathbf{y}_n \mid \mathbf{z}_{n,L+1}, \boldsymbol{\Sigma}_{L+1}\right) \prod_{n=1}^N \prod_{l=1}^L \mathrm{N}\left(\mathbf{a}_{n,l} \mid \boldsymbol{\gamma}_{n,l} \odot \mathbf{z}_{n,l}, \boldsymbol{\Sigma}_l\right)$$

$$\times \prod_{d=1}^{D_l} \frac{\exp(z_{n,l,d}/T)^{\gamma_{n,l,d}}}{1+\exp(z_{n,l,d}/T)}.$$

Then using the Equation (8), the last term can be written as:

$$\frac{\exp(z_{n,l,d}/T)^{\gamma_{n,l,d}}}{1+\exp(z_{n,l,d}/T)} = 2^{-1}\exp\left(\frac{\kappa_{n,l,d}z_{n,l,d}}{T}\right)\int_0^\infty \exp\left(-\frac{\omega_{n,l,d}z_{n,l,d}^2}{2T^2}\right)p(\omega_{n,l,d})d\omega_{n,l,d},$$

where $\omega_{n,l,d} \sim PG(1,0)$ and $\kappa_{n,l,d} = \gamma_{n,l,d} - 1/2$. Thus, introducing the additional augmented variables $\boldsymbol{\omega} = (\omega_{n,l,d})$, we arrive at the augmented model:

$$p(\mathbf{y},\mathbf{a},\boldsymbol{\gamma},\boldsymbol{\omega}|\boldsymbol{\theta}) \propto \prod_{n=1}^N \mathrm{N}\left(\mathbf{y}_n \mid \mathbf{z}_{n,L+1}, \boldsymbol{\Sigma}_{L+1}\right) \prod_{n=1}^N \prod_{l=1}^L \mathrm{N}\left(\mathbf{a}_{n,l} \mid \boldsymbol{\gamma}_{n,l} \odot \mathbf{z}_{n,l}, \boldsymbol{\Sigma}_l\right)$$

$$\times \prod_{d=1}^{D_l} \exp\left(\frac{\kappa_{n,l,d}z_{n,l,d}}{T}\right)\exp\left(-\frac{\omega_{n,l,d}z_{n,l,d}^2}{2T^2}\right)p(\omega_{n,l,d}).$$

To ease computations, the covariance matrices are assumed to be diagonal $\boldsymbol{\Sigma}_l = \mathrm{diag}(\eta_{l,1}^2, \ldots \eta_{l,D_l}^2)$, and the variances are denoted by $\boldsymbol{\eta}_l = (\eta_{l,1}^2, \ldots \eta_{l,D_l}^2)$. Additionally, we assume conjugate priors for the variances $\eta_{l,d}^2 \overset{iid}{\sim} \mathrm{IG}(\alpha_0^h, \beta_0^h)$ for $l = 1, \ldots, L$ and $\eta_{L+1,d}^2 \overset{iid}{\sim} \mathrm{IG}(\alpha_0, \beta_0)$ and for the biases $b_{l,d} \overset{iid}{\sim} \mathrm{N}(0, s_0^2)$. Here, we consider different prior parameters α_0^h, β_0^h for the variance terms associated to the hidden layers in comparison to the prior parameters α_0, β_0 for the final layer. In particular, α_0, β_0 are chosen to reflect uncertainty in noise, while α_0^h, β_0^h are chosen so that prior concentrates on small values and realizations of the stochastic activation function are more similar to the ReLU.

A graphical model of the bow tie BNN with stochastic relaxation and shrinkage priors is displayed in Figure 3, and the posterior distribution over both the model parameters and latent variables is:

$$p(\mathbf{a},\boldsymbol{\gamma},\boldsymbol{\omega},\mathbf{W},\mathbf{b},\boldsymbol{\eta},\boldsymbol{\psi},\boldsymbol{\tau}) \propto \prod_{n=1}^N \mathrm{N}\left(\mathbf{y}_n \mid \mathbf{z}_{n,L+1}, \boldsymbol{\Sigma}_{L+1}\right) \prod_{n=1}^N \prod_{l=1}^L \mathrm{N}\left(\mathbf{a}_{n,l} \mid \boldsymbol{\gamma}_{n,l} \odot \mathbf{z}_{n,l}, \boldsymbol{\Sigma}_l\right)$$
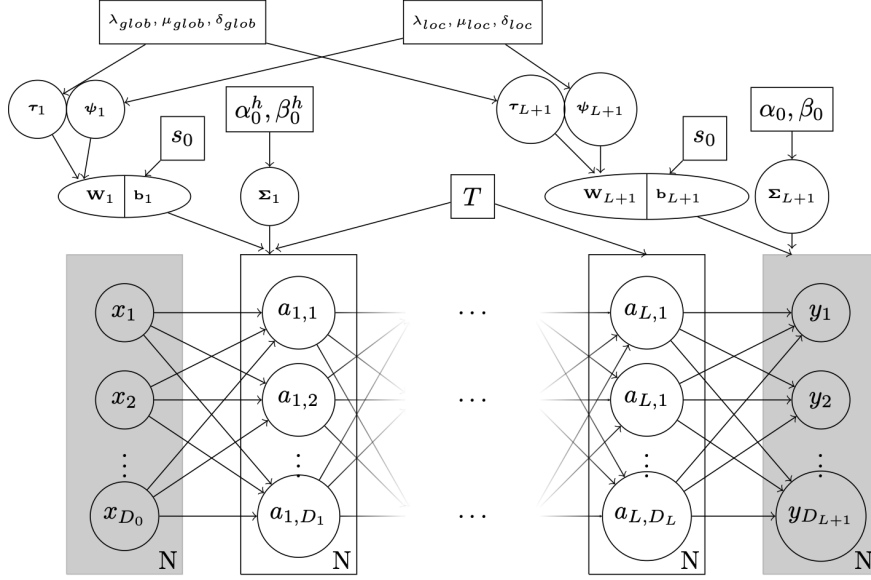
Figure 3: Directed Acyclic Graph of the model.

$$
\times \prod_{n=1}^{N} \prod_{l=1}^{L} \prod_{d=1}^{D_l} \exp\left(\frac{\kappa_{n,l,d} z_{n,l,d}}{T}\right) \exp\left(-\frac{\omega_{n,l,d} z_{n,l,d}^2}{2T^2}\right) p(\omega_{n,l,d})
$$

$$
\times \prod_{n=1}^{N} \prod_{l=1}^{L} \prod_{d=1}^{D_l} \mathrm{Bern}\left(\gamma_{n,l,d} \mid \sigma\left(\frac{z_{n,l,d}}{T}\right)\right)
$$

$$
\times \prod_{d=1}^{D_{L+1}} \mathrm{IG}(\eta_{L+1,d}^2 \mid \alpha_0, \beta_0) \times \prod_{l=1}^{L} \prod_{d=1}^{D_l} \mathrm{IG}(\eta_{l,d}^2 \mid \alpha_0^h, \beta_0^h)
$$

$$
\times \prod_{l=1}^{L} \left(\prod_{d=1}^{D_l} \left(\mathrm{N}(b_{l,d} \mid 0, s_0^2) \times \prod_{d'=1}^{D_{l-1}} \mathrm{N}\left(W_{l,d,d'} \mid 0, \tau_l \psi_{l,d,d'}\right) \right) \right)
$$

$$
\times \prod_{l=1}^{L} \left(\mathrm{GIG}\left(\tau_l \mid \nu_{\mathrm{glob}}, \delta_{\mathrm{glob}}, \lambda_{\mathrm{glob}}\right) \prod_{d=1}^{D_l} \prod_{d'=1}^{D_{l-1}} \mathrm{GIG}\left(\psi_{l,d,d'} \mid \nu_{\mathrm{loc},l}, \delta_{\mathrm{loc},l}, \lambda_{\mathrm{loc},l}\right) \right).
$$

3   Inference

3.1   Variational Bayes

While Markov chain Monte Carlo is considered the gold-standard tool for approximating posterior distributions in Bayesian modelling due its asymptotic guarantees, MCMC algorithms can be prohibitively slow when the model dimension and sample size are large. An alternative fast approximate Bayesian inference method known as *variational inference* has gained popularity in the literature Ormerod and Wand (2010); Zhang et al. (2018), due to both the explosion in the amount of data collected and use of highly parametrized models for increased flexibility. VI has been shown to yield reasonably accurate approximations in several problems as well as desirable frequentist properties. Namely, consistency and asymptotic normality of VI are studied in Wang and Blei (2019), theoretical guarantees for optimal contraction rates of variational posteriors under certain assumptions appear in several recent works Zhang and Gao (2020); Alquier and Ridgway (2020); Bhattacharya et al. (2023); Yang et al. (2020) and contraction rates as well as model selection consistency in mixtures are considered in Chérief-Abdellatif (2020).

Several works focus on properties of VI approximation in sparse models: oracle contraction rates of a

mean-field VI in the case of high-dimensional regression are established in Ray and Szabó (2022) and in the case of the neural networks with spike-and-slab priors, contraction rates of variational posteriors are studied in Bai et al. (2020). More recently, near-optimal contraction rates of the mean-field VI approximations are obtained in the context of neural networks with heavy-tailed priors on the weights Castillo and Egels (2024). We refer readers interested in the caveats of VI to Yao et al. (2018).

Consider fitting a model parameterized by $\boldsymbol{\theta}$ to the observed data $\mathcal{D}$. In variational inference, the true posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ is approximated by a density $q(\boldsymbol{\theta})$ taken from a family of distributions $\mathcal{F}$ that minimizes the Kullback-Leibler divergence between the approximate and true posterior, or equivalently maximizes the *evidence lower bound* (ELBO)

$$\text{ELBO} = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\log\left(\frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})}\right)\right]. \tag{10}$$

A common choice for $\mathcal{F}$ is the mean-field family on a partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J\}$ of $\boldsymbol{\theta}$, assuming that the variational posterior factorizes over (blocks) of latent variables: $q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\boldsymbol{\theta}_j)$, where $J \leq \dim(\boldsymbol{\theta})$. Without any further parametric assumptions, it has been shown Hinton and van Camp (1993); MacKay (1995); Jordan et al. (1999) that the optimal choice for each product component q_j is

$$q_j(\boldsymbol{\theta}_j) \propto \exp\left[\mathbb{E}_{-\boldsymbol{\theta}_j} \log\left(p(\boldsymbol{\theta}, \mathcal{D})\right)\right], \tag{11}$$

where the above expectation is taken with respect to $\prod_{j' \neq j} q_{j'}(\boldsymbol{\theta}_{j'})$. The process of sweeping through the components of the partition and updating one at a time via Equation (11) is known as *coordinate ascent variational inference* (CAVI). Wand et al. (2011) studies CAVI's performance and improvement techniques in the case of elaborate distributions. Limitations of the mean-field assumption can be found in Wand et al. (2011); Neville et al. (2014); Coker et al. (2022).

We specify the mean-field family for the approximate variational posterior:

$$q(\mathbf{a}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{W}, \mathbf{b}, \boldsymbol{\eta}, \boldsymbol{\psi}, \boldsymbol{\tau}) = q(\mathbf{a})q(\boldsymbol{\gamma})q(\boldsymbol{\omega})q(\mathbf{W}, \mathbf{b})q(\boldsymbol{\eta})q(\boldsymbol{\psi})q(\boldsymbol{\tau}).$$

Note that the assumption on the family above could be referred to as the *strucutred mean-field assumtion*. Importantly, unlike existing variational algorithms for BNNs, we do not make any assumptions on independence of parameters between layers. The calculation of each component of the variational posterior is given in the Appendix A, where using Equation (11) we obtain the following update steps.

Global shrinkage parameters: the parameters $\boldsymbol{\tau}$ are independent across layers (and can be updated in parallel) with a GIG variational posterior:

$$q(\boldsymbol{\tau}) = \prod_{l}^{L+1} \text{GIG}\left(\tau_l \mid \hat{\nu}_{\text{glob},l}, \hat{\delta}_{\text{glob},l}, \lambda_{\text{glob}}\right),$$

where for $l = 1, \ldots, L+1$,

$$\hat{\nu}_{\text{glob},l} = \nu_{\text{glob}} - \frac{D_l D_{l-1}}{2} \quad \text{and} \quad \hat{\delta}_{\text{glob},l} = \sqrt{\delta_{\text{glob}}^2 + \sum_{d}^{D_l} \sum_{d'}^{D_{l-1}} \mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right] \mathbb{E}\left[W_{l,d,d'}^2\right]}.$$

Local shrinkage parameters: the parameters $\boldsymbol{\psi}$ are independent across and within layers (and can be updated in parallel) with a GIG variational posterior:

$$q(\boldsymbol{\psi}) = \prod_{l}^{L+1} \prod_{d}^{D_l} \prod_{d'}^{D_{l-1}} \text{GIG}\left(\psi_{l,d,d'} \mid \hat{\nu}_{\text{loc},l,d,d'}, \hat{\delta}_{\text{loc},l,d,d'}, \lambda_{\text{loc},l}\right),$$

where for $l = 1, \ldots, L+1$, $d = 1, \ldots, D_l$, $d' = 1, \ldots, D_{l-1}$,

$$\hat{\nu}_{\text{loc},l,d,d'} = \nu_{\text{loc},l} - \frac{1}{2} \quad \text{and} \quad \hat{\delta}_{\text{loc},l,d,d'} = \sqrt{\mathbb{E}\left[\frac{1}{\tau_l}\right] \mathbb{E}\left[W_{l,d,d'}^2\right] + \delta_{\text{loc},l}^2}.$$

Covariance matrix: the diagonal elements of the covariance matrix $\boldsymbol{\eta}_l$ are independent across and within layers (and can be updated in parallel) with an inverse-Gamma variational posterior:

$$q(\boldsymbol{\eta}) = \prod_l^{L+1} \prod_d^{D_l} \mathrm{IG}(\eta_{l,d}^2 \mid \alpha_{l,d}, \beta_{l,d}),$$

where for the hidden layers $l = 1, \ldots, L$, the updated variational parameters for $d = 1, \ldots, D_l$ are given by

$$\alpha_{l,d} = \alpha_0^h + \frac{N}{2},$$

$$\beta_{l,d} = \beta_0^h + \frac{1}{2} \sum_n^N \left(\mathbb{E}\left[a_{n,l,d}\right] - \mathbb{E}\left[\gamma_{n,l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right] \right)^2 + \mathbb{E}\left[a_{n,l,d}^2\right] - \mathbb{E}\left[a_{n,l,d}\right]^2$$

$$+ \frac{1}{2} \sum_n^N \mathbb{E}\left[\gamma_{n,l,d}\right] \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1} \widetilde{\mathbf{a}}_{n,l-1}^T\right] \right) - \mathbb{E}\left[\gamma_{n,l,d}\right]^2 \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T\right] \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}^T\right] \right).$$

And for the final layer, the updated variational parameters for $d = 1, \ldots, D_{L+1}$ are given by

$$\alpha_{L+1,d} = \alpha_0 + \frac{N}{2},$$

$$\beta_{L+1,d} = \beta_0 + \frac{1}{2} \sum_n^N \left(y_{n,d} - \mathbb{E}\left[\widetilde{\mathbf{W}}_{L+1,d}\right] \mathbb{E}[\widetilde{\mathbf{a}}_{n,L}] \right)^2$$

$$+ \frac{1}{2} \sum_n^N \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{L+1,d}^T \widetilde{\mathbf{W}}_{L+1,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L} \widetilde{\mathbf{a}}_{n,L}^T\right] \right) - \mathrm{Tr}\left(\mathbb{E}[\widetilde{\mathbf{W}}_{L+1,d}]^T \mathbb{E}[\widetilde{\mathbf{W}}_{L+1,d}] \mathbb{E}[\widetilde{\mathbf{a}}_{n,L}] \mathbb{E}[\widetilde{\mathbf{a}}_{n,L}^T] \right).$$

Note that in the above, $\mathbf{W}_{l,d}$ represents the d-th row of the weight matrix $\mathbf{W}_l$. Additionally, for $l = 1, \ldots, L+1$ we introduce the notation $\widetilde{\mathbf{W}}_{l,d} = (b_{l,d}, \mathbf{W}_{l,d})$ and $\widetilde{\mathbf{W}} = (\mathbf{b}, \mathbf{W})$, and let the vector $\widetilde{\mathbf{a}}_{n,l}$ represent the stochastic activation augmented with an entry of one, i.e. $\widetilde{\mathbf{a}}_{n,l} = (1, \mathbf{a}_{n,l}^T)^T$.

Weights and biases: the weights and biases are independent across layers and within layer, independent across the D_l regression problems, with a Gaussian variational posterior:

$$q(\mathbf{b}, \mathbf{W}) = \prod_l^{L+1} \prod_d^{D_l} \mathrm{N}\left(\widetilde{\mathbf{W}}_{l,d} \mid \mathbf{m}_{l,d}, \mathbf{B}_{l,d} \right),$$

where for the hidden layers $l = 1, \ldots, L$, the updated variational parameters for $d = 1, \ldots, D_l$ are given by

$$\mathbf{B}_{l,d}^{-1} = \mathbf{D}_{l,d}^{-1} + \sum_n^N \left(\frac{1}{T^2} \mathbb{E}\left[\omega_{n,l,d}\right] + \mathbb{E}\left[(\eta_{l,d})^{-2}\right] \mathbb{E}\left[\gamma_{n,l,d}\right] \right) \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1} \widetilde{\mathbf{a}}_{n,l-1}^T\right],$$

$$\mathbf{m}_{l,d}^T = \mathbf{B}_{l,d} \left(\sum_n^N \mathbb{E}\left[(\eta_{l,d})^{-2}\right] \mathbb{E}\left[\gamma_{n,l,d}\right] \mathbb{E}\left[a_{n,l,d} \widetilde{\mathbf{a}}_{n,l-1}\right] + \frac{1}{T} \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right] \left(\mathbb{E}\left[\gamma_{n,l,d}\right] - \frac{1}{2} \right) \right),$$

and for the final layer, the updated variational parameters for $d = 1, \ldots, D_{L+1}$ are given by

$$\mathbf{B}_{L+1,d}^{-1} = \mathbf{D}_{L+1,d}^{-1} + \mathbb{E}\left[(\eta_{L+1,d})^{-2}\right] \sum_n^N \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L+1} \widetilde{\mathbf{a}}_{n,L+1}^T\right],$$

$$\mathbf{m}_{L+1,d}^T = \mathbf{B}_{L+1,d} \mathbb{E}\left[(\eta_{L+1,d})^{-2}\right] \left(\sum_n^N y_n \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L+1}\right] \right),$$

where for $l = 1, \ldots, L+1$ and $d = 1, \ldots, D_l$,

$$\mathbf{D}_{l,d}^{-1} = \mathrm{diag}\left(s_0^{-2}, \mathbb{E}\left[\tau_l^{-1}\right] \mathbb{E}\left[\psi_{l,d,1}^{-1}\right], \ldots, \mathbb{E}\left[\tau_l^{-1}\right] \mathbb{E}\left[\psi_{l,d,D_{l-1}}^{-1}\right] \right).$$

Polya-Gamma augmented variables: $\boldsymbol{\omega}$ are independent across observations $n = 1, \ldots, N$, layers $l = 1, \ldots, L$, and width $d = 1, \ldots, D_l$, with a Polya-Gamma variational posterior:

$$q(\boldsymbol{\omega}) = \prod_n^N \prod_l^L \prod_d^{D_l} \mathrm{PG}(\omega_{n,l,d} \mid 1, A_{n,l,d}), \tag{12}$$

with updated variational parameters:

$$A_{n,l,d} = \frac{1}{T} \sqrt{\left(\mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d} \right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1} \widetilde{\mathbf{a}}_{n,l-1}^T \right] \right) \right)}.$$

Note that simulating from or evaluating the density of the PG is not necessary, and the CAVI updates of the other parameters only require computing the expectation of $\boldsymbol{\omega}$ with respect to the variational posterior in Equation (12), which is straightforward to compute (Equation (9)).

Binary activations: $\boldsymbol{\gamma}$ are independent across observations $n = 1, \ldots, N$, layers $l = 1, \ldots, L$, and width $d = 1, \ldots, D_l$, with a Bernoulli variational posterior:

$$q(\boldsymbol{\gamma}) = \prod_n^N \prod_l^L \prod_d^{D_l} \mathrm{Bern}\left(\gamma_{n,l,d} \mid \rho_{n,l,d} \right), \tag{13}$$

with

$$\rho_{n,l,d} = \sigma\left(-\frac{\mathbb{E}\left[\eta_{l,d}^{-2} \right]}{2} \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d} \right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1} \widetilde{\mathbf{a}}_{n,l-1}^T \right] \right) + \mathbb{E}\left[\eta_{l,d}^{-2} \right] \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d} \right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1} a_{n,l,d} \right] + \frac{1}{T} \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d} \right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1} \right] \right).$$

We illustrate the variational posterior of $\boldsymbol{\gamma}$ for the toy example of Section 4.1 in Figure 4.

Stochastic activations: $\mathbf{a}$ are independent across observations $n = 1, \ldots, N$ and conditionally Gaussian given the previous layer with variational posterior:

$$q(\mathbf{a}) = \prod_{n=1}^N \prod_{l=1}^L \mathrm{N}\left(\mathbf{a}_{n,l} \mid \mathbf{t}_{n,l} + \mathbf{M}_{n,l} \mathbf{a}_{n,l-1}, \mathbf{S}_{n,l} \right), \tag{14}$$

where denote $\mathbf{a}_{n,0} := \mathbf{x}_n, \mathbf{S}_{n,L} := \mathbf{S}_L$ and the updated variational parameters for $n = 1, \ldots, N$ and $l = 1, \ldots, L-1$ are

$$\mathbf{S}_{n,l}^{-1} = \hat{\boldsymbol{\Sigma}}_l^{-1} - \mathbf{M}_{n,l+1}^T \mathbf{S}_{n,l+1}^{-1} \mathbf{M}_{n,l+1} + \sum_{d=1}^{D_{l+1}} \left(\mathbb{E}\left[\frac{1}{\eta_{l+1,d}^2} \right] \mathbb{E}\left[\gamma_{n,l+1,d} \right] + \frac{1}{T^2} \mathbb{E}\left[\omega_{n,l+1,d} \right] \right) \mathbb{E}\left[\mathbf{W}_{l+1,d}^T \mathbf{W}_{l+1,d} \right],$$

$$\mathbf{t}_{n,l} = \mathbf{S}_{n,l} \left(\mathbf{M}_{n,l+1}^T \mathbf{S}_{n,l+1}^{-1} \mathbf{t}_{n,l+1} + \hat{\boldsymbol{\Sigma}}_l^{-1} \mathbb{E}\left[\boldsymbol{\gamma}_{n,l} \right] \odot \mathbb{E}\left[\mathbf{b}_l \right] + \frac{1}{T} \sum_{d=1}^{D_{l+1}} \mathbb{E}\left[\mathbf{W}_{l+1,d}^T \right] \left(\mathbb{E}\left[\gamma_{n,l+1,d} \right] - \frac{1}{2} \right) - \right.$$

$$\left. - \sum_{d=1}^{D_{l+1}} \left(\mathbb{E}\left[\frac{1}{\eta_{l+1,d}^2} \right] \mathbb{E}\left[\gamma_{n,l+1,d} \right] + \frac{1}{T^2} \mathbb{E}\left[\omega_{n,l+1,d} \right] \right) \mathbb{E}\left[\mathbf{W}_{l+1,d} b_{l+1,d} \right] \right),$$

$$\mathbf{M}_{n,l} = \mathbf{S}_{n,l} \hat{\boldsymbol{\Sigma}}_l^{-1} \mathbb{E}\left[\boldsymbol{\gamma}_{n,l} \right] \mathbf{1}_{D_{l-1}}^T \odot \mathbb{E}\left[\mathbf{W}_l \right],$$

$$\hat{\boldsymbol{\Sigma}}_l^{-1} = \mathrm{diag}\left(\mathbb{E}\left[\eta_{l,1}^{-2} \right], \ldots, \mathbb{E}\left[\eta_{l,D_l}^{-2} \right] \right).$$

And for the final layer with $n = 1, \ldots, N$ and $l = L$,

$$\mathbf{S}_L^{-1} = \hat{\boldsymbol{\Sigma}}_L^{-1} + \sum_{d=1}^{D_{L+1}} \mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2} \right] \mathbb{E}\left[\mathbf{W}_{L+1,d}^T \mathbf{W}_{L+1,d} \right],$$
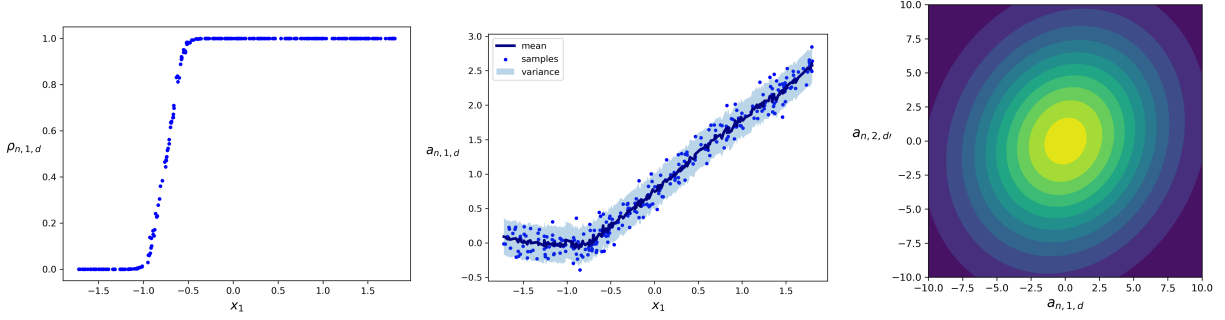
10

Figure 4: Variational posterior for $\rho_{n,1,d}$ (on the left) and $a_{n,1,d}$ (in the middle), joint distribution of $(a_{n,1,d}, a_{n,2,d'})$ (on the right) in the case of the toy example of Section 4 and particular values of d, d'.

$$\mathbf{t}_{n,L} = \mathbf{S}_L \left(\left(\sum_{d=1}^{D_{l+1}} \mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right] \left(-\mathbb{E}\left[\mathbf{W}_{L+1,d}^T b_{L+1,d}\right] + \mathbb{E}\left[\mathbf{W}_{L+1,d}^T\right] y_{n,d} \right) \right) + \hat{\mathbf{\Sigma}}_L^{-1} \mathbb{E}\left[\boldsymbol{\gamma}_{n,L}\right] \odot \mathbb{E}\left[\mathbf{b}_L\right] \right),$$

$$\mathbf{M}_{n,L} = \mathbf{S}_L \hat{\mathbf{\Sigma}}_L^{-1} \mathbb{E}\left[\boldsymbol{\gamma}_{n,L}\right] \mathbf{1}_{D_{L-1}}^T \odot \mathbb{E}\left[\mathbf{W}_L\right],$$

$$\hat{\mathbf{\Sigma}}_L^{-1} = \mathrm{diag}\left(\mathbb{E}\left[\eta_{L,1}^{-2}\right], \ldots, \mathbb{E}\left[\eta_{L,D_L}^{-2}\right] \right).$$

Figure 4 illustrates on the toy example of Section 4.1 how the variational posterior of the stochastic activations (middle) resembles a smoothed, noisy ReLU. Due the independence assumption between the stochastic and binary activations, the potentially bimodal bow tie distribution (Equation (2)) is approximated with a unimodal Gaussian in the variational framework, which may better approximate the true posterior when the temperature is not too large relative to the noise (see Figure 1). In addition, the proposed approximation has the advantage of avoiding explicit assumptions of independence between layers, allowing to capture the dependence between the stochastic activations of across layers visualized for the toy example in Figure 4 (right).

The corresponding optimization objective, i.e. the ELBO in Equation (10), is available in the closed form and provided in the Appendix B.

3.2   VI with EM

The hyperparameters can play a crucial role in Bayesian neural networks. When dealing with the sparsity-inducing priors setting an excessively large scale parameter weakens the shrinkage effects, whilst choosing a scale parameter that is too small may wipe out the effects of the important hidden nodes. Manually picking suitable values is challenging, and instead, we seek a more efficient strategy, utilizing the similarity between the variational and *expectation-maximization (EM) algorithms*. Specifically, we investigate the hybrid scheme combining VI with an EM step Osborne et al. (2022) so that the steps of the CAVI algorithm proceed with the EM update to set the hyperparameter for global shrinkage variable $\boldsymbol{\tau}$. Due to weak identifiability, we do not jointly update global and local hyperparameters. Let h_{glob} represent δ_{glob} or λ_{glob} and consider the ELBO treated as a function of h_{glob}, then the optimal values as approximate MAP estimates are:

$$h_{\mathrm{glob}} = \arg\max \mathbb{E}_{\mathrm{glob}}[\mathrm{ELBO}],$$

where

$$\mathbb{E}_{\mathrm{glob}}[\mathrm{ELBO}] = \mathbb{E}\left[\sum_{l=1}^{L+1} \log(\mathrm{GIG}(\tau_l \mid \nu_{\mathrm{glob}}, \delta_{\mathrm{glob}}, \lambda_{\mathrm{glob}}) \right]$$

$$= (L+1)\left(\nu_{\mathrm{glob}}\left(\log(\lambda_{\mathrm{glob}}) - \log(\delta_{\mathrm{glob}})\right) - \log\left(2K_{\nu_{\mathrm{glob}}}(\lambda_{\mathrm{glob}}\delta_{\mathrm{glob}})\right) \right)$$

$$+ \sum_{l=1}^{L+1} (\nu_{\mathrm{glob}} - 1)\mathbb{E}\left[\log \tau_l\right] - \frac{1}{2}\left(\delta_{\mathrm{glob}}^2 \mathbb{E}\left[\frac{1}{\tau_l}\right] + \lambda_{\mathrm{glob}}^2 \mathbb{E}\left[\tau_l\right] \right).$$

In the case of the IG priors, one's aim is to set optimal δ_{glob}, in the case of the Gamma and IGauss priors, the parameters of interest are λ_{glob}. We provide specific examples of the shrinkage parameters and the corresponding optimal values in Appendix D.2. The result of combining CAVI and the EM algorithm is described in Algorithm 1.

Algorithm 1 VI with EM

Require: Initialize hyperparameters
  while ELBO has not converged **do**
    for $l = 1, \ldots, L$ **do**
      update $\nu_{\text{glob},l}$ and $\delta_{\text{glob},l}$ {parameters of τ_l}
      update $\nu_{\text{loc},l,d,d'}$ and $\delta_{\text{loc},l,d,d'}$ {parameters of $\psi_{l,d,d'}$ for $d = 1 \ldots D_l$, $d' = 1 \ldots D_{l-1}$}
      update $\alpha_{l,d}$ and $\beta_{l,d}$ {parameters of $\eta_{l,d}$ for $d = 1 \ldots D_l$}
      update $A_{n,l,d}$ {parameter of $\omega_{n,l,d}$ for $d = 1 \ldots D_l$, $n = 1 \ldots N$}
    end for
    update $\nu_{\text{glob},L+1}$ and $\delta_{\text{glob},L+1}$ {parameters of τ_{L+1}}
    update $\nu_{\text{loc},L+1,d,d'}$ and $\delta_{\text{loc},L+1,d,d'}$ {parameters of $\psi_{L+1,d,d'}$ for $d = 1 \ldots D_y$, $d' = 1 \ldots D_L$}
    update $\alpha_{L+1,d}$ and $\beta_{L+1,d}$ for $d = 1 \ldots D_y$
    for $l = 1, \ldots, L$ **do**
      update $\mathbf{S}_{n,l}$, $\mathbf{M}_{n,l}$ and $\mathbf{t}_{n,l}$ {parameters of $\mathbf{a}_{n,l}$ for $n = 1 \ldots N$ }
    end for
    for $l = 1, \ldots, L$ **do**
      update $\mathbf{B}_{l,d}$ and $\mathbf{m}_{l,d}$ {parameters of $(b_{l,d}, \mathbf{W}_{l,d})$ for $d = 1 \ldots D_l$}
      update $\rho_{n,l,d}$ {parameter of $\gamma_{n,l,d}$ for $d = 1 \ldots D_l$, $n = 1 \ldots N$}
    end for
    update $\mathbf{B}_{L+1,d}$ and $\mathbf{m}_{L+1,d}$ {parameters of $(b_{L+1,d}, \mathbf{W}_{L+1,d})$ for $d = 1 \ldots D_y$}
    update h_{glob} {EM for global hyperparameter}
  end while

3.3 Inferring the Network Structure

The choice of the network architecture has significant practical implications on the generalization of the model, and so sparsity-promoting priors for the network weights have emerged as promising approach to allow for a data-driven choice of architecture. Instead of the classical two-group discrete mixture priors, in this article, we focus on a class of continuous shrinkage priors. While this results in more tractable computations, it also implies non-zero posterior means and draws and doesn't lead to automatic network architecture selection. Several post-processing methods have been proposed to yield a sparse solution (see e.g.Piironen et al. (2020); Li and Pati (2017); Griffin (2024)). The method known as decoupling shrinkage and selection (DSS) Hahn and Carvalho (2015) obtains sparse estimates of the weights by minimizing the sum of the predictive loss function with a parsimony-inducing penalty. An alternative approach is the the penalized credible regions (PenCR) method Zhang et al. (2021), which identifies the "sparsest" solution in posterior credible regions corresponding to different levels; it is shown to perform well in the case of global-local shrinkage priors and under certain assumptions, PenCR produces the same results as DSS. Similarly, we propose to make use of credible intervals to select nodes. Following Li and Lin (2010), we implement an automatic *credible interval criterion* which selects a node as long as its credible interval doesn't cover zero. Specifically, recall that the variational posterior of the weights is $W_{l,d,d'} \sim \text{N}(m_{l,d,d'}^W, B_{l,d,d'}^W)$ for $l = 1, \ldots, L, d = 1, \ldots, D_l, d' = 1, \ldots, D_{l-1}$, where $\mathbf{m}_{l,d}^W$ and $\mathbf{B}_{l,d}^W$ denote the subsets of the mean $\mathbf{m}_{l,d}$ and covariance matrix $\mathbf{B}_{l,d}$ corresponding to the weights. Then, we obtain sparse weights $\widehat{W}_{l,d,d'}$ with sparse variational distribution $\widehat{q}(b_{l,d}, \widehat{\mathbf{W}}_{l,d})$ for some $l \in \mathcal{L} \subseteq \{1, \ldots, L+1\}$, $d \in \mathcal{D}_l \subseteq \{1, \ldots, D_l\}, d' \in \mathcal{D}_{l-1} \subseteq \{1, \ldots, D_{l-1}\}$, defined by setting:

$$\widehat{W}_{l,d,d'} \sim \begin{cases} \text{N}\left(m_{l,d,d'}^W, \left(B_{l,d}^W\right)_{d'd'}\right) & \text{if } \max\left(Q(W_{l,d,d'} > 0), Q(W_{l,d,d'} < 0)\right) \geq \kappa, \\ \delta_0 & \text{otherwise,} \end{cases}$$

where $Q(W_{l,d,d'} < 0) = 1 - Q(W_{l,d,d'} > 0) = \Phi(-m^W_{l,d,d'}/\sqrt{(B^W_{l,d})_{d'd'}})$.

The threshold κ is chosen to control the *Bayesian false discovery rate*, which is calculated as

$$\widehat{\text{FDR}}(\kappa) = \frac{\sum_{l,d,d'}(1 - \mathcal{Q}_{l,d.d'})\mathbf{1}(\mathcal{Q}_{l,d.d'} > \kappa)}{\sum_{l,d,d'}\mathbf{1}(\mathcal{Q}_{l,d.d'} > \kappa)},$$

with $\mathcal{Q}_{l,d.d'} = \max\left(Q(W_{l,d,d'} > 0), Q(W_{l,d,d'} < 0)\right)$. Specifically, for a specified error rate α, κ is set to satisfy $\widehat{\text{FDR}}(\kappa) < \alpha$. Algorithm 2 describes the node selection procedure which begins with ordering $\mathcal{Q}_{l,d.d'}$ in the descending order and going down through the thresholds to assign κ to the smallest $\mathcal{Q}_{l,d.d'}$ such that its false discovery rate doesn't exceed α.

Algorithm 2 Node selection algorithm

Require: $\mathcal{I} = \{\mathcal{Q}_{l,d.d'} \mid l = 1 \ldots L, \ d = 1 \ldots D_{l+1}, \ d' = 1 \ldots D_l\}$.
  $\hat{\kappa} = \max(\mathcal{I})$
  $\mathcal{I} = \mathcal{I} \setminus \hat{\kappa}$
  if $\widehat{\text{FDR}}(\max(\mathcal{I})) < \alpha$ **then**
    $\hat{\kappa} = \max(\mathcal{I})$
    $\mathcal{I} = \mathcal{I} \setminus \hat{\kappa}$
  else
    break
  end if
  for $l = 1 \ldots L, \ d = 1 \ldots D_{l+1}, \ d' = 1 \ldots D_l$ **do**
    if $\mathcal{Q}_{l,d,d'} \geq \hat{\kappa}$ **then**
      $\widehat{W}_{l,d,d'} \sim \text{N}\left(m^W_{l,d,d'}, \left(B^W_{l,d}\right)_{d'd'}\right)$
    else
      $\widehat{W}_{l,d,d'} = 0$ a.s.
    end if
  end for
  for $l = L+1, \ldots 2, d = 1, \ldots D_l,$ **do**
    if $\widehat{W}_{l,d} = 0$ a.s. **then**
      $\widehat{W}_{l-1,d',d} = 0$ a.s. $\forall \, d' = 1, \ldots, D_{l-1}$
    else
      if $\widehat{W}_{l-1,d',d} = 0$ a.s. $\forall \, d' = 1, \ldots, D_{l-1}$ **then**
        $\widehat{W}_{l,d} = 0$ a.s.
      end if
    end if
  end for
Ensure: $\hat{q}(b_{l,d}, \widehat{\mathbf{W}}_{l,d}), \ l \in \mathcal{L}, \ d \in \mathcal{D}_l, \ d' \in \mathcal{D}_{l-1}$.

Once we sweep through $\mathcal{Q}_{l,d.d'}$, we do a backwards pass to remove the nodes with no connections. If the node has no outgoing connections then all the incoming connections need to be removed, and conversely, if the node has no incoming connections, then all the outgoing connections can be removed. An example of the network resulting after applying the Algorithm 2 is illustrated by the Figure 7.

3.4   Predictions

For a new $\mathbf{x}_*$, the predictive distribution of $\mathbf{y}_*$ given the data is approximated as:

$$p(\mathbf{y}_* \mid \mathbf{x}_*, \mathcal{D}) = \int p(\mathbf{y}_* \mid \mathbf{x}_*, \boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta} \mid \mathcal{D}, \mathbf{x}_*) \, d\boldsymbol{\theta}$$

$$= \int p(\mathbf{y}_* \mid \mathbf{a}_*, \mathbf{W}, \mathbf{b}, \boldsymbol{\eta})p(\mathbf{a}_*, \mathbf{W}, \mathbf{b}, \boldsymbol{\eta} \mid \mathcal{D}, \mathbf{x}_*) \, d\mathbf{a}_* \, d\mathbf{W} \, d\mathbf{b} \, d\boldsymbol{\eta}$$

$$\approx \int p(\mathbf{y}_* \mid \mathbf{a}_*, \mathbf{W}, \mathbf{b}, \boldsymbol{\eta}) q(\mathbf{a}_*) q(\mathbf{W}, \mathbf{b}) q(\boldsymbol{\eta}) \, d\mathbf{a}_* \, d\mathbf{W} \, d\mathbf{b} \, d\boldsymbol{\eta}$$

$$= \int \mathrm{N}\left(\mathbf{y}_* \mid \mathbf{W}_{L+1}\mathbf{a}_{*,L} + \mathbf{b}_{L+1}, \boldsymbol{\Sigma}_{L+1}\right) q(\mathbf{a}_{*,L}) q(\mathbf{W}_{L+1}, \mathbf{b}_{L+1}) q(\boldsymbol{\eta}_{L+1}) \, d\mathbf{a}_{*,L} \, d\mathbf{W}_{L+1} \, d\mathbf{b}_{L+1} \, d\boldsymbol{\eta}_{L+1}. \tag{15}$$

Equation (15) requires first computing the approximate variational predictive distributions $q(\mathbf{a}_*), q(\boldsymbol{\gamma}_*)$ and $q(\boldsymbol{\omega}_*)$, which are updated in a similar way to Section 3.1.

Specifically, the stochastic activations are conditionally Gaussian given the previous layer with variational predictive distribution:

$$q(\mathbf{a}_*) = \prod_{l=1}^{L} \mathrm{N}\left(\mathbf{a}_{*,l} \mid \mathbf{t}_{*,l} + \mathbf{M}_{*,l}\mathbf{a}_{*,l-1}, \mathbf{S}_{*,l}\right),$$

where $\mathbf{a}_{*,0} = \mathbf{x}_*$. For the final layer, we have:

$$\mathbf{S}_{*,L}^{-1} = \hat{\boldsymbol{\Sigma}}_L^{-1}; \quad \mathbf{t}_{*,L} = \mathbb{E}\left[\boldsymbol{\gamma}_{*,L}\right] \odot \mathbb{E}\left[\mathbf{b}_L\right]; \quad \mathbf{M}_{*,L} = \mathbb{E}\left[\boldsymbol{\gamma}_{*,L}\right] \mathbf{1}_{D_{L-1}}^T \odot \mathbb{E}\left[\mathbf{W}_L\right].$$

For all other layers $l = 1, \dots, L - 1$, we have:

$$\mathbf{S}_{*,l}^{-1} = \hat{\boldsymbol{\Sigma}}_l^{-1} - \mathbf{M}_{*,l+1}^T \mathbf{S}_{*,l+1}^{-1} \mathbf{M}_{*,l+1} + \sum_{d=1}^{D_{l+1}} \left(\mathbb{E}\left[\frac{1}{\eta_{l+1,d}^2}\right] \mathbb{E}\left[\gamma_{*,l,d}\right] + \frac{1}{T^2} \mathbb{E}\left[\omega_{*,l+1,d}\right] \right) \mathbb{E}\left[\mathbf{W}_{l+1,d}^T \mathbf{W}_{l+1,d}\right],$$

$$\mathbf{t}_{*,l} = \mathbf{S}_{*,l} \left(\mathbf{M}_{*,l+1}^T \mathbf{S}_{*,l+1}^{-1} \mathbf{t}_{*,l+1} + \hat{\boldsymbol{\Sigma}}_l^{-1} \mathbb{E}\left[\boldsymbol{\gamma}_{*,l}\right] \odot \mathbb{E}\left[\mathbf{b}_l\right] - \sum_{d=1}^{D_{l+1}} \mathbb{E}\left[\frac{1}{\eta_{l+1,d}^2}\right] \mathbb{E}\left[\gamma_{*,l,d}\right] \mathbb{E}\left[\mathbf{W}_{l+1,d}^T b_{l+1,d}\right] \right.$$

$$\left. + \frac{1}{T} \sum_{d=1}^{D_{l+1}} \mathbb{E}\left[\mathbf{W}_{l+1,d}^T\right] \left(\mathbb{E}\left[\gamma_{*,l+1,d}\right] - \frac{1}{2} \right) - \frac{1}{T^2} \sum_{d=1}^{D_{l+1}} \mathbb{E}\left[\omega_{*,l+1,d}\right] \mathbb{E}\left[\mathbf{W}_{l+1,d} b_{l+1,d}\right] \right),$$

$$\mathbf{M}_{*,l} = \mathbf{S}_{*,l} \hat{\boldsymbol{\Sigma}}_l^{-1} \mathbb{E}\left[\boldsymbol{\gamma}_{*,l}\right] \mathbf{1}_{D_{l-1}}^T \odot \mathbb{E}\left[\mathbf{W}_l\right].$$

The binary activations are independent across layers $l = 1, \dots, L$ and width $d = 1, \dots, D_l$, with a Bernoulli variational predictive distribution:

$$q(\boldsymbol{\gamma}_*) = \prod_{l}^{L} \prod_{d}^{D_l} \mathrm{Bern}\left(\gamma_{*,l,d} \mid \rho_{*,l,d}\right), \tag{16}$$

with

$$\rho_{*,l,d} = \sigma\left(-\frac{\mathbb{E}\left[\eta_{l,d}^{-2}\right]}{2} \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{*,l-1} \widetilde{\mathbf{a}}_{*,l-1}^T\right] \right) + \mathbb{E}\left[\eta_{l,d}^{-2}\right] \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{*,l-1} a_{*,l,d}\right] + \frac{1}{T} \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{*,l-1}\right] \right).$$

Finally, the Polya-Gamma augmented variables are independent across layers $l = 1, \dots, L$ and width $d = 1, \dots, D_l$, with a Polya-Gamma variational predictive distribution:

$$q(\boldsymbol{\omega}_*) = \prod_{l}^{L} \prod_{d}^{D_l} \mathrm{PG}(\omega_{*,l,d} \mid 1, A_{*,l,d}), \tag{17}$$

with updated variational parameters:

$$A_{*,l,d} = \frac{1}{T} \sqrt{\left(\mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{*,l-1} \widetilde{\mathbf{a}}_{*,l-1}^T\right] \right) \right)}.$$

Thus, before computing predictions, we first iterate to update the variational predictive distributions of $\mathbf{a}_*$, $\boldsymbol{\gamma}_*$, and $\boldsymbol{\omega}_*$. The corresponding ELBO (derived in the Appendix B.2) is monitored for convergence. While a closed-form expression for the integral in Equation (15) is unavailable, generating samples from the variational predictive is straightforward;

$$\mathbf{y}_*^{(j)} \sim \mathrm{N}\left(\mathbf{y}_* \mid \mathbf{W}_{L+1}^{(j)} \mathbf{a}_{*,L}^{(j)} + \mathbf{b}_{L+1}^{(j)}, \boldsymbol{\Sigma}_{L+1}^{(j)}\right), \tag{18}$$

for $j = 1, \ldots J$, where $(\mathbf{W}_{L+1}^{(j)}, \mathbf{b}_{L+1}^{(j)}) \sim q(\mathbf{W}_{L+1}, \mathbf{b}_{L+1})$, $\boldsymbol{\eta}_{L+1}^{(j)} \sim q(\boldsymbol{\eta}_{L+1})$, and $\mathbf{a}_{*,l}^{(j)} \mid \mathbf{a}_{*,l-1}^{(j)} \sim q(\mathbf{a}_{*,l}|\mathbf{a}_{*,l-1})$ for $l = 1, \ldots, L$, are iid draws from the variational posterior. These samples can be used to obtain a Monte Carlo approximation to investigate potential non-normality in the predictive distribution in Equation (15) and to compute credible intervals based on the highest posterior density region.

We can also compute the expectation and variance of $\mathbf{y}_*$ in closed-form. Specifically, the expectation of $\mathbf{y}_*$ under the variational predictive distribution is:

$$\mathbb{E}[\mathbf{y}_* \mid \mathbf{x}_*, \mathcal{D}] \approx \mathbb{E}_{q_{L+1}}[\mathbf{W}_{L+1}]\mathbb{E}_{q_{*,L}}[\mathbf{a}_{*,L}] + \mathbb{E}_{q_{L+1}}[\mathbf{b}_{L+1}], \tag{19}$$

where recursively

$$\mathbb{E}_{q_{*,L}}[\mathbf{a}_{*,L}] = \mathbb{E}_{q_{*,L-1}}\left[\mathbb{E}_{q(\mathbf{a}_{*,L}|\mathbf{a}_{*,L-1})}[\mathbf{a}_{*,L}]\right] = \mathbf{t}_{*,L} + \mathbf{M}_{*,L}E_{q_{*,L-1}}[\mathbf{a}_{*,L-1}].$$

Similarly, the variational variance of $\mathbf{y}_*$ is

$$\text{Var}[y_{*,d} \mid \mathbf{x}_*, \mathcal{D}] \approx \text{Var}_{q_{L+1}}[\mathbf{W}_{L+1,d}\mathbf{a}_{*,L} + \mathbf{b}_{L+1,d}] + \mathbb{E}_{q_{L+1}}\left[\boldsymbol{\eta}_{L+1,d}^2\right],$$

where the first term represents the signal variance and is computed as

$$\begin{aligned}\text{Var}_{q_{L+1}}[\mathbf{W}_{L+1,d}\mathbf{a}_{*,L} + \mathbf{b}_{L+1,d}] &= \mathbb{E}_{q_{L+1}}\left[(\mathbf{W}_{L+1,d}\mathbf{a}_{*,L} + \mathbf{b}_{L+1,d})^2\right] - \left(\mathbb{E}_{q_{L+1}}[\mathbf{W}_{L+1,d}]\mathbb{E}_{q_L}[\mathbf{a}_{*,L}] + \mathbb{E}_{q_{L+1}}[\mathbf{b}_{L+1,d}]\right)^2 \\ &= \text{Tr}\left(\mathbb{E}_{q_{L+1}}\left[\mathbf{W}_{L+1,d}^T\mathbf{W}_{L+1,d}\right]\mathbb{E}_{q_L}\left[\mathbf{a}_{*,L}\mathbf{a}_{*,L}^T\right] - \mathbb{E}_{q_{L+1}}[\mathbf{W}_{L+1,d}]^T\mathbb{E}_{q_{L+1}}[\mathbf{W}_{L+1,d}]\mathbb{E}_{q_L}[\mathbf{a}_{*,L}]\mathbb{E}_{q_L}[\mathbf{a}_{*,L}]^T\right) \\ &\quad + \text{Var}_{q_{L+1}}(\mathbf{b}_{L+1,d}) + 2\text{Cov}_{q_{L+1}}(\mathbf{W}_{L+1,d}, \mathbf{b}_{L+1,d})\mathbb{E}_{q_L}[\mathbf{a}_{*,L}],\end{aligned}$$

which requires the recursive computation:

$$\mathbb{E}_{q_L}\left[\mathbf{a}_{*,L}\mathbf{a}_{*,L}^T\right] = \mathbf{S}_{*,L} + \mathbf{t}_{*,L}\mathbf{t}_{*,L}^T + 2\mathbf{M}_{*,L}\mathbb{E}_{q_{L-1}}[\mathbf{a}_{*,L-1}]\mathbf{t}_{*,L}^T + \mathbf{M}_{*,L}\mathbb{E}_{q_{L-1}}\left[\mathbf{a}_{*,L-1}\mathbf{a}_{*,L-1}^T\right]\mathbf{M}_{*,L}^T.$$

Sparse prediction. To save on both computation and storage, the variational predictive distribution can be computed based on the sparse variational posterior (Section 3.3). For a new data point $\mathbf{x}_*$, we obtain expectation and variance of $\mathbf{y}_*$ by first computing the sparse versions of variational predictive distributions $\widehat{q}(\mathbf{a}_*), \widehat{q}(\boldsymbol{\gamma}_*)$ and $\widehat{q}(\boldsymbol{\omega}_*)$ as in Equations (14), (16) and (17) by plugging $\widehat{q}(b_{l,d}, \widehat{\mathbf{W}}_{l,d})$ instead of the $q(b_{l,d}, \mathbf{W}_{l,d})$, which only requires updates for the subset of nodes with nonzero weights.

3.5 Ensembles of Variational Approximations

While the variational algorithm described in Section 3.1 increases the ELBO at each epoch, the ELBO is a non-convex function of the variational parameters and only convergence to a local optimum is guaranteed. Due to identifiability issues, the posterior distribution of a Bayesian neural network is highly multimodal, and exploring this posterior is notoriously challenging Papamarkou et al. (2022). A single variational approximation tends to concentrate around one mode and can understate posterior uncertainty. Several approaches have been proposed to overcome such issues. Recently, Ohn and Lin (2024) introduce adaptive variational inference which achieves optimal posterior contraction rate and model selection consistency by considering several variational approximations obtained in different models. Yao et al. (2022) introduce an approach which uses parallel runs of inference algorithms to cover as many modes of the posterior distribution as possible and then combines these using Bayesian stacking.

In a similar but simpler fashion to the above proposal, we consider an ensemble of variational approximations, obtained by running in parallel the variational algorithm multiple times with different random starting points. In this case, letting $k = 1, \ldots, K$ index the different variational approximations, we compute the weight w_k associated with each approximation in accordance with the optimization objective, the ELBO:

$$w_k \propto \exp\left(\text{ELBO}_k\right).$$

We can interpret this as a Bayesian model averaging across the K different models/approximations. While in an ideal setting, the weights would be proportional to the marginal likelihood for each model, the use of the ELBO is motivated as it provides a lower bound to the marginal likelihood and can be computed

in closed form. Next, we compute predictions by taking a weighted average of the predictive distributions of each model (given in Equation (19)), that is

$$\mathbb{E}[\mathbf{y}_* \mid \mathbf{x}_*, \mathcal{D}] \approx \sum_{k=1}^{K} w_k \mathbb{E}_{q_k}[\mathbf{y}_* \mid \mathbf{x}_*, \mathcal{D}],$$

where each expectation is taken with respect to q_k (the kth variational approximation). Similarly, we can compute the variance as

$$\mathrm{Var}(\mathbf{y}_* \mid \mathbf{x}_*, \mathcal{D}) \approx \sum_{k=1}^{K} w_k \, \mathrm{Var}_{q_k}(\mathbf{y}_* \mid \mathbf{x}_*, \mathcal{D}) + \sum_{k=1}^{K} w_k \left(\mathbb{E}_{q_k}[\mathbf{y}_* \mid \mathbf{x}_*, \mathcal{D}]\right)^2 - \left(\sum_{k=1}^{K} w_k \mathbb{E}_{q_k}[\mathbf{y}_* \mid \mathbf{x}_*.\mathcal{D}]\right)^2.$$

The following approach has the potential to improve both predictive accuracy and uncertainty quantification. Once again, we can investigate the variational predictive distribution (beyond the mean and variance) by first sampling a model with probability $(w_1, \ldots, w_K)$ and then given that selected model k, generating a sample $\mathbf{y}_*$ from the kth variational predictive distribution (as described in Equation (18)).

4 Experiments

We evaluate the variational bow tie neural network (VBNN) on several datasets. First, we consider a simple nonlinear synthetic example to compare with a ground truth. We then validate VBNN on the diabetes dataset, first considered in Efron et al. (2004) to demonstrate the least angle regression (LARS) algorithm for variable selection, and subsequently, used in different proposals for sparsity-promoting priors and algorithms (e.g. Park and Casella (2008); Li and Lin (2010)). Lastly, we consider a range of popular regression datasets from the UCI Machine Learning Repository M. et al. (2007).

The importance of suitable initialization choice in NNs is well known Wenzel et al. (2020); Daniely et al. (2016); He et al. (2015), and we design two possible random initialization schemes of the VBNN, which are described in Appendix C.1 and used in all experiments. Convergence of the ELBO is monitored during the training and prediction stages, where if three consecutive measurements of ELBO for training differ by less than the specified threshold, the phase is stopped and the model moves to the prediction stage, where we proceed similarly. In most experiments, the thresholds during the training and prediction stages are set, respectively, to $1e-5$ and $1e-4$. We compare the performance of VBNN to two popular variational frameworks under the mean-field assumption: the stochastic variational inference approximation (SVI) implemented in Numpyro Phan et al. (2019), and Bayes by Backprop (BBB) Blundell et al. (2015) implemented with Pytorch. For all the datasets, we evaluate the performance over 10 random splits, where we use 90% of the data for the training and 10% for testing the model. We normalize the input but do not re-scale the output, we record the root mean squared error (RMSE), the predictive negative log-likelihood of the test data (NLL) and the empirical coverage, see Appendix C for additional implementation details and the definition of each metric.

4.1 Simulated example

We construct a synthetic dataset generated by first uniformly sampling a two-dimensional input vector $\mathbf{x}_n = (x_{n,1}, x_{n,2})$, with $x_{n,d} \sim \mathrm{Unif}([-2, 2])$, and assume only the first feature influences the output: $y_n = f(x_{n,1}) + \epsilon_n = 0.1 x_{n,1}^2 + 10 \sin(x_{n,1}) + \epsilon_n$, where $\epsilon_n \sim \mathrm{N}(0, 0.5)$. Then, the dataset consisting of $N = 300$ observations is used to investigate the performance of VBNN compared to the SVI and BBB baselines as we increase the number of hidden layers, setting $L = 1, 2$ or 5, whilst keeping the number of hidden units in each layer fixed to $D_H = 20$.

In general, for this simple non-linear example, the performance tends to deteriorate with increasing architecture complexity (larger depth), and the VBNN is the most robust to this choice (see Figure 5, which compares the performance of three of the models as a function of depth). In addition, as the number of hidden layers increases, SVI and BBB provide overly wide credible intervals while VBNN more closely aligns with the desired coverage (see Figure 6, which illustrates the empirical coverages for the observations).
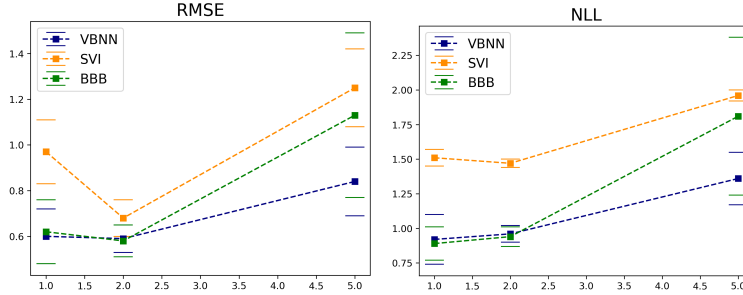
Figure 5: Performance on the simulated dataset as the depth increases. VBNN is more robust to the choice of depth and overparameterization.
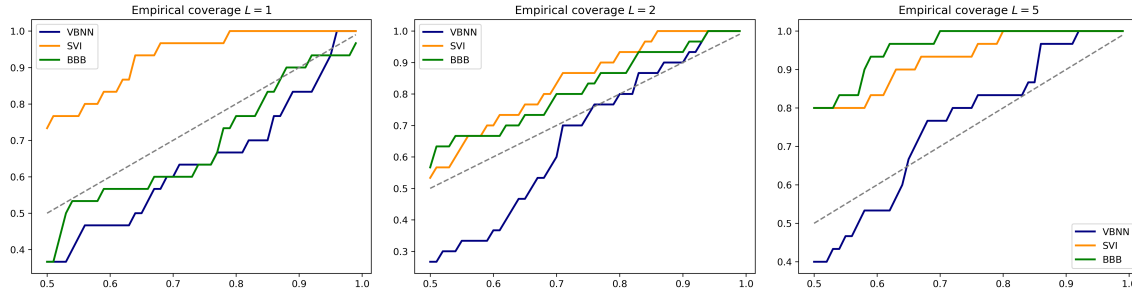

Figure 6: Empirical coverage for the observations for the simulated dataset for three different settings of network's depth.

For each depth, Figure 7 illustrates the predictive means and uncertainties computed for the observations as well as DAGs of networks' structures obtained after the post-process node selection algorithm described in Section 3.3, where the Bayesian false discovery rate is constrained by setting the error rate to $\alpha = 0.01$. The sparsity-promoting prior combined with the node selection algorithm can effectively prune the over-parametrized neural networks; for example, the sparse one-layer neural network contains only 11 hidden nodes with 30 total edges/weights from the initial $D_H = 20$ with 60 total edges/weights. Moreover, the estimated regression function and credible intervals both from the variational predictive and the sparse variation predictive distribution, recover the true function well. In this way, VBNN provides an effective tool to reduce predictive computational complexity and storage as well as ease interpretation. Note that the predictions show no relation with the coordinate x_2 (Figure 7d), recovering the true function, but some of the connections from x_2 are still present in the sparse network, due to identifiability issues, although with overall low weight (Figure 7e).

4.2 Diabetes example

The diabetes data consists of $n = 442$ entries obtained for $p = 10$ input variables, the response is a quantitative measurement of disease progression. The predictors are age, sex, body mass index, average blood pressure and six blood serum measurements and the goal is to determine which of these are relevant for forecasting diabetes progression.

We fit a neural network with one hidden layer $L = 1$ and $D_H = 20$ and perform the node selection algorithm with the FDR bounded by $\alpha = 0.01$. Figure 8 illustrates the shrinkage and node selection and compares the coefficients of the Lasso linear model with cross-validation (LassoCV) to the original and the sparsified weights of our model. Predictors with considerable effect obtained by both models coincide, whilst some of the variables the Lasso model excludes (e.g. age) are still present in the VBNN's estimates. Compared with Lasso, VBNN has the advtange of learning potential nonlinear relationships between disease progression and the predictors, which is explored in Figure 9, illustrating the predictive means and uncertainty of the observations of VBNN for four of the predictors (with all other predictors are fixed to their mean) compared to

(a) $L = 1$ (b) $L = 2$ (c) $L = 5$

(d) The predictive mean as a function of the second coordinate for $L = 1$.

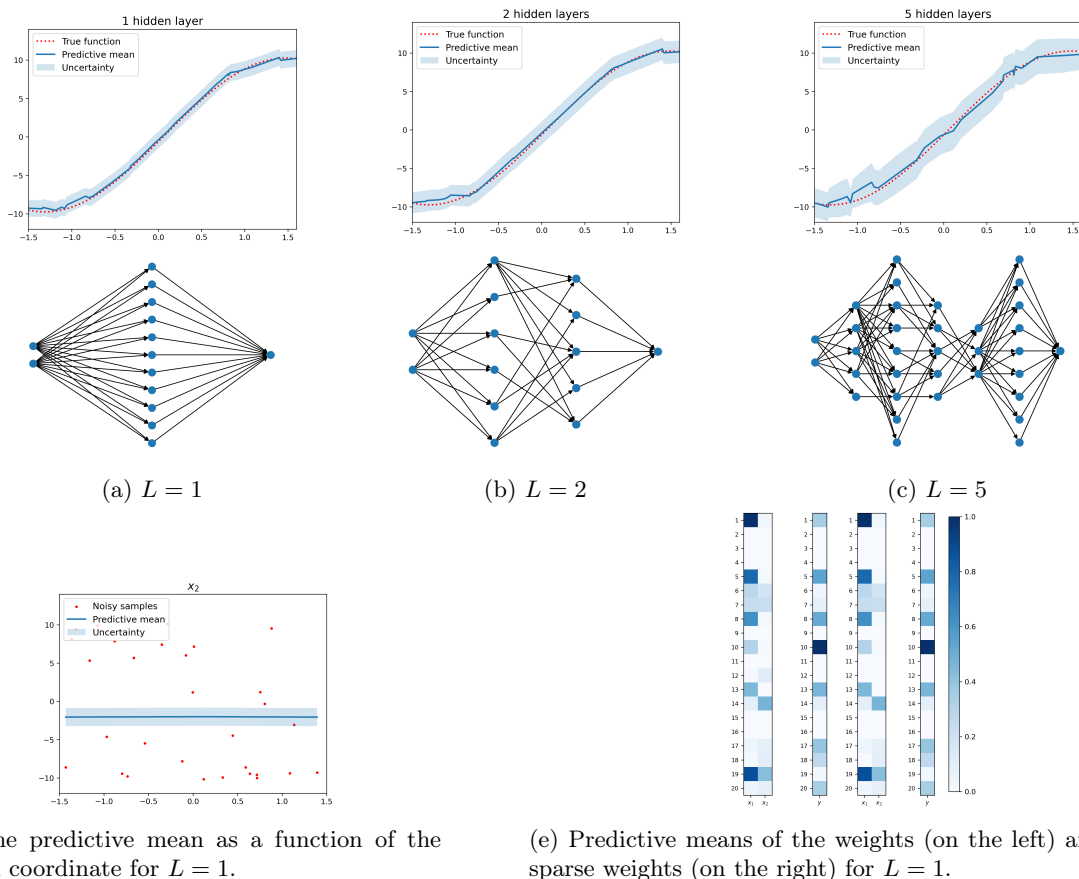(e) Predictive means of the weights (on the left) and of the sparse weights (on the right) for $L = 1$.

Figure 7: Predictive means and uncertainty estimates computed for the observations and the architecture of the network for the bound on the FDR $\alpha = 0.01$ for different settings of network's depth.

the predictions of Lasso. While the uncertainty is wide, the results suggest potential nonlinear relationships, e.g. with lamotrigine and age, the latter of which is not selected in Lasso.

Moreover, Figure 9 highlights how predictions obtained from the sparse version of the variational predictive distribution almost overlap, thus providing a reasonable, cheaper approximation. However, we note that predictive performance is only slightly improved with VBNN (see Table 2 and supplementary Figure 10 in Appendix C.3).

Table 2: RMSE, NLL and coverage for diabetes dataset.

	RMSE	NLL	Coverage
LassoCV	$53.7 \pm .5$	$5.4 \pm .13$	$.96 \pm .03$
VBNN	52.9 ± 4	$5.4 \pm .09$	$.96 \pm .03$
SVI	$55. \pm 5$	$6.5 \pm .5$	$.63 \pm .07$
BBB	$54.2 \pm .1$	$5.4 \pm .08$	$.94 \pm .02$

4.3 UCI regression datasets

Lastly, we consider publicly available datasets from the UCI Machine Learning Repository M. et al. (2007): Boston housing Harrison and Rubinfeld (1978), Energy Tsanas and Xifara (2012), Yacht dynamics J. et al. (2013), Concrete compressive strength Yeh (2007) and Concrete slump test Yeh (2009) (see Appendix C.4 for the description of the datasets). For all of the UCI regression tasks, we fit a neural network with one hidden
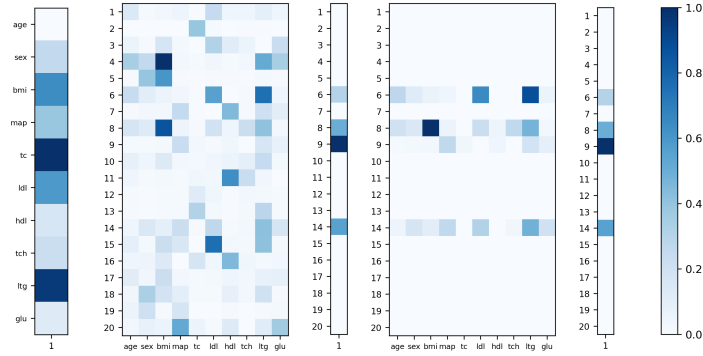
Figure 8: Coefficients of LassoCV regression (on the left), predictive means of the weights of the neural network (in the middle) and predictive means of the sparse weights obtained for $\alpha = 0.01$ (on the right).
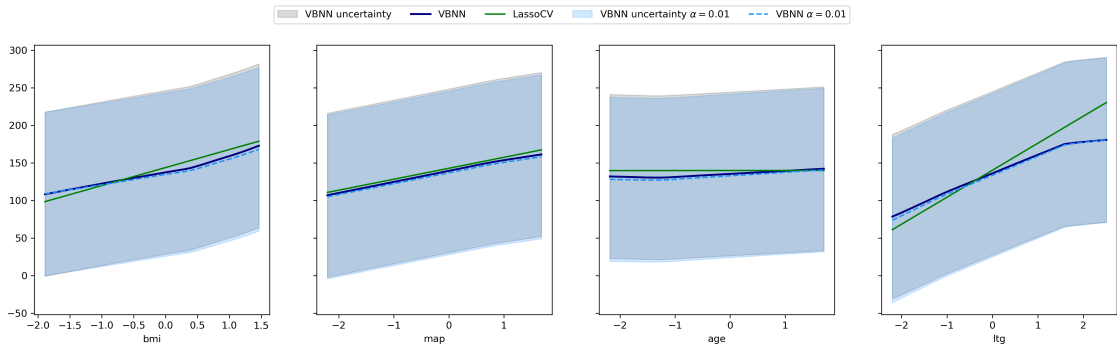

Figure 9: Slices of the predictive mean and uncertainty for observations for four predictors obtained by VBNN with and without node selection and by Lasso with cross-validation.

layer and $D_H = 50$ hidden units. Table 3 compares RMSE, NLL and empirical coverage of the observations, for VBNN, SVI and BBB baselines. Our model is more consistent in performance across various datasets than the SVI baseline and is comparable to BBB. In terms of the coverage and uncertainty quantification, VBNN performs consistently well across all datasets in contrast to the other models.

Table 3: RMSE, NLL and Coverage for UCI datasets.

Dataset	RMSE			NLL			Coverage		
	VBNN	SVI	BBB	VBNN	SVI	BBB	VBNN	SVI	BBB
Slump	6.99 ± 1.4	7.02 ± 1.0	$\mathbf{6.83 \pm 1.52}$	$\mathbf{3.37 \pm .2}$	$3.97 \pm .6$	4.33 ± 1.3	$\mathbf{.92 \pm .06}$	$.75 \pm .1$	$.84 \pm .14$
Yacht	$1.34 \pm .3$	$1.73 \pm .44$	$\mathbf{1.02 \pm .32}$	$1.74 \pm .19$	$2.08 \pm .12$	$\mathbf{.9 \pm .49}$	$.96 \pm .03$	$.99 \pm .02$	$.99 \pm .02$
Boston	$\mathbf{3.17 \pm .5}$	$3.25 \pm .54$	3.18 ± 1.13	$\mathbf{2.59 \pm .14}$	$2.95 \pm .4$	$2.69 \pm .42$	$.96 \pm .03$	$.82 \pm .04$	$.87 \pm .04$
Energy	$\mathbf{1.29 \pm .19}$	$2.35 \pm .26$	$1.88 \pm .33$	$1.7 \pm .14$	$2.31 \pm .14$	$\mathbf{1.43 \pm .19}$	$.98 \pm .01$	$.89 \pm .03$	$.99 \pm .01$
Concrete	$6.6 \pm .45$	$6.04 \pm .66$	$\mathbf{5.7 \pm .71}$	$3.31 \pm .07$	$3.22 \pm .09$	$\mathbf{3.09 \pm 0.22}$	$.96 \pm .02$	$.97 \pm .02$	$.94 \pm .03$

5 Discussion

In this paper, we presented a variational bow tie neural network (VBNN) that is amendable to Polya-gamma data augmentation so that the variational inference can be performed via the CAVI algorithm. While the idea of the stochastic relaxation described in Section 2.1 was introduced in Smith et al. (2021), the novelty of our model is in the employment of the variational inference techniques as well as sparsity-inducing priors. Namely, we implement continuous global-local shrinkage priors and propose a post-process technique for node selection. Additionally, we consider an improvement of the classical CAVI algorithm by adding EM steps for critical hyperparameters. In this way, we enrich the class of models which are handled within the

structured mean-field paradigm. We provide all the necessary (computations), techniques, and illustrative experiments demonstrating the utility of the model.

At each iteration, CAVI has to cycle through the entire data set, which can be computationally expensive and inefficient for large sample sizes. An alternative to coordinate ascent is gradient-based optimization and a future direction of this research will extend the algorithm by employing Stochastic variational inference Hoffman et al. (2013) and subsampling. Moreover, an extension to other output types, such as classification tasks, will be developed through additional Polya-gamma augmentation techniques Durante and Rigon (2019).

References

Devanshu Agrawal, Theodore Papamarkou, and Jacob Hinkle. 2020. Wide neural networks with bottlenecks are deep Gaussian processes. *Journal of Machine Learning Research* 21, 175 (2020).

Pierre Alquier and James Ridgway. 2020. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics* 48, 3 (2020), 1475–1497.

Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. 2023. A primer on Bayesian neural networks: review and debates. arXiv:2309.16314

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. 2020. Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning. In *International Conference on Learning Representations*.

Jincheng Bai, Qifan Song, and Guang Cheng. 2020. Efficient variational inference for sparse deep learning with theoretical guarantee. *Advances in Neural Information Processing Systems* 33 (2020), 466–476.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. 2019. Lasso meets horseshoe. *Statist. Sci.* 34, 3 (2019), 405–427.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. 2023. On the Convergence of Coordinate Ascent Variational Inference. arXiv:2306.01122

Christopher M Bishop. 2016. *Pattern Recognition and Machine Learning*. Springer New York.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* 112, 518 (2017), 859–877.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International Conference on Machine Learning*. PMLR, 1613–1622.

Ginevra Carbone, Luca Bortolussi, and Guido Sanguinetti. 2022. Resilience of Bayesian Layer-Wise Explanations under Adversarial Attacks. In *2022 International Joint Conference on Neural Networks*. 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892788

Ginevra Carbone, Matthew Wicker, Luca Laurenti, Andrea Patane, Luca Bortolussi, and Guido Sanguinetti. 2020. Robustness of Bayesian neural networks to gradient-based attacks. In *Advances in Neural Information Processing Systems*, Vol. 33. 15602–15613.

François Caron and Arnaud Doucet. 2008. Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning* (Helsinki, Finland) *(ICML '08)*. Association for Computing Machinery, 88–95. https://doi.org/10.1145/1390156.1390168

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. 2009. Handling Sparsity via the Horseshoe. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 5)*, David van Dyk and Max Welling (Eds.). PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 73–80.

Ismaël Castillo and Paul Egels. 2024. Posterior and variational inference for deep neural networks with heavy-tailed weights. arXiv:2406.03369

Neil K Chada, Ajay Jasra, Kody JH Law, and Sumeetpal S Singh. 2022. Multilevel Bayesian Deep Neural Networks. *Computing Research Repository* (2022).

Badr-Eddine Chérief-Abdellatif. 2020. Convergence rates of variational inference in sparse deep learning. In *International Conference on Machine Learning*. PMLR, 1831–1842.

Beau Coker, Wessel P Bruinsma, David R Burt, Weiwei Pan, and Finale Doshi-Velez. 2022. Wide mean-field bayesian neural networks ignore the data. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 5276–5333.

Amit Daniely, Roy Frostig, and Yoram Singer. 2016. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. In *Advances in Neural Information Processing Systems*, Vol. 29. Curran Associates, Inc.

Luc Devroye. 2006. Nonuniform random variate generation. *Handbooks in operations research and management science* 13 (2006), 83–121.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Daniele Durante and Tommaso Rigon. 2019. Conditionally conjugate mean-field variational Bayes for logistic models. *Statist. Sci.* 34, 3 (2019), 472 – 485. https://doi.org/10.1214/19-STS712

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. Least Angle Regression. *Annals of Statistics* (2004), 407–451.

Vincent Fortuin. 2022. Priors in bayesian deep learning: A review. *International Statistical Review* 90, 3 (2022), 563–591.

Yarin Gal. 2016. *Uncertainty in Deep Learning*. Ph. D. Dissertation. University of Cambridge.

Edward I George and Robert E McCulloch. 1993. Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* 88, 423 (1993), 881–889.

Soumya Ghosh, Jiayu Yao, and Finale Doshi-Velez. 2018. Structured variational learning of Bayesian neural networks with horseshoe priors. In *International Conference on Machine Learning*. PMLR, 1744–1753.

Jim E Griffin. 2024. Expressing and visualizing model uncertainty in Bayesian variable selection using Cartesian credible sets. arXiv:2402.12323

Jim E Griffin and Philip J Brown. 2010. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* 5, 1 (2010), 171–188.

Jim E Griffin and Philip J Brown. 2021. Bayesian global-local shrinkage methods for regularisation in the high dimension linear model. *Chemometrics and Intelligent Laboratory Systems* 210 (2021).

Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. 2023. Protein design with guided discrete diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 12489–12517.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1321–1330.

P Richard Hahn and Carlson M Carvalho. 2015. Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective. , 435–448 pages.

David Harrison and Daniel L Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 1 (1978), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. 1026–1034.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. 2019. Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 41–50.

Geoffrey E. Hinton and Drew van Camp. 1993. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory* (Santa Cruz, California, USA) *(COLT '93)*. Association for Computing Machinery, New York, NY, USA, 5–13. https://doi.org/10.1145/168304.168306

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research* (2013).

Gerritsma J., Onnink R., and Versluis A. 2013. Yacht Hydrodynamics. https://archive.ics.uci.edu/dataset/243/yacht+hydrodynamics

Antoran Javier. 2019. Bayesian Neural Networks. https://github.com/JavierAntoran/Bayesian-Neural-Networks

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An Introduction to Variational Methods for Graphical Models. *Machine Learning* 37, 2 (1999), 183–233. https://doi.org/10.1023/A:1007665907178

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. 2022. Hands-on Bayesian neural networks — A tutorial for deep learning users. *IEEE Computational Intelligence Magazine* 17, 2 (2022), 29–48.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), 583–589.

Leo Klarner, Tim G. J. Rudner, Michael Reutlinger, Torsten Schindler, Garrett M Morris, Charlotte Deane, and Yee Whye Teh. 2023. Drug Discovery under Covariate Shift with Domain-Informed Prior Distributions over Functions. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 17176–17197.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

Kody JH Law and Vitaly Zankin. 2022. Sparse online variational Bayesian regression. *SIAM/ASA Journal on Uncertainty Quantification* 10, 3 (2022), 1070–1100.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. 2018. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations*.

Kyeongwon Lee and Jaeyong Lee. 2022. Asymptotic properties for bayesian neural network in besov space. *Advances in Neural Information Processing Systems* 35 (2022), 5641–5653.

Hanning Li and Debdeep Pati. 2017. Variable selection using shrinkage priors. *Computational Statistics & Data Analysis* 107, C (2017), 107–119.

Qing Li and Nan Lin. 2010. The Bayesian elastic net. *Bayesian Analysis* 5, 1 (2010), 151 – 170. https://doi.org/10.1214/10-BA506

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

Kelly M., Longjohn R., and Nottingham K. 2007. The UCI Machine Learning Repository. https://archive.ics.uci.edu

David JC MacKay. 1995. Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* 6, 3 (1995), 469.

Alexander G de G Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. 2018. Gaussian Process Behaviour in Wide Deep Neural Networks. In *International Conference on Learning Representations*.

Rowan McAllister, Yarin Gal, Alex Kendall, Mark Van Der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. 2017. Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 4745–4753.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc.

Toby J Mitchell and John J Beauchamp. 1988. Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* 83, 404 (1988), 1023–1032.

Eric Nalisnick, José Miguel Hernández-Lobato, and Padhraic Smyth. 2019. Dropout as a structured shrinkage prior. In *International Conference on Machine Learning*. PMLR, 4712–4722.

Radford M Neal. 2012. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media.

Sarah E. Neville, John T. Ormerod, and Matthew P Wand. 2014. Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electronic Journal of Statistics* 8, 1 (2014). https://doi.org/10.1214/14-ejs910

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 427–436.

Ilsang Ohn and Lizhen Lin. 2024. Adaptive variational Bayes: Optimality, computation and applications. *The Annals of Statistics* 52, 1 (2024), 335–363.

John T. Ormerod and Matthew P Wand. 2010. Explaining variational approximations. *The American Statistician* 64, 2 (2010).

Nathan Osborne, Christine B Peterson, and Marina Vannucci. 2022. Latent network estimation and variable selection for compositional data via variational EM. *Journal of Computational and Graphical Statistics* 31, 1 (2022), 163–175.

Theodore Papamarkou, Jacob Hinkle, M. Todd Young, and David Womble. 2022. Challenges in Markov chain Monte Carlo for Bayesian neural networks. *Statist. Sci.* 37, 3 (2022), 425 – 442. https://doi.org/10.1214/21-STS840

Trevor Park and George Casella. 2008. The Bayesian Lasso. *J. Amer. Statist. Assoc.* 103, 482 (2008), 681–686.

Stefano Peluchetti, Stefano Favaro, and Sandra Fortini. 2020. Stable behaviour of infinitely wide deep neural networks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Vol. 108. 1137–1146.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. 2019. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. arXiv:1912.11554

Juho Piironen, Markus Paasiniemi, and Aki Vehtari. 2020. Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics* 14, 1 (2020), 2155 – 2197. https://doi.org/10.1214/20-EJS1711

Juho Piironen and Aki Vehtari. 2017. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 905–913.

Nicholas G Polson and Veronika Ročková. 2018. Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems* 31 (2018).

Nicholas G Polson and James G Scott. 2012. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian statistics* 9, 501-538 (2012), 105.

Nicholas G Polson, James G Scott, and Jesse Windle. 2013. Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* 108, 504 (2013), 1339–1349.

Kolyan Ray and Botond Szabó. 2022. Variational Bayes for High-Dimensional Linear Regression With Sparse Priors. *J. Amer. Statist. Assoc.* 117, 539 (2022), 1270–1281. https://doi.org/10.1080/01621459.2020.1847121

Christian Schäfer and Nicolas Chopin. 2013. Sequential Monte Carlo on large binary sampling spaces. *Statistics and Computing* 23 (2013), 163–184.

Torben Sell and Sumeetpal Sidhu Singh. 2023. Trace-class Gaussian priors for Bayesian learning of neural networks with MCMC. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 85, 1 (2023), 46–66.

Jimmy TH Smith, Dieterich Lawson, and Scott W Linderman. 2021. Bayesian Inference in Augmented Bow Tie Networks. *Proceedings of Bayesian Deep Learning Workshop* (2021).

Qifan Song and Faming Liang. 2023. Nearly optimal Bayesian shrinkage for high-dimensional regression. *Science China Mathematics* 66, 2 (2023), 409–442.

Yan Sun, Qifan Song, and Faming Liang. 2022. Learning sparse deep neural networks with a spike-and-slab prior. *Statistics & Probability Letters* 180 (2022).

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations*. ICLR.

Michael E. Tipping. 2001. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* 1 (Jan. 2001), 211–244. https://doi.org/10.1162/15324430152748236

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971

Athanasios Tsanas and Angeliki Xifara. 2012. Energy Efficiency. https://archive.ics.uci.edu/dataset/242/energy+efficiency

Mariia Vladimirova, Julyan Arbel, and Stéphane Girard. 2021. Dependence between Bayesian neural network units. In *BDL 2021-Workshop. Bayesian Deep Learning NeurIPS*. 1–9.

Matthew P Wand, John T Ormerod, Simone A Padoan, and Rudolf Frührwirth. 2011. Mean Field Variational Bayes for Elaborate Distributions. *Bayesian Analysis* 6, 4 (2011), 1–48.

Yixin Wang and David M Blei. 2019. Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* 114, 527 (2019), 1147–1161.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. 2020. How Good is the Bayes Posterior in Deep Neural Networks Really?. In *International Conference on Machine Learning*. PMLR, 10248–10259.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* 10, 2 (jan 2019).

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. 2024. SneakyPrompt: Jailbreaking Text-to-image Generative Models. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 123–123.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. 2020. α-variational inference with statistical guarantees. *The Annals of Statistics* 48, 2 (2020), 886–905.

Yuling Yao, Aki Vehtari, and Andrew Gelman. 2022. Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. *The Journal of Machine Learning Research* 23, 1 (2022), 3426–3471.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. Yes, but Did It Work?: Evaluating Variational Inference. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 5581–5590.

I-Cheng Yeh. 2007. Concrete Compressive Strength. https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength

I-Cheng Yeh. 2009. Concrete Slump Test. https://archive.ics.uci.edu/dataset/182/concrete+slump+test

Hang Yu, Laurence T. Yang, Qingchen Zhang, David Armstrong, and M. Jamal Deen. 2021. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing* 444 (2021), 92–110.

Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. 2018. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2018), 2008–2026.

Fengshuo Zhang and Chao Gao. 2020. Convergence rates of variational posterior distributions. *The Annals of Statistics* 48, 4 (2020), 2180–2207.

Tianren Zhang, Chujie Zhao, Guanyu Chen, Yizhou Jiang, and Feng Chen. 2024. Feature Contamination: Neural Networks Learn Uncorrelated Features and Fail to Generalize. In *Proceedings of The 41st International Conference on Machine Learning*.

Yan Dora Zhang, Weichang Yu, and Howard D Bondell. 2021. Variable Selection with Shrinkage Priors via Sparse Posterior Summaries. In *Handbook of Bayesian Variable Selection*. Chapman and Hall/CRC, 179–198.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. 2024. Fool Your Vision and Language Model with Embarrassingly Simple Permutations. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*. PMLR, 62892–62913.

A Derivations of the Variational Posterior

Global shrinkage parameters. Using [Equation (11)](#), the variational posterior for the global shrinkage parameters is:

$$q(\boldsymbol{\tau}) \propto \exp\left(\mathbb{E}\left[\log \prod_l^{L+1} \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \mathrm{N}\left(W_{l,d,d'}|0, \tau_l \psi_{l,d,d'}\right)\right] + \log \prod_l^{L+1} \mathrm{GIG}\left(\tau_l \mid \nu_{\mathrm{glob}}, \delta_{\mathrm{glob}}, \lambda_{\mathrm{glob}}\right)\right)$$

$$\propto \prod_l^{L+1} \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \exp \mathbb{E}\left[\log\left(\frac{1}{\sqrt{\tau_l \psi_{l,d,d'}}} \exp\left(-\frac{W_{l,d,d'}^2}{2\tau_l \psi_{l,d,d'}}\right)\right)\right] \times \prod_l^{L+1} \tau_l^{\nu_{\mathrm{glob}}-1} \exp\left(-\frac{1}{2}\left(\frac{\delta_{\mathrm{glob}}^2}{\tau_l} + \lambda_{\mathrm{glob}}^2 \tau_l\right)\right)$$

$$\propto \prod_l^{L+1} \tau_l^{\nu_{\mathrm{glob}}-1} \exp\left(-\frac{\delta_{\mathrm{glob}}^2}{2\tau_l} - \frac{\lambda_{\mathrm{glob}}^2 \tau_l}{2}\right) \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \tau_l^{-\frac{1}{2}} \exp\left(-\frac{\mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right] \mathbb{E}\left[W_{l,d,d'}^2\right]}{2\tau_l}\right)$$

$$\propto \prod_l^{L+1} \tau_l^{\nu_{\mathrm{glob}}-\frac{D_l D_{l-1}}{2}-1} \exp\left(-\frac{1}{2}\left(\frac{1}{\tau_l}\left(\sum_d^{D_l} \sum_{d'}^{D_{l-1}} \mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right] \mathbb{E}\left[W_{l,d,d'}^2\right] + \delta_{\mathrm{glob}}^2\right) + \lambda_{\mathrm{glob}}^2 \tau_l\right)\right)$$

$$\propto \prod_l^{L+1} \mathrm{GIG}\left(\tau_l \mid \hat{\nu}_{\mathrm{glob},l}, \hat{\delta}_{\mathrm{glob},l}, \lambda_{\mathrm{glob}}\right),$$

where for $l = 1, \ldots, L+1$

$$\hat{\nu}_{\mathrm{glob},l} = \nu_{\mathrm{glob}} - \frac{D_l D_{l-1}}{2},$$

$$\hat{\delta}_{\mathrm{glob},l} = \sqrt{\delta_{\mathrm{glob}}^2 + \sum_d^{D_l} \sum_{d'}^{D_{l-1}} \mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right] \mathbb{E}\left[w_{l,d,d'}^2\right]}.$$

Local shrinkage parameters. Similarly, the variational posterior for the local shrinkage parameters is:

$$q(\boldsymbol{\psi}) \propto \prod_l^{L+1} \exp\left(\mathbb{E}\left[\log \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \mathrm{N}\left(W_{l,d,d'}|0, \tau_l \psi_{l,d,d'}\right)\right] + \log \prod_{d=1}^{D_l} \prod_{d'}^{D_{l-1}} \mathrm{GIG}\left(\psi_{l,d,d'} \mid \nu_{\mathrm{loc},l}, \delta_{\mathrm{loc},l}, \lambda_{\mathrm{loc},l}\right)\right)$$

$$\propto \prod_l^{L+1} \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \exp\left(\frac{1}{2}\log\psi_{l,d,d'} - \frac{\mathbb{E}\left[\frac{1}{\tau_l}\right] \mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right] \mathbb{E}\left[W_{l,d,d'}^2\right]}{2}\right) \psi_{l,d,d'}^{\nu_{\mathrm{loc},l}-1} \exp\left(-\frac{1}{2}\left(\frac{\delta_{\mathrm{loc},l}^2}{\psi_{l,d,d'}} + \lambda_{\mathrm{loc},l}^2 \psi_{l,d,d'}\right)\right)$$

$$\propto \prod_l^{L+1} \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \psi_{l,d,d'}^{\nu_{\mathrm{loc},l}-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\frac{1}{\psi_{l,d,d'}}\left(\mathbb{E}\left[\frac{1}{\tau_l}\right] \mathbb{E}\left[W_{l,d,d'}^2\right] + \delta_{\mathrm{loc},l}^2\right) + \lambda_{\mathrm{loc},l}^2 \psi_{l,d,d'}\right)\right)$$

$$\propto \prod_l^{L+1} \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \mathrm{GIG}\left(\psi_{l,d,d'} \mid \hat{\nu}_{\mathrm{loc},l,d,d'}, \hat{\delta}_{\mathrm{loc},l,d,d'}, \lambda_{\mathrm{loc},l}\right),$$

where for $l = 1, \ldots, L+1$, $d = 1, \ldots, D_l$, D_{l-1} $d' = 1, \ldots, D_{l-1}$

$$\hat{\nu}_{\mathrm{loc},l,d,d'} = \nu_{\mathrm{loc},l} - \frac{1}{2},$$

$$\hat{\delta}_{\mathrm{loc},l,d,d'} = \sqrt{\mathbb{E}\left[\frac{1}{\tau_l}\right] \mathbb{E}\left[W_{l,d,d'}^2\right] + \delta_{\mathrm{loc},l}^2}.$$

Covariance matrix. Under the assumption of a diagonal covariance matrix, with parameters $\boldsymbol{\eta}_l = (\eta_{l,1}^2, \ldots \eta_{l,D_l}^2)$, the variational posterior is:

$$q(\boldsymbol{\eta}) \propto \exp\left(\mathbb{E}\left[\log \prod_n^N \mathrm{N}\left(\mathbf{y}_n \mid \mathbf{z}_{n,L+1}, \boldsymbol{\Sigma}_{L+1}\right) + \log \prod_n^N \prod_l^L \mathrm{N}\left(\mathbf{a}_{n,l} \mid \boldsymbol{\gamma}_{n,l} \odot \mathbf{z}_{n,l}, \boldsymbol{\Sigma}_l\right)\right]\right)$$

$$\times \prod_l^L \prod_d^{D_l} \mathrm{IG}(\eta_{l,d}^2 \mid \alpha_0^h, \beta_0^h) \prod_d^{D_{L+1}} \mathrm{IG}(\eta_{l,d}^2 \mid \alpha_0, \beta_0)$$

$$\propto \exp\left(-\frac{1}{2}\mathbb{E}\left[\sum_n^N \sum_d^{D_{L+1}} (\eta_{L+1,d})^{-2}\left(y_{n,d} - z_{n,L+1,d}\right)^2\right]\right) \prod_d^{D_{L+1}} \left((\eta_{L+1,d}^2)^{-\alpha_0 - 1 - \frac{N}{2}} \exp\left(-\frac{\beta_0}{\eta_{L+1,d}^2}\right)\right)$$

$$\times \prod_l^L \exp\left(-\frac{1}{2}\mathbb{E}\left[\sum_n^N \sum_d^{D_l}(\eta_{l,d})^{-2}\left(a_{n,l,d} - \gamma_{n,l,d} \odot z_{n,l,d}\right)^2\right]\right) \times \prod_l^L \prod_d^{D_l} \left((\eta_{l,d}^2)^{-\alpha_0^h - 1 - \frac{N}{2}} \exp\left(-\frac{\beta_0^h}{\eta_{l,d}^2}\right)\right)$$

$$\propto \prod_d^{D_{L+1}} \left((\eta_{L+1,d}^2)^{-\alpha_0 - 1 - \frac{N}{2}}\right) \exp\left(-\frac{1}{\eta_{L+1,d}^2}\left(\beta_0 + \frac{1}{2}\sum_n^N \mathbb{E}\left[(y_{n,d} - z_{n,L+1,d})^2\right]\right)\right)$$

$$\times \prod_l^L \prod_d^{D_l}(\eta_{l,d}^2)^{-\alpha_0^h - 1 - \frac{N}{2}} \exp\left(-\frac{1}{\eta_{l,d}^2}\left(\beta_0^h + \frac{1}{2}\sum_n^N \mathbb{E}\left[(a_{n,l,d} - \gamma_{n,l,d} \odot z_{n,l,d})^2\right]\right)\right).$$

Thus, $q(\boldsymbol{\eta}) \propto \prod_l^{L+1} \prod_d^{D_l} \mathrm{IG}(\alpha_{l,d}, \beta_{l,d})$, where

$$\alpha_{l,d} = \alpha_0^h + \frac{N}{2}, \quad d = 1, \ldots, D_l, \quad l = 1, \ldots, L,$$

$$\alpha_{L+1,d} = \alpha_0 + \frac{N}{2}, \quad d = 1, \ldots, D_{L+1},$$

$$\beta_{l,d} = \beta_0^h + \frac{1}{2}\sum_n^N \mathbb{E}\left[(a_{n,l,d} - \gamma_{n,l,d} \odot z_{n,l,d})^2\right], \quad d = 1, \ldots, D_l, \quad l = 1, \ldots, L,$$

$$\beta_{L+1,d} = \beta_0 + \frac{1}{2}\sum_n^N \mathbb{E}\left[(y_{n,d} - z_{n,L+1,d})^2\right], \quad d = 1, \ldots, D_{L+1}.$$

For the parameters $\beta_{l,d}$, we must compute the sum of squares terms. For the last layer $l = L + 1$, this term, for each data point n, is given by:

$$\mathbb{E}\left[(y_{n,d} - z_{n,L+1,d})^2\right] = \sum_n^N \left(y_{n,d} - \mathbb{E}\left[\mathbf{W}_{L+1,d}\right]\mathbb{E}\left[\mathbf{a}_{n,L}\right] - \mathbb{E}\left[b_{L+1,d}\right]\right)^2$$

$$+ \sum_n^N \mathbb{E}\left[b_{L+1,d}^2\right] - \mathbb{E}\left[b_{L+1,d}\right]^2 + 2\mathbb{E}\left[b_{L+1,d}\mathbf{W}_{L+1,d}\right]\mathbb{E}\left[\mathbf{a}_{n,L}\right] - 2\mathbb{E}\left[b_{L+1,d}\right]\mathbb{E}\left[\mathbf{W}_{L+1,d}\right]\mathbb{E}\left[\mathbf{a}_{n,L}\right]$$

$$+ 2\sum_n^N \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{W}_{L+1,d}^T \mathbf{W}_{L+1,d}\right]\mathbb{E}\left[\mathbf{a}_{n,L}\mathbf{a}_{n,L}^T\right]\right) - \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{W}_{L+1,d}^T\right]\mathbb{E}\left[\mathbf{W}_{L+1,d}\right]\mathbb{E}\left[\mathbf{a}_{n,L}\right]\mathbb{E}\left[\mathbf{a}_{n,L}^T\right]\right).$$

Instead, for an intermediate layer $l = 1, \ldots, L$, the sum of squares term, for each data point n, is given by:

$$\mathbb{E}\left[(a_{n,l,d} - \gamma_{n,l,d} \cdot z_{n,l,d})^2\right] = \sum_n^N \left(\mathbb{E}\left[a_{n,l,d}\right] - \mathbb{E}\left[\gamma_{n,l,d}\right]\mathbb{E}\left[b_{l,d}\right] - \mathbb{E}\left[\gamma_{n,l,d}\right]\mathbb{E}\left[\mathbf{W}_{l,d}\right]\mathbb{E}\left[\mathbf{a}_{n,l-1}\right]\right)^2$$

$$+ \sum_n^N \mathbb{E}\left[a_{n,l,d}^2\right] - \mathbb{E}\left[a_{n,l,d}\right]^2 + \mathbb{E}\left[\gamma_{n,l,d}\right]\mathbb{E}\left[b_{l,d}^2\right] - \mathbb{E}\left[\gamma_{n,l,d}\right]^2 \mathbb{E}\left[b_{l,d}\right]^2$$

$$+ \sum_n^N \mathbb{E}\left[\gamma_{n,l,d}\right] \operatorname{Tr}\left(\mathbb{E}\left[\mathbf{W}_{l,d}^T \mathbf{W}_{l,d}\right] \mathbb{E}\left[\mathbf{a}_{n,l-1}\mathbf{a}_{n,l-1}^T\right]\right) - \mathbb{E}\left[\gamma_{n,l,d}\right]^2 \operatorname{Tr}\left(\mathbb{E}\left[\mathbf{W}_{l,d}^T\right]\mathbb{E}\left[\mathbf{W}_{l,d}\right]\left[\mathbf{a}_{n,l-1}\mathbf{a}_{n,l-1}^T\right]\right)$$

$$+ 2\sum_n^N \mathbb{E}\left[\gamma_{n,l,d}\right]\mathbb{E}\left[b_{l,d}\mathbf{W}_{l,d}\right]\left[\mathbf{a}_{n,l-1}\right] - \mathbb{E}\left[\gamma_{n,l,d}\right]^2 \mathbb{E}\left[b_{l,d}\right]\mathbb{E}\left[\mathbf{W}_{l,d}\right]\left[\mathbf{a}_{n,l-1}\right].$$

Weights and biases. The variational posterior for the weights and biases is:

$$q(\mathbf{b}, \mathbf{W}) \propto \exp\left(\mathbb{E}\left[\log \prod_n^N \mathrm{N}\left(y_n \mid \mathbf{W}_{L+1}\mathbf{a}_{n,L} + \mathbf{b}_{L+1}, \boldsymbol{\Sigma}_{L+1}\right) \prod_n^N \prod_l^L \mathrm{N}\left(\mathbf{a}_{n,l} \mid \boldsymbol{\gamma}_{n,l} \odot (\mathbf{W}_l \mathbf{a}_{n,l-1} + \mathbf{b}_l), \boldsymbol{\Sigma}_l\right)\right]\right)$$

$$\times \exp\left(\mathbb{E}\left[\log \prod_n^N \prod_l^L \prod_d^{D_l} \exp\left(\frac{(\gamma_{n,l,d} - \frac{1}{2})z_{n,l,d}}{T}\right) \exp\left(-\frac{\omega_{n,l,d}z_{n,l,d}^2}{2T^2}\right)\right]\right)$$

$$\times \exp\left(\mathbb{E}\left[\log \prod_l^{L+1} \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \mathrm{N}\left(W_{l,d,d'} \mid 0, \tau_l \psi_{l,d,d'}\right)\right]\right) \prod_l^{L+1} \prod_d^{D_l} \mathrm{N}(b_{l,d} \mid 0, s_0^2)$$

$$\propto \prod_n^N \exp\left(\mathbb{E}\left[\log \frac{1}{\sqrt{|\boldsymbol{\Sigma}_{L+1}|}}\right]\right) \exp\left(\mathbb{E}\left[-\frac{1}{2}\left(\mathbf{y}_n - \mathbf{W}_{L+1}\mathbf{a}_{n,L} - \mathbf{b}_{L+1}\right)^T \boldsymbol{\Sigma}_{L+1}^{-1}\left(\mathbf{y}_n - \mathbf{W}_{L+1}\mathbf{a}_{n,L} - \mathbf{b}_{L+1}\right)\right]\right)$$

$$\times \prod_l^L \prod_n^N \exp\left(\mathbb{E}\left[\log \frac{1}{\sqrt{|\boldsymbol{\Sigma}_l|}}\right]\right) \exp\left(\mathbb{E}\left[\left(-\frac{1}{2}\left(\mathbf{a}_{n,l} - \boldsymbol{\gamma}_{n,l}\mathbf{W}_l\mathbf{a}_{n,l-1} - \boldsymbol{\gamma}_{n,l}\mathbf{b}_l\right)^T \boldsymbol{\Sigma}_l^{-1}\left(\mathbf{a}_{n,l} - \boldsymbol{\gamma}_{n,l}\mathbf{W}_l\mathbf{a}_{n,l-1} - \boldsymbol{\gamma}_{n,l}\mathbf{b}_l\right)\right)\right]\right)$$

$$\times \prod_l^L \prod_n^N \prod_d^{D_l} \exp\left(\mathbb{E}\left[\frac{(\gamma_{n,l,d} - \frac{1}{2})\left(\mathbf{W}_{l,d}\mathbf{a}_{n,l-1} + b_{l,d}\right)}{T}\right]\right) \exp\left(\mathbb{E}\left[-\frac{\omega_{n,l,d}\left(\mathbf{W}_{l,d}\mathbf{a}_{n,l-1} + b_{l,d}\right)^2}{2T^2}\right]\right)$$

$$\times \prod_l^{L+1} \prod_d^{D_l} \prod_{d'}^{D_{l-1}} \exp\left(-\frac{W_{l,d,d'}^2}{2}\mathbb{E}\left[\frac{1}{\tau_l}\right]\mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right]\right) \prod_l^{L+1} \prod_d^{D_l} \exp\left(-\frac{b_{l,d}^2}{2s_0^2}\right).$$

Therefore, using also the fact that $\boldsymbol{\Sigma}_l$ is diagonal, we have that the variational posterior factorizes as $q(\mathbf{b}, \mathbf{W}) = \prod_l^{L+1} \prod_{d=1}^{D_l} q(b_{l,d}, \mathbf{W}_{l,d})$. We consider the terms $q(b_{l,d}, \mathbf{W}_{l,d})$ for the intermediate layers $l = 1, \ldots, L$ and $q(b_{L+1,d}, \mathbf{W}_{L+1,d})$ for the last layer separately.

Starting with the last layer $L + 1$, we first introduce the matrix

$$\mathbf{D}_{L+1,d}^{-1} = \operatorname{diag}\left(s_0^{-2}, \mathbb{E}\left[\tau_{L+1}^{-1}\right]\mathbb{E}\left[\psi_{L+1,d,1}^{-1}\right], \ldots, \mathbb{E}\left[\tau_{L+1}^{-1}\right]\mathbb{E}\left[\psi_{L+1,d,D_L}^{-1}\right]\right).$$

Then, for the variational posterior of the weights and biases for the dth dimension of the final layer, we only need to consider the relevant terms:

$$q(b_{L+1,d}, \mathbf{W}_{L+1,d}) \propto \exp\left(-\frac{1}{2}\widetilde{\mathbf{W}}_{L+1,d}\mathbf{D}_{L+1,d}^{-1}\widetilde{\mathbf{W}}_{L+1,d}^T - \frac{1}{2}\mathbb{E}\left[(\eta_{L+1,d})^{-2}\right]\sum_n^N \mathbb{E}\left[\left(y_{n,d} - \widetilde{\mathbf{W}}_{L+1,d}\widetilde{\mathbf{a}}_{n,L}\right)^2\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\widetilde{\mathbf{W}}_{L+1,d}\mathbf{D}_{L+1,d}^{-1}\widetilde{\mathbf{W}}_{L+1,d}^T - \frac{\mathbb{E}\left[\eta_{L+1,d}^{-2}\right]}{2}\left(\widetilde{\mathbf{W}}_{L+1,d}\left(\sum_n^N \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L}\widetilde{\mathbf{a}}_{n,L}^T\right]\right)\widetilde{\mathbf{W}}_{L+1,d}^T - 2\widetilde{\mathbf{W}}_{L+1,d}\left(\sum_n^N y_n \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L}\right]\right)\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\widetilde{\mathbf{W}}_{L+1,d}\left(\mathbf{D}_{L+1,d}^{-1} + \mathbb{E}\left[\eta_{L+1,d}^{-2}\right]\sum_n^N \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L}\widetilde{\mathbf{a}}_{n,L}^T\right]\right)\widetilde{\mathbf{W}}_{L+1,d}^T - 2\mathbb{E}\left[\eta_{L+1,d}^{-2}\right]\widetilde{\mathbf{W}}_{L+1,d}\left(\sum_n^N y_n \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L}\right]\right)\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\widetilde{\mathbf{W}}_{L+1,d}\mathbf{B}_{L+1,d}^{-1}\widetilde{\mathbf{W}}_{L+1,d}^T - 2\widetilde{\mathbf{W}}_{L+1,d}\mathbf{B}_{L+1,d}^{-1}\mathbf{m}_{L+1,d}^T\right)\right),$$

where

$$\mathbf{B}_{L+1,d}^{-1} = \mathbf{D}_{L+1,d}^{-1} + \mathbb{E}\left[(\eta_{L+1,d})^{-2}\right]\sum_n^N \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L}\widetilde{\mathbf{a}}_{n,L}^T\right],$$

$$\mathbf{m}_{L+1,d}^T = \mathbf{B}_{L+1,d} \mathbb{E}\left[(\eta_{L+1,d})^{-2}\right] \left(\sum_n^N y_n \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L}\right]\right).$$

Thus, completing the square, we have that

$$q(b_{L+1,d}, \mathbf{W}_{L+1,d}) = \mathrm{N}\left(\widetilde{\mathbf{W}}_{L+1,d} \mid \mathbf{m}_{L+1,d}, \mathbf{B}_{L+1,d}\right).$$

Next, for the intermediate layers $l = 1, \ldots, L$, we can similarly obtain the variational posterior of the weights and biases $q(b_{l,d}, \mathbf{W}_{l,d})$ for dimensions $d = 1, \ldots, D_l$. We introduce the matrices

$$\mathbf{D}_{l,d}^{-1} = \mathrm{diag}\left(s_0^{-2}, \mathbb{E}\left[\tau_l^{-1}\right] \mathbb{E}\left[\psi_{l,d,1}^{-1}\right], \ldots, \mathbb{E}\left[\tau_l^{-1}\right] \mathbb{E}\left[\psi_{l,d,D_{l-1}}^{-1}\right]\right),$$

and consider the terms relevant to derive each $q(b_{l,d}, \mathbf{W}_{l,d})$ separately:

$$q(b_{l,d}, \mathbf{W}_{l,d}) \propto \exp\left(-\frac{1}{2}\widetilde{\mathbf{W}}_{l,d}\mathbf{D}_{l,d}^{-1}\widetilde{\mathbf{W}}_{l,d}^T - \frac{1}{2T^2}\widetilde{\mathbf{W}}_{l,d}\left(\sum_n^N \mathbb{E}\left[\omega_{n,l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\widetilde{\mathbf{a}}_{n,l-1}^T\right]\right)\widetilde{\mathbf{W}}_{l,d}^T\right.$$

$$-\frac{\mathbb{E}\left[\eta_{l,d}^{-2}\right]}{2}\widetilde{\mathbf{W}}_{l,d}\left(\sum_n^N \mathbb{E}\left[\gamma_{n,l,d}^2\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\widetilde{\mathbf{a}}_{n,l-1}^T\right]\right)\widetilde{\mathbf{W}}_{l,d}^T + \mathbb{E}\left[\eta_{l,d}^{-2}\right]\widetilde{\mathbf{W}}_{l,d}\left(\sum_n^N \mathbb{E}\left[\gamma_{n,l,d}\right] \mathbb{E}\left[a_{n,l,d}\mathbf{a}_{n,l-1}\right]\right)$$

$$\left.+\frac{1}{T}\widetilde{\mathbf{W}}_{l,d}\left(\sum_n^N \mathbb{E}\left[\gamma_{n,l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right]\right) - \frac{1}{2T}\widetilde{\mathbf{W}}_{l,d}\left(\sum_n^N \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right]\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\widetilde{\mathbf{W}}_{l,d}\mathbf{B}_{l,d}^{-1}\widetilde{\mathbf{W}}_{l,d}^T - 2\widetilde{\mathbf{W}}_{l,d}\mathbf{B}_{l,d}^{-1}\mathbf{m}_{l,d}^T\right)\right),$$

where

$$\mathbf{B}_{l,d}^{-1} = \mathbf{D}_{l,d}^{-1} + \sum_n^N \left(\left(\frac{1}{T^2}\mathbb{E}\left[\omega_{n,l,d}\right] + \mathbb{E}\left[(\eta_{l,d})^{-2}\right]\mathbb{E}\left[\gamma_{n,l,d}\right]\right)\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\widetilde{\mathbf{a}}_{n,l-1}^T\right]\right),$$

$$\mathbf{m}_{l,d}^T = \mathbf{B}_{l,d}\left(\sum_n^N \left(\mathbb{E}\left[(\eta_{l,d})^{-2}\right]\mathbb{E}\left[\gamma_{n,l,d}\right]\mathbb{E}\left[a_{n,l,d}\widetilde{\mathbf{a}}_{n,l-1}\right] + \frac{1}{T}\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right]\left(\mathbb{E}\left[\gamma_{n,l,d}\right] - \frac{1}{2}\right)\right)\right).$$

Again, completing the square, we obtain the Gaussian variational posterior

$$q(b_{l,d}, \mathbf{W}_{l,d}) = \mathrm{N}\left((b_{l,d}, \mathbf{W}_{l,d}) \mid \mathbf{m}_{l,d}, \mathbf{B}_{l,d}\right).$$

Augmented variables. The variational posterior of the augmented variables is

$$q(\boldsymbol{\omega}) \propto \exp\left(\mathbb{E}\left[\log \prod_n^N \prod_l^L \prod_d^{D_l} \exp\left(-\frac{\omega_{n,l,d}z_{n,l,d}^2}{2T^2}\right) p(\omega_{n,l,d})\right]\right)$$

$$\propto \prod_n^N \prod_l^L \prod_d^{D_l} \exp\left(\mathbb{E}\left[-\frac{\omega_{n,l,d}z_{n,l,d}^2}{2T^2}\right]\right) p(\omega_{n,l,d}).$$

Thus, they are independent across width, depth, and observations, with

$$q(\omega_{n,l,d}) = \mathrm{PG}(\omega_{n,l,d} \mid 1, \frac{1}{T}\sqrt{\mathbb{E}\left[z_{n,l,d}^2\right]})$$

$$= \mathrm{PG}(\omega_{n,l,d} \mid 1, A_{n,l,d}),$$

where

$$A_{n,l,d} = \frac{1}{T}\sqrt{\left(\mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T\widetilde{\mathbf{W}}_{l,d}\right]\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\widetilde{\mathbf{a}}_{n,l-1}^T\right]\right)\right)}.$$

Binary activation. The variational posterior of the binary activations is:

$$q(\boldsymbol{\gamma}) \propto \exp\left(\mathbb{E}\left[\log \prod_n^N \prod_l^L \mathrm{N}\left(\mathbf{a}_{n,l} \mid \boldsymbol{\gamma}_{n,l} \odot \mathbf{z}_{n,l}, \boldsymbol{\Sigma}_l\right) + \log\left(\prod_n^N \prod_l^L \prod_d^{D_l} \exp\left(\frac{\gamma_{n,l,d} z_{n,l,d}}{T}\right) \right) \right] \right)$$

$$\propto \prod_n^N \prod_l^L \prod_l^{D_l} \exp\left(-\frac{1}{2\eta_{l,d}^2} \mathbb{E}\left[(a_{n,l,d} - \gamma_{n,l,d}(\mathbf{W}_{l,d}\mathbf{a}_{n,l-1} + b_{l,d}))^2 \right] + \mathbb{E}\left[\frac{\gamma_{n,l,d}(\mathbf{W}_{l,d}\mathbf{a}_{n,l-1} + b_{l,d})}{T} \right] \right).$$

Therefore, the variational posterior $q(\boldsymbol{\gamma})$ factories across observations $n = 1, \ldots, N$, layers $l = 1, \ldots, L$, and dimensions of the layer $d = 1, \ldots, D_l$, with each factor $q(\gamma_{n,l,d})$ given by:

$$q(\gamma_{n,l,d}) \propto \exp\left(-\frac{1}{2} \mathbb{E}\left[\eta_{l,d}^{-2}\right] \left(\gamma_{n,l,d}^2 \mathbb{E}\left[\left(\widetilde{\mathbf{W}}_{l,d}\widetilde{\mathbf{a}}_{n,l-1}\right)^2 \right] - 2\gamma_{n,l,d} \mathbb{E}\left[a_{n,l,d}\widetilde{\mathbf{W}}_{l,d}\widetilde{\mathbf{a}}_{n,l-1} \right] \right) + \frac{1}{T}\gamma_{n,l,d}\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\widetilde{\mathbf{a}}_{n,l-1}\right] \right)$$

$$\propto \exp\left(\gamma_{n,l,d}\left(-\frac{\mathbb{E}\left[\eta_{l,d}^{-2}\right]}{2} \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\widetilde{\mathbf{a}}_{n,l-1}^T\right] \right) + \mathbb{E}\left[\eta_{l,d}^{-2}\right] \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}a_{n,l,d}\right] + \frac{1}{T}\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right] \right) \right)$$

$$\propto \exp\left(\gamma_{n,l,d}\sigma^{-1}\left(\rho_{n,l,d}\right) \right),$$

where σ is the logistic function and

$$\rho_{n,l,d} = \sigma\left(\mathbb{E}\left[\eta_{l,d}^{-2}\right] \left(-\frac{1}{2}\mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\widetilde{\mathbf{a}}_{n,l-1}^T\right] \right) + \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}a_{n,l,d}\right] \right) + \frac{1}{T}\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right] \right).$$

Then noticing that $\sigma^{-1}(\rho) = \log(\rho(1-\rho)^{-1})$ and combining separate factors of the variational posterior of the binary activations, we obtain:

$$q(\boldsymbol{\gamma}) \propto \prod_n^N \prod_l^L \prod_d^{D_l} \rho_{n,l,d}^{\gamma_{n,l,d}} \left(1 - \rho_{n,l,d}\right)^{1-\gamma_{n,l,d}}$$

$$\propto \prod_n^N \prod_l^L \prod_d^{D_l} \mathrm{Bern}\left(\gamma_{n,l,d} \mid \rho_{n,l,d}\right).$$

Stochastic activation. The variational posterior of the stochastic activation is

$$q(\mathbf{a}) \propto \exp\left(\mathbb{E}\left[\log \prod_n^N \mathrm{N}\left(\mathbf{y}_n \mid \mathbf{z}_{n,L+1}, \boldsymbol{\eta}_{L+1}\right) + \log \prod_n^N \prod_l^L \mathrm{N}\left(\mathbf{a}_{n,l} \mid \boldsymbol{\gamma}_{n,l} \odot \mathbf{z}_{n,l}, \boldsymbol{\eta}_l\right) \right. \right.$$

$$\left. \left. + \log \prod_n^N \prod_l^L \prod_d^{D_l} \exp\left(\frac{(\gamma_{n,l,d} - \frac{1}{2})z_{n,l,d}}{T}\right) \exp\left(-\frac{\omega_{n,l,d} z_{n,l,d}^2}{2T^2}\right) \right] \right)$$

$$\propto \prod_n^N \exp\left(-\frac{1}{2} \sum_d^{D_{L+1}} \mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right] \mathbb{E}\left[(y_{n,d} - \mathbf{W}_{L+1,d}\mathbf{a}_{n,L} - b_{L+1,d})^2 \right] \right)$$

$$\times \prod_n^N \exp\left(-\frac{1}{2} \sum_l^L \sum_d^{D_l} \mathbb{E}\left[\frac{1}{\eta_{l,d}^2}\right] \mathbb{E}\left[(a_{n,l,d} - \gamma_{n,l,d}(\mathbf{W}_{l,d}\mathbf{a}_{n,l-1} + b_{l,d}))^2 \right] \right)$$

$$\times \prod_n^N \exp\left(\sum_l^L \sum_d^{D_l} \mathbb{E}\left[\frac{(\gamma_{n,l,d} - \frac{1}{2})z_{n,l,d}}{T} - \frac{\omega_{n,l,d} z_{n,l,d}^2}{2T^2} \right] \right).$$

Therefore, the variational posterior of the stochastic activations factories across observations $n = 1, \ldots, N$ and we derive $q(\mathbf{a}_n)$ separately. For each layer $l = 1, \ldots, L$, we introduce the following diagonal matrix $\hat{\boldsymbol{\Sigma}}_l^{-1} = \mathrm{diag}\left(\mathbb{E}\left[\eta_{l,1}^{-2}\right], \ldots, \mathbb{E}\left[\eta_{l,D_l}^{-2}\right] \right)$ and consider the relevant terms of the variational posterior:

$$q(\mathbf{a}_n) \propto \exp\left(-\frac{1}{2}\mathbf{a}_{n,L}^T \left(\sum_d^{D_{L+1}} \mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right] \mathbb{E}\left[\mathbf{W}_{L+1,d}^T \mathbf{W}_{L+1,d}\right] \mathbf{a}_{n,L} \right) \right)$$

$$\times \exp\left(-\mathbf{a}_{n,L}^T \left(\sum_d^{D_{L+1}} \mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right] \left(\mathbb{E}\left[\mathbf{W}_{L+1,d}^T b_{L+1,d}\right] - \mathbb{E}\left[\mathbf{W}_{L+1,d}^T\right] y_{n,d}\right)\right)\right)$$

$$\times \exp\left(-\frac{1}{2}\left(\mathbf{a}_{n,L}^T \hat{\mathbf{\Sigma}}_L^{-1} \mathbf{a}_{n,L} - 2\mathbf{a}_{n,L}^T \hat{\mathbf{\Sigma}}_L^{-1} \left(\left(\mathbb{E}\left[\boldsymbol{\gamma}_{n,L}\right] \mathbf{1}_{D_{L-1}}^T \odot \mathbb{E}\left[\mathbf{W}_L\right]\right) \mathbf{a}_{n,L-1} + \mathbb{E}\left[\boldsymbol{\gamma}_{n,L}\right] \odot \mathbb{E}\left[\mathbf{b}_L\right]\right)\right)\right)$$

$$\times \prod_{l=1}^{L-1} \exp\left(-\frac{1}{2}\left(\mathbf{a}_{n,l}^T \hat{\mathbf{\Sigma}}_l^{-1} \mathbf{a}_{n,l} - 2\mathbf{a}_{n,l}^T \hat{\mathbf{\Sigma}}_l^{-1} \left(\left(\mathbb{E}\left[\boldsymbol{\gamma}_{n,l}\right] \mathbf{1}_{D_{l-1}}^T \odot \mathbb{E}\left[\mathbf{W}_l\right]\right) \mathbf{a}_{n,l-1} + \mathbb{E}\left[\boldsymbol{\gamma}_{n,l}\right] \odot \mathbb{E}\left[\mathbf{b}_l\right]\right)\right)\right)$$

$$\times \prod_{l=1}^{L} \exp\left(-\frac{1}{2}\left(\mathbf{a}_{n,l-1}^T \left(\sum_{d=1}^{D_l} \mathbb{E}\left[\frac{1}{\eta_{l,d}^2}\right] \mathbb{E}\left[\gamma_{n,l,d}\right] \mathbb{E}\left[\mathbf{W}_{l,d}^T \mathbf{W}_{l,d}\right] \mathbf{a}_{n,l-1}\right)\right)\right) \times$$

$$\times \prod_{l=1}^{L} \exp\left(-\mathbf{a}_{n,l-1}^T \left(\sum_{d=1}^{D_l} \mathbb{E}\left[\frac{1}{\eta_{l,d}^2}\right] \mathbb{E}\left[\gamma_{n,l,d}\right] \mathbb{E}\left[\mathbf{W}_{l,d}^T \mathbf{b}_{l,d}\right]\right)\right)$$

$$\times \prod_{l=1}^{L} \exp\left(-\frac{1}{2}\left(\mathbf{a}_{n,l-1}^T \left(\frac{1}{T^2}\sum_{d=1}^{D_l} \mathbb{E}\left[\omega_{n,l,d}\right] \mathbb{E}\left[\mathbf{W}_{l,d}^T \mathbf{W}_{l,d}\right]\right) \mathbf{a}_{n,l-1}\right)\right)$$

$$\times \prod_{l=1}^{L} \exp\left(\mathbf{a}_{n,l-1}^T \left(\frac{1}{T}\sum_{d=1}^{D_l} \mathbb{E}\left[\mathbf{W}_{l,d}^T\right]\left(\mathbb{E}\left[\gamma_{n,l,d}\right] - \frac{1}{2}\right) - \frac{1}{T^2}\sum_{d=1}^{D_l} \mathbb{E}\left[\omega_{n,l,d}\right] \mathbb{E}\left[\mathbf{W}_{l,d}^T \mathbf{b}_{l,d}\right]\right)\right).$$

The variational posterior of the stochastic activations does not factories into independent blocks, however it does have a structured sequential factorization $q(\mathbf{a}_n) = \prod_{l=1}^{L} q(\mathbf{a}_{n,l} \mid \mathbf{a}_{n,l-1})$. And, we can derive the variational factor $q(\mathbf{a}_{n,L} \mid \mathbf{a}_{n,L-1})$ by only considering the terms with $\mathbf{a}_{n,L}$. First, introduce the matrices $\mathbf{S}_{n,L}$ and $\mathbf{M}_{n,L}$ and a vectors $\mathbf{t}_{n,L}$:

$$\mathbf{S}_{n,L}^{-1} = \hat{\mathbf{\Sigma}}_L^{-1} + \sum_{d=1}^{D_{L+1}} \mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right] \mathbb{E}\left[\mathbf{W}_{L+1,d}^T \mathbf{W}_{L+1,d}\right],$$

$$\mathbf{t}_{n,L} = \mathbf{S}_{n,L}\left(\left(\sum_{d=1}^{D_{l+1}} \mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right]\left(-\mathbb{E}\left[\mathbf{W}_{L+1,d}^T b_{L+1,d}\right] + \mathbb{E}\left[\mathbf{W}_{L+1,d}^T\right] y_{n,d}\right)\right) + \hat{\mathbf{\Sigma}}_l^{-1}\mathbb{E}\left[\boldsymbol{\gamma}_{n,L}\right] \odot \mathbb{E}\left[\mathbf{b}_L\right]\right),$$

$$\mathbf{M}_{n,L} = \mathbf{S}_{n,L}\hat{\mathbf{\Sigma}}_L^{-1}\mathbb{E}\left[\boldsymbol{\gamma}_{n,L}\right] \mathbf{1}_{D_{L-1}}^T \odot \mathbb{E}\left[\mathbf{W}_L\right].$$

Then we consider relevant terms of the variational posterior:

$$q(\mathbf{a}_{n,L} \mid \mathbf{a}_{n,L-1}) \propto \exp\left(-\frac{1}{2}\left(\mathbf{a}_{n,L}^T \mathbf{S}_{n,L}^{-1} \mathbf{a}_{n,L} - 2\mathbf{a}_{n,L}^T \mathbf{S}_{n,L}^{-1}\left(\mathbf{t}_{n,L} + \mathbf{M}_{n,L}\mathbf{a}_{n,L-1}\right)\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\mathbf{a}_{n,L} - \left(\mathbf{t}_{n,L} + \mathbf{M}_{n,L}\mathbf{a}_{n,L-1}\right)\right)^T \mathbf{S}_{n,L}^{-1}\left(\mathbf{a}_{n,L} - \left(\mathbf{t}_{n,L} + \mathbf{M}_{n,L}\mathbf{a}_{n,L-1}\right)\right)\right) \times$$

$$\times \exp\left(\frac{1}{2}\left(\mathbf{t}_{n,L} + \mathbf{M}_{n,L}\mathbf{a}_{n,L-1}\right)^T \mathbf{S}_{n,L}^{-1}\left(\mathbf{t}_{n,L} + \mathbf{M}_{n,L}\mathbf{a}_{n,L-1}\right)\right)$$

$$\propto \mathrm{N}\left(\mathbf{a}_{n,L} \mid \mathbf{t}_{n,L} + \mathbf{M}_{n,L}\mathbf{a}_{n,L-1}, \mathbf{S}_{n,L}\right) \times \exp\left(\frac{1}{2}\left(\mathbf{t}_{n,L} + \mathbf{M}_{n,L}\mathbf{a}_{n,L-1}\right)^T \mathbf{S}_{n,L}^{-1}\left(\mathbf{t}_{n,L} + \mathbf{M}_{n,L}\mathbf{a}_{n,L-1}\right)\right),$$

where the first term in the equation above provides $q(\mathbf{a}_{n,L} \mid \mathbf{a}_{n,L-1})$ and the second terms is relevant for computing the subsequent $q(\mathbf{a}_{n,L-1} \mid \mathbf{a}_{n,L-2})$. Recursively repeating a similar procedure for $l = L-1, \ldots, 1$, we are then able to obtain each of the variational posteriors $q(\mathbf{a}_{n,l} \mid \mathbf{a}_{n,l-1})$. Each time we define $\mathbf{S}_{n,l}, \mathbf{M}_{n,l}$ and $\mathbf{t}_{n,l}$ as follows:

$$\mathbf{S}_{n,l}^{-1} = \hat{\mathbf{\Sigma}}_l^{-1} - \mathbf{M}_{n,l+1}^T \mathbf{S}_{n,l+1}^{-1} \hat{\mathbf{M}}_{n,l+1} + \sum_{d=1}^{D_{l+1}}\left(\mathbb{E}\left[\frac{1}{\eta_{l+1,d}^2}\right] \mathbb{E}\left[\gamma_{n,l+1,d}\right] + \frac{1}{T^2}\sum_{d=1}^{D_{l+1}} \mathbb{E}\left[\omega_{n,l+1,d}\right]\right) \mathbb{E}\left[\mathbf{W}_{l+1,d}^T \mathbf{W}_{l+1,d}\right]$$

$$\mathbf{t}_{n,l} = \mathbf{S}_{n,l}\left(\mathbf{M}_{n,l+1}^T \mathbf{S}_{n,l+1}^{-1}\mathbf{t}_{n,l+1} + \hat{\mathbf{\Sigma}}_l^{-1}\mathbb{E}\left[\boldsymbol{\gamma}_{n,l}\right] \odot \mathbb{E}\left[\mathbf{b}_l\right] + \frac{1}{T}\sum_{d=1}^{D_{l+1}} \mathbb{E}\left[\mathbf{W}_{l+1,d}^T\right]\left(\mathbb{E}\left[\gamma_{n,l+1,d}\right] - \frac{1}{2}\right)\right.$$

31

$$-\sum_{d=1}^{D_{l+1}}\left(\mathbb{E}\left[\frac{1}{\eta_{l+1,d}^2}\right]\mathbb{E}\left[\gamma_{n,l+1,d}\right]+\frac{1}{T^2}\mathbb{E}\left[\omega_{n,l+1,d}\right]\right)\mathbb{E}\left[\mathbf{W}_{l+1,d}^T b_{l+1,d}\right]\right),$$

$$\mathbf{M}_{n,l}=\mathbf{S}_{n,l}\hat{\mathbf{\Sigma}}_l^{-1}\mathbb{E}\left[\boldsymbol{\gamma}_{n,l}\right]\mathbf{1}_{D_{l-1}}^T\odot\mathbb{E}\left[\mathbf{W}_l\right].$$

Then substituting the above into the terms of the variational posterior containing $\mathbf{a}_{n,l}$:

$$q(\mathbf{a}_{n,l}\mid\mathbf{a}_{n,l-1})\propto\exp\left(-\frac{1}{2}\left(\mathbf{a}_{n,l}-(\mathbf{t}_{n,l}+\mathbf{M}_{n,l}\mathbf{a}_{n,l-1})\right)^T\mathbf{S}_{n,l}^{-1}\left(\mathbf{a}_{n,l}-(\mathbf{t}_{n,l}+\mathbf{M}_{n,l}\mathbf{a}_{n,l-1})\right)\right)$$

$$\times\exp\left(\frac{1}{2}\left(\mathbf{t}_{n,l}+\mathbf{M}_{n,l}\mathbf{a}_{n,l-1}\right)^T\mathbf{S}_{n,l}^{-1}\left(\mathbf{t}_{n,l}+\mathbf{M}_{n,l}\mathbf{a}_{n,l-1}\right)\right)$$

$$\propto\mathrm{N}\left(\mathbf{a}_{n,l}\mid\mathbf{t}_{n,l}+\mathbf{M}_{n,l}\mathbf{a}_{n,l-1},\mathbf{S}_{n,l}\right)\times\exp\left(\frac{1}{2}\left(\mathbf{t}_{n,l}+\mathbf{M}_{n,l}\mathbf{a}_{n,l-1}\right)^T\mathbf{S}_{n,l}^{-1}\left(\mathbf{t}_{n,l}+\mathbf{M}_{n,l}\mathbf{a}_{n,l-1}\right)\right).$$

Finally, we combine the terms $q(\mathbf{a}_{n,l}\mid\mathbf{a}_{n,l-1})$ for $l=1,\ldots,L+1$ and get the variational posterior of the stochastic activation

$$q(\mathbf{a})\propto\prod_{n=1}^{N}\prod_{l=1}^{L}\mathrm{N}\left(\mathbf{a}_{n,l}\mid\mathbf{t}_{n,l}+\mathbf{M}_{n,l}\mathbf{a}_{n,l-1},\mathbf{S}_{n,l}\right).$$

B ELBO computation

B.1 ELBO for training

Recall, that optimal variational parameters maximize the ELBO function of Equation (10) which for our model is:

$$\mathrm{ELBO}=\mathbb{E}\left[\log p(\mathbf{y},\mathbf{a},\boldsymbol{\gamma},\boldsymbol{\omega}|\mathbf{W},\mathbf{b},\mathbf{\Sigma})\right]+\mathbb{E}\left[\log p(\mathbf{W}|\boldsymbol{\psi},\boldsymbol{\tau})\right]+\mathbb{E}\left[\log p(\boldsymbol{\psi})\right]+\mathbb{E}\left[\log p(\boldsymbol{\tau})\right]$$

$$+\mathbb{E}\left[\log p(\mathbf{b})\right]+\mathbb{E}\left[\log p(\mathbf{\Sigma})\right]-\mathbb{E}\left[\log q(\mathbf{a})\right]-\mathbb{E}\left[\log q(\boldsymbol{\gamma})\right]-\mathbb{E}\left[\log q(\boldsymbol{\omega})\right]$$

$$-\mathbb{E}\left[\log q(\mathbf{W},\mathbf{b})\right]-\mathbb{E}\left[\log q(\boldsymbol{\eta})\right]-\mathbb{E}\left[\log q(\boldsymbol{\psi})\right]-\mathbb{E}\left[\log q(\boldsymbol{\tau})\right].$$

Similar to the variational update, we compute the terms of the ELBO corresponding to different blocks of parameters separately.

ELBO of $\boldsymbol{\tau}$. First, consider the terms of the ELBO containing the global shrinkage parameters:

$$\mathbb{E}\left[\log p(\boldsymbol{\tau})-\log q(\boldsymbol{\tau})\right]=\sum_{l=1}^{L+1}\mathbb{E}\left[\log\mathrm{GIG}\left(\tau_l\mid\nu_{\mathrm{glob}},\delta_{\mathrm{glob}},\lambda_{\mathrm{glob}}\right)-\log\mathrm{GIG}\left(\tau_l|\hat{\nu}_{\mathrm{glob},l},\hat{\delta}_{\mathrm{glob},l},\lambda_{\mathrm{glob}}\right)\right]$$

$$=C_\tau+\sum_{l=1}^{L+1}\mathbb{E}\left[\log\tau_l^{\nu_{\mathrm{glob}}-1}\exp\left(-\frac{1}{2}\left(\frac{\delta_{\mathrm{glob}}^2}{\tau_l}+\lambda_{\mathrm{glob}}^2\tau_l\right)\right)\right]-\mathbb{E}\left[\log\left(\tau_l^{\hat{\nu}_{\mathrm{glob},l}-1}\right)\exp\left(-\frac{1}{2}\left(\frac{\hat{\delta}_{\mathrm{glob},l}^2}{\tau_l}+\lambda_{\mathrm{glob}}^2\tau_l\right)\right)\right]$$

$$=C_\tau+\frac{1}{2}\sum_{l=1}^{L+1}D_lD_{l-1}\mathbb{E}\left[\log\tau_l\right]+\mathbb{E}\left[\frac{1}{\tau_l}(\hat{\delta}_{\mathrm{glob},l}^2-\delta_{\mathrm{glob}}^2)\right],$$

where the normalizing constant is

$$C_\tau=\sum_{l=1}^{L+1}(\nu_{\mathrm{glob}}-\hat{\nu}_{\mathrm{glob},l})\log(\lambda_{\mathrm{glob}})+\hat{\nu}_{\mathrm{glob},l}\log(\hat{\delta}_{\mathrm{glob},l})-\nu_{\mathrm{glob}}\log(\delta_{\mathrm{glob}})$$

$$+\sum_{l=1}^{L+1}\log(K_{\hat{\nu}_{\mathrm{glob},l}}(\lambda_{\mathrm{glob}}\hat{\delta}_{\mathrm{glob},l}))-\log(K_{\nu_{\mathrm{glob}}}(\lambda\delta_{\mathrm{glob}})).$$

ELBO of ψ. Similarly, the terms of the ELBO containing the local shrinkage parameters are

$$\mathbb{E}\left[\log p(\psi) - \log q(\psi)\right] = C_\psi + \sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\sum_{d'=1}^{D_{l-1}} \mathbb{E}\left[\log \text{GIG}\left(\psi_{l,d,d'} \mid \nu_{\text{loc},l}, \delta_{\text{loc},l}, \lambda_{\text{loc},l}\right)\right]$$

$$- \sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\sum_{d'=1}^{D_{l-1}} \mathbb{E}\left[\log \text{GIG}\left(\psi_{l,d,d'}|\hat{\nu}_{\text{loc},l,d,d'}, \hat{\delta}_{\text{loc},l,d,d'}, \lambda_{\text{loc},l}\right)\right]$$

$$= C_\psi + \sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\sum_{d'=1}^{D_{l-1}} \mathbb{E}\left[\log \psi_{l,d,d'}^{\nu_{\text{loc},l}-1} \exp\left(-\frac{1}{2}\left(\frac{\delta_{\text{loc},l}^2}{\psi_{l,d,d'}} + \lambda_{\text{loc},l}^2 \psi_{l,d,d'}\right)\right)\right]$$

$$- \sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\sum_{d'=1}^{D_{l-1}} \mathbb{E}\left[\log\left(\psi_{l,d,d'}^{\hat{\nu}_{\text{loc},l,d,d'}-1}\right) \exp\left(-\frac{1}{2}\left(\frac{\hat{\delta}_{\text{loc},l,d,d'}^2}{\psi_{l,d,d'}} + \lambda_{\text{loc},l}^2 \psi_{l,d,d'}\right)\right)\right]$$

$$= C_\psi + \frac{1}{2}\sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\sum_{d'=1}^{D_{l-1}} \mathbb{E}\left[\log \psi_{l,d,d'}\right] + \mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right]\left(\hat{\delta}_{\text{loc},l,d,d'}^2 - \delta_{\text{loc},l}^2\right),$$

where the normalizing constant is

$$C_\psi = \sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\sum_{d'=1}^{D_{l-1}} (\nu_{\text{loc},l} - \hat{\nu}_{\text{loc},l,d,d'})\log(\lambda_{\text{loc},l}) + \hat{\nu}_{\text{loc},l,d,d'}\log(\hat{\delta}_{\text{loc},l,d,d'}) - \nu_{\text{loc},l}\log(\delta_{\text{loc},l})$$

$$+ \sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\sum_{d'=1}^{D_{l-1}} \log(K_{\hat{\nu}_{\text{loc},l,d,d'}}(\lambda_{\text{glob}}\hat{\delta}_{\text{loc},l,d,d'})) - \log(K_{\nu_{\text{loc},l}}(\lambda_{\text{loc},l}\delta_{\text{loc},l})).$$

ELBO of η. As before, the covariance matrix matrix is assumed to be diagonal so that the relevant ELBO is:

$$\mathbb{E}\left[\log p(\Sigma) - \log q(\eta)\right]$$

$$= C_\eta + \sum_{l=1}^{L}\sum_{d=1}^{D_l} \mathbb{E}\left[\log \text{IG}(\eta_{l,d}^2|\alpha_0^h, \beta_0^h)\right] + \sum_{d=1}^{D_{L+1}} \mathbb{E}\left[\log \text{IG}(\eta_{l,d}^2|\alpha_0, \beta_0)\right] - \sum_{l}^{L+1}\sum_{d}^{D_l} \mathbb{E}\left[\log \text{IG}(\eta_{l,d}^2 \mid \alpha_{l,d}, \beta_{l,d})\right]$$

$$= C_\eta + \sum_{l}^{L}\sum_{d}^{D_l} \left(\alpha_{l,d} - \alpha_0^h\right)\mathbb{E}\left[\log \eta_{l,d}^2\right] + \sum_{d=1}^{D_{L+1}} (\alpha_{L+1,d} - \alpha_0)\mathbb{E}\left[\log \eta_{L+1,d}^2\right]$$

$$+ \sum_{l=1}^{L}\sum_{d=1}^{D_l} \left(\beta_{l,d} - \beta_0^h\right)\mathbb{E}\left[\frac{1}{\eta_{l,d}^2}\right] + \sum_{d=1}^{D_{L+1}} (\beta_{L+1,d} - \beta_0)\mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right]$$

$$= C_\eta + \frac{N}{2}\sum_{l}^{L+1}\sum_{d}^{D_l} \mathbb{E}\left[\log \eta_{l,d}^2\right] + \sum_{l=1}^{L}\sum_{d=1}^{D_l} \left(\beta_{l,d} - \beta_0^h\right)\mathbb{E}\left[\frac{1}{\eta_{l,d}^2}\right] + \sum_{d=1}^{D_{L+1}} (\beta_{L+1,d} - \beta_0)\mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right],$$

where the normalizing constant is

$$C_\eta = \sum_{l=1}^{L}\sum_{d=1}^{D_l} \alpha_0^h \log \beta_0^h - \alpha_{l,d}\log \beta_{l,d} + \log\Gamma(\alpha_{l,d}) - \log\Gamma(\alpha_0^h)$$

$$+ \sum_{d=1}^{D_{L+1}} \alpha_0 \log \beta_0 - \alpha_{L+1,d}\log \beta_{L+1,d} + \log\Gamma(\alpha_{L+1,d}) - \log\Gamma(\alpha_0).$$

ELBO of $(\mathbf{W}, \mathbf{b})$.

Recall, previously introduced matrices $\mathbf{D}_{l,d} = \text{diag}\left(s_0^{-2}, \mathbb{E}\left[\tau_l^{-1}\right]\mathbb{E}\left[\psi_{l,d,1}^{-1}\right], \ldots, \mathbb{E}\left[\tau_l^{-1}\right]\mathbb{E}\left[\psi_{l,d,D_{l-1}}^{-1}\right]\right)$ and denote further $\mathbf{D}_{l,d}^0 = \text{diag}\left(s_0^2, \tau_l\psi_{l,d,1}, \ldots, \tau_l\psi_{l,d,D_{l-1}}\right)$. Then the ELBO of weights and biases is:

$$\mathbb{E}\left[\log p(\mathbf{W}|\psi, \tau)\right] + \mathbb{E}\left[\log p(\mathbf{b})\right] - \mathbb{E}\left[\log q(\mathbf{W}, \mathbf{b})\right]$$

$$= \sum_{l=1}^{L+1} \sum_{d=1}^{D_l} \sum_{d'=1}^{D_{l-1}} \mathbb{E}\left[\log \mathrm{N}\left(\widetilde{\mathbf{W}}_{l,d} | 0, \mathbf{D}_{l,d}^0\right)\right] - \sum_{l}^{L+1} \sum_{d}^{D_l} \mathbb{E}\left[\log \mathrm{N}\left(\widetilde{\mathbf{W}}_{l,d} \mid \mathbf{m}_{l,d}, \mathbf{B}_{l,d}\right)\right]$$

$$= \sum_{l=1}^{L+1} \sum_{d=1}^{D_l} \mathbb{E}\left[\log\left(|\mathbf{D}_{l,d}|\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\widetilde{\mathbf{W}}_{l,d}\left(\mathbf{D}_{l,d}^0\right)^{-1}\widetilde{\mathbf{W}}_{l,d}^T\right)\right]$$

$$- \sum_{l}^{L+1} \sum_{d}^{D_l} \mathbb{E}\left[\log |\mathbf{B}_{l,d}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\widetilde{\mathbf{W}}_{l,d} - \mathbf{m}_{l,d}\right)\mathbf{B}_{l,d}^{-1}\left(\widetilde{\mathbf{W}}_{l,d} - \mathbf{m}_{l,d}\right)^T\right)\right]$$

$$= \frac{1}{2} \sum_{l}^{L+1} \sum_{d}^{D_l} \mathbb{E}\left[\log |\mathbf{B}_{l,d}|\right] - \mathbb{E}\left[\log\left(|\mathbf{D}_{l,d}^0|\right)\right] - \mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}(\mathbf{D}_{l,d}^0)^{-1}\widetilde{\mathbf{W}}_{l,d}^T\right] + \mathbb{E}\left[\left(\widetilde{\mathbf{W}}_{l,d} - \mathbf{m}_{l,d}\right)\mathbf{B}_{l,d}^{-1}\left(\widetilde{\mathbf{W}}_{l,d} - \mathbf{m}_{l,d}\right)^T\right]$$

$$= \frac{1}{2} \sum_{l}^{L+1} \sum_{d}^{D_l} \log |\mathbf{B}_{l,d}| - \mathbb{E}\left[\log\left(|\mathbf{D}_{l,d}^0|\right)\right] - \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d}\right] \mathbb{E}\left[(\mathbf{D}_{l,d}^0)^{-1}\right]\right) + \sum_{l}^{L+1} \frac{D_l}{2}$$

$$= \frac{1}{2} \sum_{l}^{L+1} \sum_{d}^{D_l} \left(\log |\mathbf{B}_{l,d}| - \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T \widetilde{\mathbf{W}}_{l,d}\right] \mathbf{D}_{l,d}\right) - \sum_{d'=1}^{D_{l-1}} \mathbb{E}\left[\log \psi_{l,d,d'}\right]\right) - \frac{1}{2} \sum_{l}^{L+1} D_l\left(\log s_0^2 + D_{l-1}\mathbb{E}\left[\log \tau_l\right] - 1\right).$$

ELBO of a, $\boldsymbol{\gamma}$ and $\boldsymbol{\omega}$. The remaining terms of the ELBO are the ones with stochastic and binary activations and additional augmented variables:

$$\mathbb{E}\left[\log p(\mathbf{y}, \mathbf{a}, \boldsymbol{\gamma}, \boldsymbol{\omega} | \mathbf{W}, \mathbf{b}, \boldsymbol{\Sigma})\right] - \mathbb{E}\left[\log q(\mathbf{a})\right] - \mathbb{E}\left[\log q(\boldsymbol{\gamma})\right] - \mathbb{E}\left[\log q(\boldsymbol{\omega})\right]$$

$$= \sum_{n=1}^{N} \sum_{d=1}^{D_{L+1}} \mathbb{E}\left[\log \mathrm{N}\left(y_{n,d} \mid \mathbf{z}_{n,L+1,d}, \boldsymbol{\Sigma}_{L+1,d}\right)\right] + \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E}\left[\log \mathrm{N}\left(a_{n,l,d} \mid \boldsymbol{\gamma}_{n,d} \odot \mathbf{z}_{n,l,d}, \boldsymbol{\Sigma}_{l,d}\right)\right]$$

$$+ \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E}\left[\log\left(\exp\left(\frac{\kappa_{n,l,d}z_{n,l,d}}{T}\right)\exp\left(-\frac{\omega_{n,l,d}z_{n,l,d}^2}{2T^2}\right)\mathrm{PG}(\omega_{n,l,d} \mid 1, 0)\right)\right]$$

$$- \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbb{E}\left[\log \mathrm{N}\left(\mathbf{a}_{n,l} \mid \mathbf{t}_{n,l} + \mathbf{M}_{n,l}\mathbf{a}_{n,l-1}, \mathbf{S}_{n,l}\right)\right]$$

$$- \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E}\left[\log \mathrm{Bern}\left(\gamma_{n,l,d} \mid \rho_{n,l,d}\right)\right] + \mathbb{E}\left[\log \mathrm{PG}(\omega_{n,l,d} \mid 1, A_{n,l,d})\right]$$

$$= \sum_{n=1}^{N} \sum_{d=1}^{D_{L+1}} \mathbb{E}\left[\log(\eta_{L+1,d}^2)^{-1/2}\exp\left(-\frac{1}{2\eta_{L+1,d}^2}\left(y_{n,d} - \mathbf{W}_{L+1,d}\mathbf{a}_{n,L} - b_{L+1,d}\right)^2\right)\right] - \frac{ND_{L+1}}{2}\log(2\pi)$$

$$+ \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E}\left[\log(\eta_{l,d}^2)^{-1/2}\exp\left(-\frac{1}{2\eta_{l,d}^2}\left(a_{n,l,d} - \gamma_{n,l,d} \odot \left(\mathbf{W}_{l,d}\mathbf{a}_{n,l-1} + b_{l,d}\right)\right)^2\right)\right] - N\sum_{l=1}^{L} D_l \log(2)$$

$$+ \frac{1}{T} \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E}\left[\left(\gamma_{n,d} - \frac{1}{2}\right)\left(\mathbf{W}_{l,d}\mathbf{a}_{n,l-1} + b_{l,d}\right)\right] - \frac{1}{2T^2} \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E}\left[\omega_{n,l,d}\left(\mathbf{W}_{l,d}\mathbf{a}_{n,l-1} + b_{l,d}\right)^2\right]$$

$$- \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbb{E}\left[\log |\mathbf{S}_{n,l}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\mathbf{a}_{n,l} - \mathbf{t}_{n,l} - \mathbf{M}_{n,l}\mathbf{a}_{n,l-1}\right)^T \mathbf{S}_{n,l}^{-1}\left(\mathbf{a}_{n,l} - \mathbf{t}_{n,l} - \mathbf{M}_{n,l}\mathbf{a}_{n,l-1}\right)\right)\right]$$

$$- \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \left(\rho_{n,l,d}\log\rho_{n,l,d} + (1-\rho_{n,l,d})\log(1-\rho_{n,l,d})\right) + \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E}\left[\log\frac{\mathrm{PG}(\omega_{n,l,d} \mid 1, 0)}{\mathrm{PG}(\omega_{n,l,d} \mid 1, A_{n,d})}\right]$$

$$= -\frac{1}{2} \sum_{n=1}^{N} \sum_{d=1}^{D_{L+1}} \mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right]\left(y_{n,d}^2 - 2y_{n,d}\mathbb{E}\left[\widetilde{\mathbf{W}}_{L+1,d}\right]\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L}\right] + \mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{L+1,d}^T\widetilde{\mathbf{W}}_{L+1,d}\right]\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,L}\widetilde{\mathbf{a}}_{n,L}^T\right]\right)\right)$$

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{d=1}^{D_l}\mathbb{E}\left[\frac{1}{\eta_{l,d}^2}\right]\left(\mathbb{E}\left[a_{n,l,d}^2\right]-2\mathbb{E}\left[\gamma_{n,l,d}\right]\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right]\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}a_{n,l,d}\right]\right)$$

$$-\frac{1}{2}\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{d=1}^{D_l}\mathbb{E}\left[\frac{1}{\eta_{l,d}^2}\right]\mathbb{E}\left[\gamma_{n,l,d}^2\right]\mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T\widetilde{\mathbf{W}}_{l,d}\right]\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\widetilde{\mathbf{a}}_{n,l-1}^T\right]\right)$$

$$-\frac{N}{2}\sum_{d=1}^{D_{L+1}}\mathbb{E}\left[\log\eta_{L+1,d}^2\right]-\frac{N}{2}\sum_{l=1}^{L}\sum_{d=1}^{D_l}\mathbb{E}\left[\log\eta_{l,d}^2\right]+\frac{1}{2}\sum_{n=1}^{N}\sum_{l=1}^{L}\log(|\mathbf{S}_{n,l}|)$$

$$+\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{d=1}^{D_l}\left(\frac{1}{T}\left(\rho_{n,l,d}-\frac{1}{2}\right)\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}\right]\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\right]-\frac{1}{2T^2}\mathbb{E}\left[\omega_{n,l,d}\right]\left(\mathrm{Tr}\left(\mathbb{E}\left[\widetilde{\mathbf{W}}_{l,d}^T\widetilde{\mathbf{W}}_{l,d}\right]\mathbb{E}\left[\widetilde{\mathbf{a}}_{n,l-1}\widetilde{\mathbf{a}}_{n,l-1}^T\right]\right)\right)\right)$$

$$-\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{d=1}^{D_l}\left(\rho_{n,l,d}\log\rho_{n,l,d}+(1-\rho_{n,l,d})\log(1-\rho_{n,l,d})-\frac{A_{n,l,d}^2}{2}\mathbb{E}\left[\omega_{n,l,d}\right]+\log(\cosh(\frac{A_{n,l,d}}{2}))\right)+C_a,$$

where the normalizing constant is

$$C_a=-\frac{ND_{L+1}}{2}\log(2\pi)-N\sum_{l=1}^{L}D_l\log(2).$$

Total ELBO   Then, we can sum the derived above parts to get the total ELBO of our model:

$$\mathrm{ELBO}=\mathrm{const.}+\sum_{l=1}^{L+1}\frac{1}{2}\mathbb{E}\left[\frac{1}{\tau_l}\right]\left(\hat{\delta}_{\mathrm{glob},l}^2-\delta_{\mathrm{glob}}^2\right)+(\hat{\nu}_{\mathrm{glob},l}\log(\hat{\delta}_{\mathrm{glob},l})+\log(K_{\hat{\nu}_{\mathrm{glob},l}}(\lambda_{\mathrm{glob}}\hat{\delta}_{\mathrm{glob},l}))$$

$$+\sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\sum_{d'=1}^{D_{l-1}}\frac{1}{2}\mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right]\left(\hat{\delta}_{\mathrm{loc},l,d,d'}^2-\delta_{\mathrm{loc},l}^2\right)+\hat{\nu}_{\mathrm{loc},l,d,d'}\log(\hat{\delta}_{\mathrm{loc},l,d,d'})+\log(K_{\hat{\nu}_{\mathrm{loc},l,d,d'}}(\lambda_{\mathrm{loc},l}\hat{\delta}_{\mathrm{loc},l,d,d'}))$$

$$+\sum_{l=1}^{L+1}\sum_{d=1}^{D_l}\mathbb{E}\left[\frac{1}{\eta_{l,d}^2}\right]\left(\beta_{l,d}-\beta_0^l\right)-\alpha_{l,d}\log\beta_{l,d}+\frac{1}{2}\log|\mathbf{B}_{l,d}|-\frac{1}{2}\left(\frac{1}{s_0^2}\mathbb{E}[b_{l,d}^2]+\sum_{d'=1}^{D_{l-1}}\mathbb{E}\left[\frac{1}{\tau_l}\right]\mathbb{E}\left[\frac{1}{\psi_{l,d,d'}}\right]\mathbb{E}[w_{l,d,d'}^2]\right)$$

$$+\frac{1}{2}\sum_{n=1}^{N}\sum_{l=1}^{L}\log(\mathbf{S}_{n,l})-\frac{1}{2}\sum_{d=1}^{D_y}\mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right]\left(\sum_{n=1}^{N}\mathbb{E}\left[\left(y_{n,d}-\mathbb{E}\left[\tilde{\mathbf{W}}_{L+1,d}\right]\mathbb{E}\left[\tilde{\mathbf{a}}_{n,L}\right]\right)^2\right]\right)$$

$$-\frac{1}{2}\sum_{d=1}^{D_y}\mathbb{E}\left[\frac{1}{\eta_{L+1,d}^2}\right]\left(\sum_{n=1}^{N}\mathrm{Tr}\left(\left(\mathbf{B}_{L+1,d}+\mathbf{m}_{L+1,d}\mathbf{m}_{L+1,d}^T\right)\mathbb{E}\left[\tilde{\mathbf{a}}_{n,L}\tilde{\mathbf{a}}_{n,L}^T\right]\right)-\mathrm{Tr}\left(\mathbf{m}_{L+1,d}\mathbf{m}_{L+1,d}^T\mathbb{E}\left[\tilde{\mathbf{a}}_{n,L}\right]\mathbb{E}\left[\tilde{\mathbf{a}}_{n,L}^T\right]\right)\right)$$

$$-\frac{1}{2}\sum_{l=1}^{L}\sum_{d=1}^{D_l}\mathbb{E}\left[\frac{1}{\eta_{1,d}^2}\right]\left(\sum_{n=1}^{N}\left(\rho_{n,l,d}\mathbb{E}\left[\tilde{\mathbf{W}}_{l,d}\right]\mathbb{E}\left[\tilde{\mathbf{a}}_{n,l-1}\right]-\mathbb{E}\left[\mathbf{a}_{n,l,d}\right]\right)^2+\mathbb{E}\left[\mathbf{a}_{n,l,d}^2\right]-\mathbb{E}\left[\mathbf{a}_{n,l,d}\right]^2\right)$$

$$-\frac{1}{2}\sum_{l=1}^{L}\sum_{d=1}^{D_l}\mathbb{E}\left[\frac{1}{\eta_{1,d}^2}\right]\left(\sum_{n=1}^{N}\rho_{n,l,d}\mathrm{Tr}\left(\left(\mathbf{B}_{l,d}+\mathbf{m}_{l,d}\mathbf{m}_{l,d}^T\right)\mathbb{E}\left[\tilde{\mathbf{a}}_{n,l-1}\tilde{\mathbf{a}}_{n,l-1}^T\right]\right)-\rho_{n,l,d}^2\mathrm{Tr}\left(\mathbf{m}_{l,d}\mathbf{m}_{l,d}^T\mathbb{E}\left[\tilde{\mathbf{a}}_{n,l-1}\right]\mathbb{E}\left[\tilde{\mathbf{a}}_{n,l-1}^T\right]\right)\right)$$

$$-\sum_{l=1}^{L}\sum_{d=1}^{D_l}\mathbb{E}\left[\frac{1}{\eta_{1,d}^2}\right]\left(\sum_{n=1}^{N}\rho_{n,l,d}\mathbb{E}\left[\tilde{\mathbf{W}}_{l,d}\right]\left(\mathbb{E}\left[\mathbf{a}_{n,l,d}\right]\mathbb{E}\left[\tilde{\mathbf{a}}_{n,l-1}\right]-\mathbb{E}\left[\mathbf{a}_{n,l,d}\tilde{\mathbf{a}}_{n,l-1}\right]\right)\right)$$

$$+\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{d=1}^{D_l}\frac{1}{T}\left(\rho_{n,l,d}-\frac{1}{2}\right)\left(\mathbb{E}\left[\tilde{\mathbf{W}}_{l,d}\right]\mathbb{E}\left[\tilde{\mathbf{a}}_{n,l-1}\right]\right)-\frac{1}{2T^2}\mathbb{E}\left[\omega_{n,l,d}\right]\left(\mathrm{Tr}\left(\left(\mathbf{B}_{l,d}+\mathbf{m}_{l,d}\mathbf{m}_{l,d}^T\right)\mathbb{E}\left[\tilde{\mathbf{a}}_{n,l-1}\tilde{\mathbf{a}}_{n,l-1}^T\right]\right)\right)$$

$$-\sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{d=1}^{D_l}(\rho_{n,l,d}\log\rho_{n,l,d}+(1-\rho_{n,l,d})\log(1-\rho_{n,l,d}))-\frac{A_{n,l,d}^2}{2}\mathbb{E}\left[\omega_{n,l,d}\right]+\log(\cosh(A_{n,l,d}/2)).$$

Note that when implementing VI with EM scheme, we adjust the formula above by adding the term which arises in the normalizing constant C_τ defined when computing the ELBO of global shrinkage parameters, specifically, we add

$$\mathrm{ELBO}_{EM}=(L+1)\left(\nu_{\mathrm{glob}}\left(\log(\lambda_{\mathrm{glob}})-\log(\delta_{\mathrm{glob}})\right)-\log\left(K_{\nu_{\mathrm{glob}}}(\lambda_{\mathrm{glob}}\delta_{\mathrm{glob}})\right)\right)$$

$$+ \sum_{l=1}^{L+1} (\nu_{\text{glob}} - 1) \mathbb{E} \left[\log \tau_l \right] - \frac{1}{2} \lambda_{\text{glob}}^2 \mathbb{E} \left[\tau_l \right] - \nu_l \log(\lambda_{\text{glob}}).$$

B.2 ELBO for prediction

To obtain the posterior predictive distribution, we compute the approximate variational predictive distributions of $\mathbf{a}_*$, $\boldsymbol{\gamma}_*$ and $\boldsymbol{\omega}_*$ with the objective function being the ELBO of Equation (10). Thus, in the predictive step of our algorithm, we monitor the convergence of the ELBO of $\mathbf{a}_*$, $\boldsymbol{\gamma}_*$ and $\boldsymbol{\omega}_*$, which we derive as follows:

$$\mathbb{E} \left[\log p(\mathbf{a}_*, \boldsymbol{\gamma}_*, \boldsymbol{\omega}_* | \mathbf{W}, \mathbf{b}, \boldsymbol{\Sigma}) \right] - \mathbb{E} \left[\log q(\mathbf{a}_*) \right] - \mathbb{E} \left[\log q(\boldsymbol{\gamma}_*) \right] - \mathbb{E} \left[\log q(\boldsymbol{\omega}_*) \right]$$

$$= \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E} \left[\log \mathrm{N} \left(a_{*,l,d} \mid \boldsymbol{\gamma}_{*,d} \odot \mathbf{z}_{*,l,d}, \boldsymbol{\Sigma}_{l,d} \right) \right] + \mathbb{E} \left[\log \left(\exp \left(\frac{\kappa_{*,l,d} z_{*,l,d}}{T} \right) \exp \left(- \frac{\omega_{*,l,d} z_{*,l,d}^2}{2T^2} \right) \mathrm{PG}(\omega_{n,l,d} \mid 1, 0) \right) \right]$$

$$- \sum_{l=1}^{L} \mathbb{E} \left[\log \mathrm{N} \left(\mathbf{a}_{*,l} \mid \mathbf{t}_{*,l} + \mathbf{M}_{*,l} \mathbf{a}_{*,l-1}, \mathbf{S}_{*,l} \right) \right] - \sum_{l=1}^{L} \sum_{d=1}^{D_l} \left(\mathbb{E} \left[\log \mathrm{Bern} \left(\gamma_{*,l,d} \mid \rho_{*,l,d} \right) \right] + \mathbb{E} \left[\log \mathrm{PG}(\omega_{*,l,d} \mid 1, A_{*,l,d}) \right] \right)$$

$$= - \frac{1}{2} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E} \left[\frac{1}{\eta_{l,d}^2} \right] \left(\mathbb{E} \left[a_{*,l,d}^2 \right] - 2 \mathbb{E} \left[\gamma_{*,l,d} \right] \mathbb{E} \left[\tilde{\mathbf{W}}_{l,d} \right] \mathbb{E} \left[\tilde{\mathbf{a}}_{*,l-1} a_{*,l,d} \right] \right)$$

$$- \frac{1}{2} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E} \left[\log \eta_{l,d}^2 \right] + \frac{1}{2} \sum_{l=1}^{L} \log(|\mathbf{S}_{*,l}|) + \frac{1}{T} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \left(\rho_{*,l,d} - \frac{1}{2} \right) \mathbb{E} \left[\widetilde{\mathbf{W}}_{l,d} \right] \mathbb{E} \left[\widetilde{\mathbf{a}}_{*,l-1} \right]$$

$$- \frac{1}{2} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \left(\mathbb{E} \left[\frac{1}{\eta_{l,d}^2} \right] \mathbb{E} \left[\gamma_{*,l,d}^2 \right] + \frac{1}{T^2} \mathbb{E} \left[\omega_{*,l,d} \right] \right) \mathrm{Tr} \left(\mathbb{E} \left[\tilde{\mathbf{W}}_{l,d}^T \tilde{\mathbf{W}}_{l,d} \right] \mathbb{E} \left[\tilde{\mathbf{a}}_{*,l-1} \tilde{\mathbf{a}}_{*,l-1}^T \right] \right)$$

$$- \sum_{l=1}^{L} \sum_{d=1}^{D_l} \left(\rho_{*,l,d} \log \rho_{*,l,d} + (1 - \rho_{*,l,d}) \log(1 - \rho_{*,l,d}) \right)$$

$$+ \sum_{l=1}^{L} \sum_{d=1}^{D_l} \frac{A_{*,l,d}^2}{2} \mathbb{E} \left[\omega_{*,l,d} \right] - \sum_{l=1}^{L} \sum_{d=l}^{D_l} \log(\cosh(\frac{A_{*,l,d}}{2})) + \mathrm{const}$$

$$= - \frac{1}{2} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E} \left[\frac{1}{\eta_{1,d}^2} \right] \left(\left(\rho_{*,l,d} \mathbb{E} \left[\tilde{\mathbf{W}}_{l,d} \right] \mathbb{E} \left[\tilde{\mathbf{a}}_{*,l-1} \right] - \mathbb{E} \left[a_{*,l,d} \right] \right)^2 + \mathbb{E} \left[a_{*,l,d}^2 \right] - \mathbb{E} \left[a_{*,l,d} \right]^2 \right)$$

$$- \frac{1}{2} \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E} \left[\frac{1}{\eta_{l,d}^2} \right] \left(\rho_{*,l,d} \mathrm{Tr} \left(\left(\mathbf{B}_{l,d} + \mathbf{m}_{l,d} \mathbf{m}_{l,d}^T \right) \mathbb{E} \left[\tilde{\mathbf{a}}_{*,l-1} \tilde{\mathbf{a}}_{*,l-1}^T \right] \right) - \rho_{*,l,d}^2 \mathrm{Tr} \left(\mathbf{m}_{l,d} \mathbf{m}_{l,d}^T \mathbb{E} \left[\tilde{\mathbf{a}}_{*,l-1} \right] \mathbb{E} \left[\tilde{\mathbf{a}}_{*,l-1}^T \right] \right) \right)$$

$$- \sum_{l=1}^{L} \sum_{d=1}^{D_l} \mathbb{E} \left[\frac{1}{\eta_{l,d}^2} \right] \left(\rho_{*,l,d} \mathbb{E} \left[\tilde{\mathbf{W}}_{l,d} \right] \left(\mathbb{E} \left[\mathbf{a}_{*,l,d} \right] \mathbb{E} \left[\tilde{\mathbf{a}}_{*,l-1} \right] - \mathbb{E} \left[a_{*,l,d} \tilde{\mathbf{a}}_{*,l-1} \right] \right) \right) + \frac{1}{2} \mathbb{E} \left[\log \eta_{l,d}^2 \right]$$

$$+ \frac{1}{2} \sum_{l=1}^{L} \log(|\mathbf{S}_{*,l}|) + \sum_{l=1}^{L} \sum_{d=1}^{D_l} \frac{1}{T} \left(\rho_{*,l,d} - \frac{1}{2} \right) \mathbb{E} \left[\widetilde{\mathbf{W}}_{l,d} \right] \mathbb{E} \left[\widetilde{\mathbf{a}}_{*,l-1} \right] + \frac{1}{2T^2} \left(\mathbb{E} \left[\omega_{*,l,d} \right] \right) \mathrm{Tr} \left(\mathbb{E} \left[\tilde{\mathbf{W}}_{l,d}^T \tilde{\mathbf{W}}_{l,d} \right] \mathbb{E} \left[\tilde{\mathbf{a}}_{*,l-1} \tilde{\mathbf{a}}_{*,l-1}^T \right] \right)$$

$$- \sum_{l=1}^{L} \sum_{d=1}^{D_l} \left(\rho_{*,l,d} \log \rho_{*,l,d} + (1 - \rho_{*,l,d}) \log(1 - \rho_{*,l,d}) - \frac{A_{*,l,d}^2}{2} \mathbb{E} \left[\omega_{*,l,d} \right] + \log(\cosh(\frac{A_{*,l,d}}{2})) \right) + \mathrm{const}.$$

C Experiments

C.1 Initialization schemes

Initialization plays an important role in the ability of Bayesian inference algorithms to effectively approximate the posterior. This is especially true in variational schemes for complex posteriors (such as for BNNs), which are only guaranteed to converge to local optimum. We design two possible variations of random yet effective initialization schemes. To simplify the exposition, we describe the procedure in the case of Inverse Gamma shrinkage priors, for which $\lambda = 0$ and the selection of the scale parameters δ determines the level of shrinkage. Note that during the training step, we employ the expectation-maximization algorithm to set an optimal δ_{glob}, whilst the value of $\delta_{loc,l}$ remains the fixed. To encourage more shrinkage for larger depth, we assume

$\delta_{\text{glob}} \propto 1/\sqrt{L}$, and to encourage shrinkage for larger width set $\delta_{\text{loc},l} \propto 1/\sqrt{D_l}$. Given specified values of $\nu_{\text{loc}}, \nu_{\text{glob}}, \delta_{\text{loc}}, \delta_{\text{glob}}, \alpha_0^h, \alpha_0, \beta_0^h, \beta_0$, we first re-scale the shrinkage parameters to scale appropriately

$$\delta_{\text{glob}} = \frac{\delta_{\text{glob}}}{\sqrt{L}}, \delta_{\text{loc},l} = \frac{\delta_{\text{loc}}}{\sqrt{D_{l-1}}}, \nu_{\text{loc},l} = \nu_{\text{loc}},$$

and the initialization steps are:

Algorithm 3 Initialization

Require: Training inputs $\mathbf{x}_n$, $n = 1, \ldots, N$; choice of mode *laplace* or *spike-slab*

  $\mathbf{z}_{n,0} = \mathbf{x}_n$

  for $l = 1 \ldots L$, **do**

    set $\Delta = 0.05 * (\max(\mathbf{z}_{n,l-1}) - \min(\mathbf{z}_{n,l-1}))$

    for $d = 1 \ldots D_l$ **do**

      if *laplace* **then**

$$m_{l,d,d'}^W \sim \text{Laplace}\left(0, \sqrt{\frac{2}{D_{l-1}}}\right),$$

      end if.

      if *spike-slab* **then**

$$m_{l,d,d'}^W \sim \pi \text{N}\left(0, \frac{2}{\sqrt{D_{l-1}}}\right) + (1-\pi)\delta_0, \text{ where } \pi = \frac{1}{1 + \sqrt{D_{l-1}}},$$

      end if.

$$\mathbf{s} = (s_1, \ldots, s_{D_{l-1}}), \text{ where } s_{d'} \sim \text{Unif}([\min(z_{n,l-1,d'}) - \Delta_{d'}, \max(z_{n,l-1,d'}) + \Delta_{d'}]),$$
$$m_{l,d}^b = -\mathbf{m}_{l,d}^W \mathbf{s}, \ \mathbf{m}_{l,d} = \left(m_{l,d}^b, \mathbf{m}_{l,d}^W\right),$$

    end for.

$$\rho_{n,l,d} = \sigma\left(\frac{m_{l,d}^b + \mathbf{m}_{l,d}^W \mathbf{z}_{n,l-1}}{T}\right) \quad d = 1, \ldots, D_l,$$

$$\mathbf{M}_{n,l} = \mathbf{m}_l^W \odot \rho_{n,l} \mathbf{1}_{D_l}^T, \text{ where by } \mathbf{1} \text{ we denote a vector of ones},$$

$$\mathbf{t}_{n,l} = m_l^b \odot \rho_{n,l},$$

$$\mathbf{z}_{n,l} = \mathbf{M}_{n,l} \mathbf{z}_{n,l-1} + \mathbf{t}_{n,l},$$

  end for.

Ensure: $\mathbf{M}_{n,l}, \mathbf{t}_{n,l}, \mathbf{m}_{l,d}$ for $l = 1, \ldots, L, d = 1, \ldots, D_l$ and $\mathbf{z}_L$.

---

1. Covariance for biases and weights: $\mathbf{B}_{l,d} = 0.01 \mathbf{I}_{D_{l-1}+1}$ for $l = 1, \ldots, L+1, d = 1, \ldots D_l$.

2. Covariance for stochastic activation: $\mathbf{S}_{n,l} = 0.01 \mathbf{I}_{D_l}$ for $n = 1, \ldots, N, \ l = 1, \ldots, L$.

3. Variational parameters for $\boldsymbol{\eta}$: Set $\alpha_{L+1,d} = \alpha_0, \alpha_{l,d} = \alpha_0^h$ and $\beta_{L+1,d} = \beta_0, \beta_{l,d} = \beta_0^h$.

4. Variational parameters for $\boldsymbol{\tau}, \boldsymbol{\psi}$:

$$\nu_{\text{loc},l,d,d'} = \nu_{\text{loc},l}, \ \nu_{\text{glob},l} = \nu_{\text{glob}},$$

$$\delta_{\text{glob},l} \sim \sqrt{2(\nu_{\text{glob},l} - 1)\text{IG}\left(\nu_{\text{glob},l}, \delta_{\text{glob}}\right)},$$

$$\delta_{\text{loc},l,d,d'} \sim \sqrt{2(\nu_{\text{loc},l,d,d'} - 1)\text{IG}\left(\nu_{\text{loc},l,d,d'}, \delta_{\text{loc},l}\right)}.$$

5. Use Algorithm 3 to initialize the variational means of the weights and biases for all intermediate layers, and the variational means of the stochastic activations and the variational parameters of the binary activations.

6. Variational mean of the weights and biases for the last layer $\mathbf{m}_{L+1}$ is obtained as a solution of fitting D_y ridge regressions with inputs $\mathbf{z}_L$ and outputs $\mathbf{y}_d$.

C.2 Implementation details

When comparing the performance of our method to already existing ones we implement the following model in Numpyro:

$$\mathbf{y} \sim \text{N}\left(\mathbf{W}_{L+1}\text{ReLU}(\mathbf{z}_L) + \mathbf{b}_L, \boldsymbol{\Sigma}\right), \quad \text{where} \quad \boldsymbol{\Sigma} \sim \text{IG}(2, \sigma_y)\mathbf{I}_{D_y},$$

$$\mathbf{z}_l = \mathbf{W}_l\text{ReLU}(\mathbf{z}_{l-1}) + \mathbf{b}_l, \ W_{l,d,d'} \sim \text{N}\left(0, \frac{\sigma_W^2\gamma}{\sqrt{D_{l-1}}}\right), \quad b_{l,d} \sim \text{N}(0, \sigma_b^2\gamma),$$

where $\mathbf{z}_{n,0} = \mathbf{x}_n$, $\gamma \sim \text{IG}(2, 1)$ and $l = 1, \ldots, L$, $d = 1, \ldots D_l$, $d' = 1, \ldots, D_{l-1}$. The choice of σ_y, σ_W and σ_b is made in accordance with α_0, s_0 and $\delta_{\text{loc},l}$, respectively. For experiments with SVI we use Adam optimizer with learning rate set to 0.001 and maximum number of iterations varying from 5000 to 20000 depending on the dataset and depth of the network. Additionally, we consider the Bayes by Backprop model of Blundell et al. (2015) and adapt its Pytorch implementation from the publicly available repository Javier (2019). For all experiments with BBB we set the learning rate to 0.01 and maximum number of epochs varies from 500 to 1000.

Suppose that the data on which we evaluate the predictive performance consists of N points and the true target is $\mathbf{y}^*$, then recorded evaluation metrics are RMSE, NLL and EC and are computed as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_n^N [(y_n^* - \mathbb{E}[y_n^o])^2]},$$

$$\text{NLL} = \frac{1}{N}\sum_n^N \log \text{N}(y_n^* \mid \mathbb{E}[y_n^o], \text{Var}(y_n^o))$$

$$\text{EC} = \frac{\#\{\mathbf{y}^* \in [q_{0.025}^o, q_{0.975}^o]\}}{N}.$$

where the predicted observations are $\mathbf{y}^o$ and the corresponding quantiles are denoted as q^o. When computing quantiles to obtain empirical coverage and illustrating the uncertainty in Section 4 and below in Appendix C.3, we rely on the Gaussian approximation.

C.3 Supplementary material to the diabetes example.

Figure 10 supplements Table 2 and the diabetes example in Section 4.2. Here, in the case of VBNN, BBB and SVI models we provide the uncertainty of the observations and in the case of the LassoCV we provide residual standard deviation. Additionally, we illustrate the sparse prediction and the uncertainty obtained from sparse weights of the VBNN, which largely coincide with the original prediction and uncertainty estimates. Whilst the coverage estimates for observations of VBNN and BBB are comparable, the SVI underestimates the uncertainty and provides a dramatically lower coverage for observations.
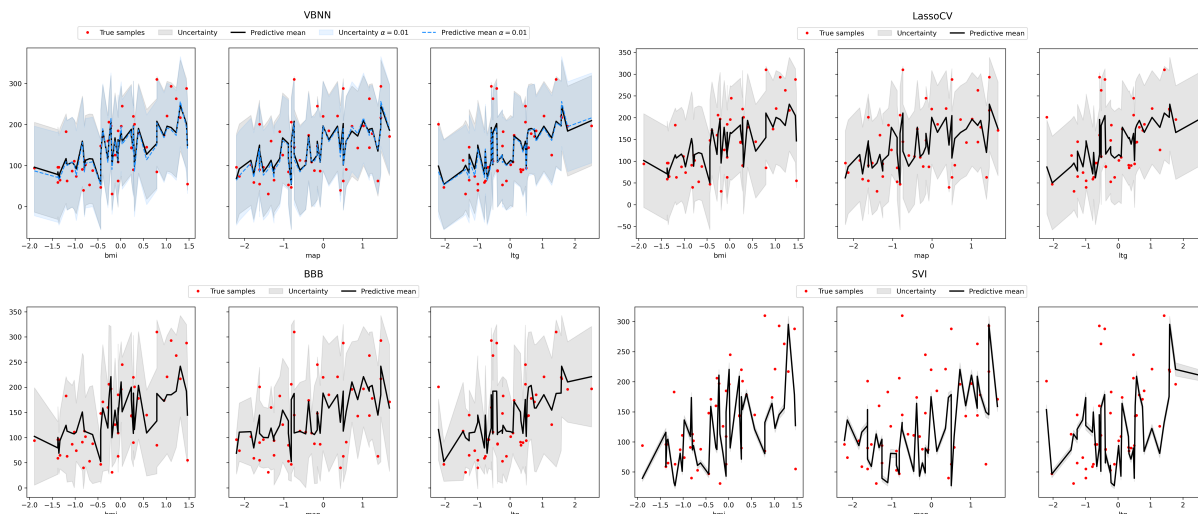
Figure 10: Predictive mean and the uncertainty estimates for the observations for three of the predictors with considerable contribution.

C.4   Supplementary information on the datasets

Boston housing Harrison and Rubinfeld (1978): $n = 506, p = 13$, the predictors are per capita crime rate by town, the proportion of residential land zoned for lots over 25,000 sq.ft., the proportion of non-retail business acres per town, Charles River dummy variable, nitrite oxides concentration, average number of rooms per dwelling, the proportion of owner-occupied, units built before 1940, weighted distances to five Boston employment centres, index of accessibility to radial highways, full-value property-tax rate, the pupil-teacher ratio by town, the quantitative measure of systemic racism as a factor in house pricing, lower status of the population; the response of interest is the median value of owner-occupied homes. The Boston housing dataset is among the most popular pip available datasets, and with respect to variable selection it was considered in e.g. Schäfer and Chopin (2013).

Energy Tsanas and Xifara (2012): $n = 768, p = 8$, the predictors are relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution, and the task is to predict the heating load of residential buildings.

Yacht dynamics J. et al. (2013): $n = 308, p = 6$, the predictors are long position, prismatic coefficient, length-displacement ratio, bean-draught ratio length-bean ratio and froude number, and the task is to model the residuary resistance per unit weight of displacement for a yacht hull.

Concrete compressive strength Yeh (2007): $n = 1030, p = 8$, the predictors are cement, furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate and the age of testing, and the response variable is the compressive strength of concrete. This is also considered from the variable selection perspective in several works including Schäfer and Chopin (2013); Griffin (2024).

Concrete slump test Yeh (2009): $n = 103, p = 7$, the predictors are concrete ingredients, namely cement, furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate and the task is to predict slump of concrete.

D   Review of relevant distributions

D.1   Generalized Inverse Gaussian

The Generalized Inverse Gaussian has density:

$$p(x \mid \nu, \delta, \lambda) = \frac{(\lambda/\delta)^\nu}{2K_\nu(\lambda\delta)} x^{\nu-1} \exp\left(-\frac{1}{2}(\delta^2/x + \lambda^2 x)\right),$$

where $K_\nu()$ is the modified Bessel function of the second kind. The GIG prior requires $\nu > 0$ if $\delta = 0$ and $\nu < 0$ if $\lambda = 0$ for a proper prior. Then the expectations arising in computations throughout this paper are:

$$\mathbb{E}[x] = \frac{\delta K_{\nu+1}(\lambda\delta)}{\lambda K_\nu(\lambda\delta)},$$

$$\mathbb{E}\left[\frac{1}{x}\right] = \frac{\lambda K_{\nu+1}(\lambda\delta)}{\delta K_\nu(\lambda\delta)} - \frac{2\nu}{\delta^2}.$$

Often, it is sensible to consider special cases of the GIG, which include:

1. Inverse Gamma: when $\lambda = 0$, the GIG reduces to the IG with density:

$$p(x \mid \nu, \delta) = \frac{2^\nu}{\delta^{2\nu}\Gamma(-\nu)}(1/x)^{-\nu+1}\exp\left(-\frac{\delta^2}{2x}\right),$$

where $\nu < 0$ and $\delta > 0$. This can also be re-written in terms of the more standard parametrization of the IG:

$$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/x)^{\alpha+1}\exp\left(-\frac{\beta}{x}\right),$$

where $\alpha = -\nu > 0$ and $\beta = \delta^2/2 > 0$. Note that if $w \sim N(0, \tau)$ and $\tau \sim IG(\alpha, \beta)$, this implies a marginal student t-prior on w with degrees of freedom $\text{dof} = 2\alpha = -2\nu$ and scale $s = \sqrt{\beta/\alpha} = \delta/\sqrt{-2\nu}$. For example, setting $\nu = -1.5$ would correspond to $\text{dof} = 3$ and $\nu = -2.5$ is equivalent to $\text{dof} = 5$.

The relevant expectations for the VI updates and ELBO computation include:

$$\mathbb{E}[x] = \frac{\beta}{\alpha - 1} = \frac{-\delta^2}{2\nu + 2},$$

$$\mathbb{E}\left[\frac{1}{x}\right] = \frac{\alpha}{\beta} = \frac{-2\nu}{\delta^2},$$

where ψ is the logarithmic derivative of the gamma function (a.k.a. digamma function).

2. Gamma: when $\delta^2 = 0$, the GIG reduces to the Gamma with density:

$$p(x \mid \nu, \lambda) = \frac{\lambda^{2\nu}}{2^\nu\Gamma(\nu)}x^{\nu-1}\exp(-\frac{\lambda^2}{2}x),$$

where $\nu > 0$, rewriting in the standard parametrization with $\alpha = \nu$ and $\beta = \lambda^2/2$:

$$p(x \mid \alpha, \beta) = \beta^\alpha\frac{1}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x),$$

where $\alpha = \nu > 0$ and $\beta = \lambda^2/2 > 0$. Similarly, the relevant expectations are:

$$\mathbb{E}[x] = \frac{\alpha}{\beta} = \frac{2\nu}{\lambda^2},$$

$$\mathbb{E}\left[\frac{1}{x}\right] = \frac{\beta}{\alpha - 1} = \frac{\lambda^2}{2(\nu - 1)}.$$

Note that if $w \sim N(0, \tau)$ and $\tau \sim \text{Gam}(1, \beta)$, this implies a marginal Laplace prior on w (i.e. Bayesian Lasso Park and Casella (2008)) with scale $s = 1/\sqrt{2\beta} = 1/\lambda$.

3. Inverse Gaussian (IGaus): when $\nu = -1/2$, the GIG reduces to the Inverse Gaussian with density:

$$p(x \mid \delta, \lambda) = \frac{\delta}{\sqrt{2\pi x^3}}\exp\left(-\frac{(\lambda x - \delta)^2}{2x}\right),$$

where setting $\alpha = \delta/\lambda > 0$ and $\beta = \delta^2 > 0$ we derive

$$p(x \mid \alpha, \beta) = \left(\frac{\beta}{2\pi x^3} \right)^{\frac{1}{2}} \exp\left(\frac{-\beta(x - \alpha)^2}{2\alpha^2 x} \right).$$

The relevant expectations for the VI updates and ELBO computation include:

$$\mathbb{E}\left[x\right] = \alpha = \frac{\delta}{\lambda},$$

$$\mathbb{E}\left[\frac{1}{x}\right] = \frac{1}{\alpha} + \frac{1}{\beta} = \frac{\lambda}{\delta} + \frac{1}{\delta^2}.$$

Note that if $w \sim \mathrm{N}(0,\tau)$ and $\tau \sim \mathrm{IGaus}(\alpha,\beta)$, the marginal distribution is of the form Caron and Doucet (2008):

$$p(w_k) = \frac{1}{\pi\alpha} \left(\frac{\beta}{\beta + w_k^2} \right)^{\frac{1}{2}} \exp\left(\frac{\beta^{\frac{1}{2}}}{\alpha} \right) K_1 \left(\frac{(\beta + w_k^2)^{\frac{1}{2}}}{\alpha} \right)$$

$$= \frac{\lambda}{\pi} \exp(\lambda) \left(\delta^2 + w_k^2 \right)^{-\frac{1}{2}} K_1 \left(\frac{\lambda}{\delta} \left(\delta^2 + w_k^2 \right)^{\frac{1}{2}} \right).$$

D.2 EM update for different cases of global-local priors

As discussed above, the special cases of the GIG include Inverse Gamma, Gamma and Inverse Gaussian distributions, we derive the EM updates in each of the special cases of priors:

1. Inverse Gamma: when the global shrinkage parameter has an Inverse Gamma distribution, then

$$\delta_{\mathrm{glob}} = \arg\max \left(\delta_{\mathrm{glob}}^2 \sum_{l=1}^{L+1} \frac{\nu_{\mathrm{glob},l}}{\delta_{\mathrm{glob},l}^2} - 2(L+1)\nu_{\mathrm{glob}} \log(\delta_{\mathrm{glob}}) \right),$$

$$\delta_{\mathrm{glob}} = ((L+1)\nu_{\mathrm{glob}})^{\frac{1}{2}} \left(\sum_{l=1}^{L+1} \frac{\nu_{\mathrm{glob},l}}{\delta_{\mathrm{glob},l}^2} \right)^{-\frac{1}{2}}.$$

2. Gamma: similarly, when global shrinkage parameter is Gamma:

$$\lambda_{\mathrm{glob}} = \arg\max \left(4(L+1)\nu_{\mathrm{glob}} \log(\lambda_{\mathrm{glob}}) - \lambda_{\mathrm{glob}}^2 \sum_{l=1}^{L+1} \frac{\delta_{\mathrm{glob},l} K_{\nu_{\mathrm{glob},l}+1}(\lambda_{\mathrm{glob},l}\delta_{\mathrm{glob},l})}{\lambda_{\mathrm{glob},l} K_{\nu_{\mathrm{glob},l}}(\lambda_{\mathrm{glob},l}\delta_{\mathrm{glob},l})} \right),$$

$$\lambda_{\mathrm{glob}} = (2(L+1)\nu_{\mathrm{glob}})^{\frac{1}{2}} \left(\sum_{l=1}^{L+1} \frac{\delta_{\mathrm{glob},l} K_{\nu_{\mathrm{glob},l}+1}(\lambda_{\mathrm{glob},l}\delta_{\mathrm{glob},l})}{\lambda_{\mathrm{glob},l} K_{\nu_{\mathrm{glob},l}}(\lambda_{\mathrm{glob},l}\delta_{\mathrm{glob},l})} \right)^{-\frac{1}{2}}.$$

3. Inverse Gaussian: if the global shrinkage parameter is Inverse Gaussian, then

$$\lambda_{\mathrm{glob}} = \arg\max \left(2(L+1)\lambda_{\mathrm{glob}}\delta_{\mathrm{glob}} - \lambda_{\mathrm{glob}}^2 \sum_{l=1}^{L+1} \frac{\delta_{\mathrm{glob},l} K_{\nu_{\mathrm{glob},l}+1}(\lambda_{\mathrm{glob},l}\delta_{\mathrm{glob},l})}{\nu_{\mathrm{glob},l} K_{\nu_{\mathrm{glob},l}}(\lambda_{\mathrm{glob},l}\delta_{\mathrm{glob},l})} \right),$$

$$\lambda_{\mathrm{glob}} = 2(L+1)\delta_{\mathrm{glob}} \left(\sum_{l=1}^{L+1} \frac{\delta_{\mathrm{glob},l} K_{\nu_{\mathrm{glob},l}+1}(\lambda_{\mathrm{glob},l}\delta_{\mathrm{glob},l})}{\nu_{\mathrm{glob},l} K_{\nu_{\mathrm{glob},l}}(\lambda_{\mathrm{glob},l}\delta_{\mathrm{glob},l})} \right)^{-1}.$$