An Improved Variational Approximate Posterior for the Deep Wishart Process

Sebastian W. Ober[1] **Ben Anson**[*2] **Edward Milsom**[*2] **Laurence Aitchison**[2]

[1]University of Cambridge
[2]University of Bristol

Abstract

Deep kernel processes are a recently introduced class of deep Bayesian models that have the flexibility of neural networks, but work entirely with Gram matrices. They operate by alternately sampling a Gram matrix from a distribution over positive semi-definite matrices, and applying a deterministic transformation. When the distribution is chosen to be Wishart, the model is called a deep Wishart process (DWP). This particular model is of interest because its prior is equivalent to a deep Gaussian process (DGP) prior, but at the same time it is invariant to rotational symmetries, leading to a simpler posterior distribution. Practical inference in the DWP was made possible in recent work ("A variational approximate posterior for the deep Wishart process" Ober and Aitchison, 2021a) where the authors used a generalisation of the Bartlett decomposition of the Wishart distribution as the variational approximate posterior. However, predictive performance in that paper was less impressive than one might expect, with the DWP only beating a DGP on a few of the UCI datasets used for comparison. In this paper, we show that further generalising their distribution to allow linear combinations of rows and columns in the Bartlett decomposition results in better predictive performance, while incurring negligible additional computation cost.

1 INTRODUCTION

Deep kernel processes (DKPs) [Aitchison et al., 2021] are a class of deep Bayesian models which have the flexibility of neural networks (NNs), but work entirely with Gram matrices. NNs have many tuneable parameters which allow them to automatically adapt to problems, and therefore learn good top-layer representations, which turns out to be very important for complex tasks like image classification [Krizhevsky et al., 2012]. On the other hand, most kernels only have a very small number of tuneable hyperparameters, meaning that the kernel matrices they produce are comparatively rigid, and do not have the ability, that NNs have, to learn flexible top-layer representations. DKPs solve this problem by alternately taking the kernel matrix from the previous layer, and sampling from a distribution over positive semi-definite matrices, centred on the previous kernel. Since DKPs never sample features (except for the final outputs), they are distinct from e.g. deep Gaussian processes [Damianou and Lawrence, 2013] (DGPs), which sample features at every layer.

A particular DKP, called the deep Wishart process (DWP), is of particular interest since Aitchison et al. [2021] showed that its prior is equivalent to the DGP prior. However, they were unable to perform inference in the DWP due to the lack of a sufficiently flexible yet tractable distribution over positive semi-definite matrices to use as an approximate posterior. The first solution to this problem was posed by Ober and Aitchison [2021a], who developed a generalisation of the Bartlett decomposition of the Wishart distribution, and used it as the basis of their approximate posterior in a series of experiments that compared DWPs to DGPs. In theory, purely kernel-based methods should have an advantage over feature-based methods, since Gram matrices are invariant to certain symmetries to which feature-based methods are not, leading to simpler posteriors (see Appendix D2 in Aitchison et al. 2021). However, the experiments in Ober and Aitchison [2021a] showed only minor advantages over DGPs on a fraction of the datasets they tested. In this paper, we extend the generalised (singular) Wishart distributions proposed by Ober and Aitchison [2021a] by introducing parameters that rotate, stretch, and mix the rows and columns of the Bartlett decomposition, and show that this added flexibility in the approximate posterior allows DWPs to consistently match or outperform DGPs on UCI datasets, while adding negligible computation cost.

[*]These authors contributed equally to this work.

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:1555–1563.

2 CONTRIBUTIONS

Concretely, our contributions are:

- We propose the A-generalised (singular) Wishart and AB-generalised (singular) Wishart distributions, two flexible distributions over positive semi-definite matrices, and we provide full derivations for the densities of these distributions in Appendix A.
- We prove both analytically and empirically that the A/AB-generalised (singular) Wishart families are proper supersets of the generalised (singular) Wishart family proposed by Ober and Aitchison [2021a].
- We show experimentally that our proposed approximate posteriors provide significant performance benefits on UCI datasets, while adding negligible computation cost.

3 RELATED WORK

Perhaps the closest prior work is Ober and Aitchison [2021a], which introduces generalised (singular) Wishart approximate posteriors for the deep Wishart process. However, the performance in that paper was less impressive than one might have expected, indicating that there may be room to further improve the family of approximate posteriors over Gram matrices. We provide such an improvement by introducing A and AB-generalised (singular) Wishart approximate posteriors, which exhibit considerably improved performance over the original approximate posterior from Ober and Aitchison [2021a].

The deep kernel process line of work emerged from Aitchison et al. [2021]. While they introduced the deep Wishart process prior, they were not able to perform inference, as they did not have a suitable approximate posterior (that approximate posterior was developed in Ober and Aitchison 2021a). Instead, they were able to do inference in the alternative deep inverse Wishart process, which (unlike the deep Wishart process) does not have any equivalences to DGPs.

The deep kernel process research direction was originally inspired by work showing that infinite-width Bayesian neural networks have GP-distributed outputs [Lee et al., 2017, Matthews et al., 2018, Novak et al., 2018, Garriga-Alonso et al., 2018]. However, this limit is problematic in that the resulting GP kernel is a fixed, deterministic function of the inputs that cannot be learned from data. Thus, this limit eliminates representation or feature learning, which is perhaps the key mechanism behind the excellent practical performance of neural networks [Yang and Hu, 2020, Aitchison, 2020]. Deep kernel processes [Aitchison et al., 2021] were inspired by infinite width NNs but designed specifically to retain flexible, learned kernels.

Another related approach that enables representation learning in infinite-width NNs is the deep kernel machine (DKM) [Yang et al., 2023, Milsom et al., 2023]. DKMs differ from DKPs slightly because they are deterministic and correspond directly to an infinitely wide DGP with an infinitely wide top layer [Yang et al., 2023].

4 BACKGROUND

In order to understand deep Wishart processes, it is necessary to first define the Wishart distribution. Our implementation of deep Wishart processes further requires a flexibile approximate posterior. This motivates our proposed A/AB-generalised (singular) Wishart distributions, which in turn are obtained by considering the Bartlett decomposition.

4.1 WISHART DISTRIBUTION

The Wishart distribution is a generalisation of the gamma distribution to positive semi-definite matrices. Suppose we take a matrix $\mathbf{F} \in \mathbb{R}^{P \times \nu}$ whose columns $\mathbf{f}_\lambda \in \mathbb{R}^P \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ are multivariate Gaussian distributed vectors with $\lambda \in \{1, \dots, \nu\}$ and $\mathbf{\Sigma} \in \mathbb{R}^{P \times P}$, where P is the number of datapoints. Then,

$$\mathbf{W} := \mathbf{F}\mathbf{F}^T = \sum_{\lambda=1}^{\nu} \mathbf{f}_\lambda \mathbf{f}_\lambda^T \tag{1}$$

is said to be Wishart distributed, denoted $\mathbf{W} \in \mathbb{R}^{P \times P} \sim \mathcal{W}(\mathbf{\Sigma}, \nu)$, with positive definite scale matrix $\mathbf{\Sigma}$ and degrees of freedom ν.

4.2 DEEP WISHART PROCESS

The deep Wishart process is a specific instantiation of a deep kernel process [Aitchison et al., 2021]. Moreover, DGPs can be reframed as deep Wishart processes. Consider a DGP model, which samples features $\mathbf{F}_\ell \in \mathbb{R}^{P \times \nu_\ell}$ at each layer sequentially from a Gaussian process,

$$P(\mathbf{F}_\ell \mid \mathbf{F}_{\ell-1}) = \prod_{\lambda=1}^{\nu_\ell} \mathcal{N}(\mathbf{f}_\lambda^\ell; \mathbf{0}, \mathbf{K}(\mathbf{F}_{\ell-1})). \tag{2}$$

The columns $\mathbf{f}_\lambda^\ell$ are IID multivariate Gaussian random variables, where we apply a kernel function $k(\cdot,\cdot) : \mathbb{R}^{\nu_\ell} \times \mathbb{R}^{\nu_\ell} \to \mathbb{R}$ pairwise to the previous layer to form the covariance matrix $\mathbf{K}(\mathbf{F}_{\ell-1}) \in \mathbb{R}^{P \times P}$.

With deep kernel processes, instead of working with features $\mathbf{F}_\ell$, we work with Gram matrices $\mathbf{G}_\ell$,

$$\mathbf{G}_\ell = \frac{1}{\nu_\ell} \mathbf{F}_\ell \mathbf{F}_\ell^T. \tag{3}$$

With $\mathbf{F}_\ell$ defined by Eq. (2), $\mathbf{G}_\ell$ is sampled using the same generative process that defines the Wishart distribution

(Sec. 4.1), so we have,

$$\mathbf{G}_\ell \mid \mathbf{F}_{\ell-1} \sim \mathcal{W}\left(\frac{1}{\nu_\ell}\mathbf{K}(\mathbf{F}_{\ell-1}), \nu_\ell\right). \tag{4}$$

The final ingredient for defining a deep Wishart process is the fact that we can often compute the kernel matrix $\mathbf{K}$ from $\mathbf{G}_{\ell-1}$, without having to know $\mathbf{F}_{\ell-1}$. This is true for many common kernels, including isotropic kernels, which only depend on the average squared euclidean distance $R_{ij}^{\ell-1}$ between datapoints,

$$R_{ij}^\ell = \frac{1}{\nu_\ell}\sum_{\lambda=1}^{\nu_\ell}(F_{i\lambda}^\ell - F_{j\lambda}^\ell)^2 \tag{5a}$$
$$= \frac{1}{\nu_\ell}\sum_{\lambda=1}^{\nu_\ell}(F_{i\lambda}^\ell)^2 - 2F_{i\lambda}^\ell F_{j\lambda}^\ell + \left(F_{j\lambda}^\ell\right)^2 \tag{5b}$$
$$= G_{ii}^\ell - 2G_{ij}^\ell + G_{jj}^\ell. \tag{5c}$$

(see Aitchison et al., 2021 for further details). Hence we can use $\mathbf{K}\left(\mathbf{G}_{\ell-1}\right)$, eliminating features to work entirely with gram matrices, defining our deep Wishart process like so,

$$\mathrm{P}\left(\mathbf{G}_\ell \mid \mathbf{G}_{\ell-1}\right) = \mathcal{W}\left(\mathbf{G}_\ell; \frac{1}{\nu_\ell}\mathbf{K}(\mathbf{G}_{\ell-1}), \nu_\ell\right), \tag{6a}$$
$$\mathrm{P}\left(\mathbf{F}_{L+1} \mid \mathbf{G_L}\right) = \prod_{\lambda}^{\nu_{L+1}} \mathcal{N}\left(\mathbf{f}_\lambda^{L+1}; \mathbf{0}, \mathbf{K}(\mathbf{G}_L)\right), \tag{6b}$$

where L is the number of hidden layers, $\mathbf{G}_0 = \frac{1}{\nu_0}\mathbf{X}\mathbf{X}^T$ for input data $\mathbf{X} \in \mathbb{R}^{P\times\nu_0}$, and at the output layer (Eq. 6b) we sample features that can be provided to a likelihood function $\mathrm{P}\left(\mathbf{Y} \mid \mathbf{F}_{L+1}\right)$, e.g. a Gaussian likelihood for regression, or a categorical likelihood for classification.

4.3 VARIATIONAL INFERENCE IN DWPs

As is the case with almost all Bayesian models of reasonable complexity, the true posterior $\mathrm{P}\left(\mathbf{G}_1, \cdots, \mathbf{G}_L \mid \mathbf{X}, \mathbf{Y}\right)$ is intractable. We therefore use variational inference (VI), which replaces the true posterior with an approximate posterior $\mathrm{Q}\left(\mathbf{G}_1, \cdots, \mathbf{G}_L\right)$. This distribution is taken from a variational family of distributions with parameters ϕ, which are optimised to maximise a lower bound on the marginal log-likelihood of the data.

We consider approximate posteriors that factorise layerwise,

$$\mathrm{Q}\left(\mathbf{G}_1, \cdots, \mathbf{G}_L\right) = \prod_{\ell=1}^{L}\mathrm{Q}\left(\mathbf{G}_\ell \mid \mathbf{G}_{\ell-1}\right), \tag{7}$$

where each term $\mathrm{Q}\left(\mathbf{G}_\ell \mid \mathbf{G}_{\ell-1}\right)$ is a distribution over positive definite matrices. Note that although the prior of each layer is Wishart distributed, the posterior is not Wishart distributed in general. The seemingly obvious choice for this variational family is the Wishart family itself, but as Aitchison et al. [2021] argued, this is not flexible enough. For $\mathbf{G} \sim \mathcal{W}(\mathbf{\Sigma}, \nu)$ (particularly in the case where ν is fixed), the mean and variance cannot be independently specified since we have

$$\mathbb{E}[\mathbf{G}] = \nu\mathbf{\Sigma}, \tag{8a}$$
$$\mathbb{V}[G_{ij}] = \nu(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma jj). \tag{8b}$$

The ability to independently specify the variance is critical for an approximate posterior to be able to capture potentially narrow true posteriors, so we need an alternative. Aitchison et al. [2021] also suggested that a non-central Wishart distribution would be flexible enough to use as an approximate posterior, but its density is too expensive to evaluate as part of the training loop. Hence Aitchison et al. [2021] ultimately did not perform inference in the DWP, instead opting to change the model. In a subsequent work, Ober and Aitchison [2021a] introduced the generalised (singular) Wishart distribution, which finally allowed practical inference in DWPs, and which our work builds upon. In order to define that distribution, we first need to recap the Bartlett decomposition.

4.4 THE BARTLETT DECOMPOSITION

The Bartlett decomposition [Bartlett, 1933] is a factorisation for Wishart random variables. Specifically, if $\mathbf{W} \sim \mathcal{W}(\mathbf{I}, \nu)$ is a standard Wishart random variable (that is, it has identity scale matrix, $\mathbf{\Sigma} = \mathbf{I}$), then we have,

$$\mathbf{W} = \mathbf{T}\mathbf{T}^T, \tag{9}$$

where $\mathbf{T}$ is lower triangular, with the square of its diagonals Gamma-distributed, and its off-diagonals Gaussian-distributed,

$$\mathbf{T} = \begin{pmatrix} T_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ T_{P1} & \cdots & T_{PP} \end{pmatrix}, \tag{10a}$$
$$\mathrm{P}\left(T_{ii}^2\right) = \mathrm{Gamma}\left(T_{ii}^2; \tfrac{\nu-i+1}{2}, \tfrac{1}{2}\right), \tag{10b}$$
$$\mathrm{P}\left(T_{i>j}\right) = \mathcal{N}\left(T_{i>j}; 0, 1\right), \tag{10c}$$

In particular each element of $\mathbf{T}$ is independent. For Wishart distributions with non-identity scale matrices, we can compute the Cholesky decomposition $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T$, so that $\mathbf{W} = \mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{L}^T$ (this follows from the canonical definition of the Wishart using Gaussian vectors).

4.5 GENERALISED (SINGULAR) WISHART DISTRIBUTION

As shown by Ober and Aitchison [2021a], a generalisation of the Wishart distribution can be obtained by allowing the Bartlett decomposition to be more flexible (and by allowing singular matrices [Srivastava, 2003], since the Wishart ordinarily only supports positive definite matrices). Namely, we can introduce parameters $\alpha_j, \beta_j, \mu_{ij}, \sigma_{ij}$ such that the decomposition $\mathbf{T}$ has distribution,

$$\mathbf{T} = \begin{pmatrix} T_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ T_{\nu 1} & \cdots & T_{\nu\nu} \\ \vdots & \ddots & \vdots \\ T_{P1} & \cdots & T_{P\nu} \end{pmatrix}, \tag{11a}$$

$$\mathrm{Q}\left(T_{ii}^2\right) = \mathrm{Gamma}\left(T_{ii}^2;\, \alpha_i, \beta_i\right),\; i \in \{1, \dots, \nu\}, \tag{11b}$$

$$\mathrm{Q}\left(T_{i>j}\right) = \mathcal{N}\left(T_{i>j};\, \mu_{ij}, \sigma_{ij}^2\right) \tag{11c}$$

where, in the singular case of $\nu < P$, $\mathbf{T}$ is now a (tall) rectangular matrix, with the upper square block being lower triangular. This more flexible distribution defines a standard generalised (singular) Wishart random variable, denoted $\mathbf{T}\mathbf{T}^T \sim \mathcal{GW}\left(\mathbf{I}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$. For the more general case of $\boldsymbol{\Sigma} \neq \mathbf{I}$, we compute the cholesky decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ and obtain $\mathbf{W} \sim \mathcal{GW}\left(\boldsymbol{\Sigma}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$ as $\mathbf{W} = \mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{L}^T$.

For the general case $\mathbf{W} = \mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{L}^T$, Ober and Aitchison [2021a] showed that the density of $\mathbf{W}$ is

$$\mathrm{Q}\left(\mathbf{W}\right) = \left(\prod_{j=1}^{P} \frac{1}{L_{jj}^{\min(j,\nu)}}\right) \prod_{j=1}^{\tilde{\nu}} \frac{\mathrm{Gamma}\left(T_{jj}^2;\, \alpha_j, \beta_j\right)}{T_{jj}^{P-j} L_{jj}^{P-j+1}} \prod_{i=j+1}^{P} \mathcal{N}\left(T_{ij};\, \mu_{ij}, \sigma_{ij}^2\right). \tag{12}$$

5 METHODS

5.1 A-$\mathcal{GW}$ AND AB-$\mathcal{GW}$ DISTRIBUTIONS

Whilst the generalised (singular) Wishart distribution represented a big step for approximate inference in DWPs, the experimental results in Ober and Aitchison [2021a] indicated much room for improvement. Despite the theoretical advantages that DWPs have over DGPs due to their invariance to certain posterior symmetries, the DGP still outperformed the DWP in a few cases. By contrast, the A-generalised / AB-generalised (singular) Wishart distributions we introduce in this paper allow the DWP to match or outperform the DGP on all datasets we tested.

One issue with the generalised (singular) Wishart distribution is that it is unclear how flexible it is with respect to linear transformations. Suppose $\mathbf{W} \sim \mathcal{GW}\left(\boldsymbol{\Sigma}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$ and consider the mapping $\mathbf{W} \mapsto \mathbf{W}' = \mathbf{R}\mathbf{W}\mathbf{R}^T$, where $\mathbf{R} \in \mathbb{R}^{P\times P}$ is some invertible matrix. With $\mathbf{W}$ constructed as in Section 4.5, i.e. $\mathbf{W} = \mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{L}^T$, where $\mathbf{T}\mathbf{T}^T \sim \mathcal{GW}\left(\mathbf{I}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$ and $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$, we have $\mathbf{W}' = \mathbf{R}\mathbf{L}\mathbf{T}\mathbf{T}^T\mathbf{L}\mathbf{R}^T$. However, since $\mathbf{R}\mathbf{L}$ is not in general lower triangular, there is no obvious form that suggests $\mathbf{W}'$ is in general distributed as a generalised (singular) Wishart. To remedy this, we introduce more flexibility into the generalised (singular) Wishart distribution. In particular, instead of parameterising the distribution in terms of $\boldsymbol{\Sigma}$, and multiplying $\mathbf{T}$ by the Cholesky of $\boldsymbol{\Sigma}$, we both parameterise the distribution and multiply $\mathbf{T}$ with an arbitrary invertible matrix of parameters $\mathbf{A} \in \mathbb{R}^{P\times P}$. We write $\mathbf{W} = \mathbf{A}\mathbf{T}(\mathbf{A}\mathbf{T})^T \sim \text{A-}\mathcal{GW}\left(\mathbf{A}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$ and say that $\mathbf{W}$ is A-generalised (singular) Wishart distributed. If $\mathbf{W}$ is A-generalised (singular) Wishart distributed, it is clear that for any transformation $\mathbf{R}$, the associated Gram matrix $\mathbf{W}'$ remains in the same family of distributions; in particular, $\mathbf{W}' = \mathbf{R}\mathbf{W}\mathbf{R}^T \sim \text{A-}\mathcal{GW}\left(\mathbf{R}\mathbf{A}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$.

We can understand $\mathbf{A}$ in $\mathbf{W} = \mathbf{A}\mathbf{T}(\mathbf{A}\mathbf{T})^T$ as mixing the rows of $\mathbf{T}$ by linear combinations. This mixing means that the elements of $\mathbf{A}\mathbf{T}$ have a more complex dependency structure than the elements of $\mathbf{T}$. It also raises the question of whether we could introduce a more complex dependency structure still. We propose that this can be done by additionally mixing the columns of $\mathbf{T}$ with a matrix $\mathbf{B}$, via $\mathbf{T}\mathbf{B}$, suggesting an additional generalisation of the generalised (singular) Wishart distribution. We write $\mathbf{W} = \mathbf{A}\mathbf{T}\mathbf{B}(\mathbf{A}\mathbf{T}\mathbf{B})^T \sim \text{AB-}\mathcal{GW}\left(\mathbf{A}, \mathbf{B}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$, where $\mathbf{B} \in \mathbb{R}^{\nu\times\nu}$ is lower triangular and invertible, and say that $\mathbf{W}$ is AB-generalised (singular) Wishart distributed.

To use the A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ distributions for VI, it is necessary to obtain expressions for their densities. Since this is non-trivial, the derivations are provided in the Appendix A.7, and we simply quote the results here. The density for the A-$\mathcal{GW}$ distribution is,

$$\text{A-}\mathcal{GW}\left(\mathbf{W};\, \mathbf{A}, \nu,\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \boldsymbol{\mu},\, \boldsymbol{\sigma}\right) = \frac{\left|\mathbf{W}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-N-1)/2}}{\left|\mathbf{A}\right|^{\nu}\left|(\mathbf{C}_A)_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-N-1)/2}} \prod_{j=1}^{\tilde{\nu}} \frac{\mathrm{Gamma}\left(T_{jj}^2;\, \alpha_j, \beta_j\right)}{T_{jj}^{N-j}} \prod_{i=j+1}^{N} \mathcal{N}\left(T_{ij}; \mu_{ij}, \sigma_{ij}^2\right), \tag{13}$$

where $\tilde{\nu} = \min\{\nu, N\}$, $\mathbf{C}_A = \mathbf{T}\mathbf{T}^T$, and the notation $\mathbf{X}_{:a,:b}$ means the submatrix of $\mathbf{X}$ obtained by taking the first a rows and b columns. The density for the AB-$\mathcal{GW}$ distribution is,

$$\text{AB-}\mathcal{GW}\left(\mathbf{W};\, \mathbf{A}, \mathbf{B}, \nu,\, \boldsymbol{\alpha},\, \boldsymbol{\beta},\, \boldsymbol{\mu},\, \boldsymbol{\sigma}\right) = \frac{\left|\mathbf{W}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-N-1)/2}}{\left|\mathbf{A}\right|^{\nu}\left|(\mathbf{C}_{AB})_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-N-1)/2}} \prod_{j=1}^{\tilde{\nu}} \frac{\mathrm{Gamma}\left(T_{jj}^2;\, \alpha_j, \beta_j\right)}{T_{jj}^{N-j} B_{jj}^{2(N-j+1)}} \prod_{i=j+1}^{N} \mathcal{N}\left(T_{ij}; \mu_{ij}, \sigma_{ij}^2\right), \tag{14}$$

where $\mathbf{C}_{AB} = (\mathbf{T}\mathbf{B})(\mathbf{T}\mathbf{B})^T$ is defined for notational convenience. Notice that the densities are defined in terms of

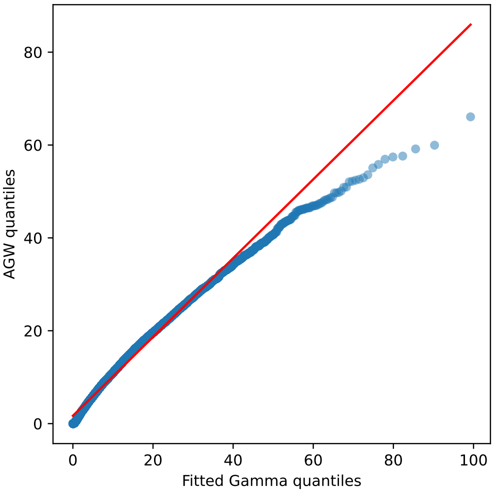

Figure 1: A probability plot comparing the probability density of the top-left element of an A-$\mathcal{GW}$ distributed matrix to the density of the top-left element of a $\mathcal{GW}$ distributed matrix. If the two were identically distributed, all the points would lie on the diagonal line. We sampled the top-left element of the A-$\mathcal{GW}$ distributed matrix 10,000 times using Eq. (18), Eq. (19) and Eq. (21). We compared this distribution against the closest fitting Gamma distribution, as the top-left element of a $\mathcal{GW}$ distributed matrix is Gamma-distributed (Sec. 5.3). We used $\mu = 3, \sigma^2 = 1$ in Eq. (18). In this probability plot, the "Fitted Gamma quantiles" are the exact quantiles of the closest fitting Gamma distribution, while the "AGW quantiles" are the sample quantiles from the top-left element of the A-$\mathcal{GW}$ distribution (i.e. the samples ordered by size). The plot shows a clear mismatch between the two distributions, confirming the theoretical result that a Gamma distribution does not capture general non-central chi-squared distributions.

both $\mathbf{W}$ and $\mathbf{T}$. Since $\mathbf{W} = (\mathbf{ATB})(\mathbf{ATB})^T$, we can see that $\mathbf{T}$ can be recovered by first computing $(\mathbf{TB})(\mathbf{TB})^T = \mathbf{A}^{-1}\mathbf{W}\mathbf{A}^{-T}$, from which we can compute $\mathbf{TB}$ as the cholesky decomposition. Thus $\mathbf{T}$ is recovered by simply right-multiplying $\mathbf{TB}$ by $\mathbf{B}^{-1}$.

5.2 A-$\mathcal{GW}$ AND AB-$\mathcal{GW}$ APPROXIMATE POSTERIORS

As discussed in Section 5.1, the A and AB-generalised (singular) Wishart distributions give us more flexible distributions over Gram matrices, which ought to be useful for VI. The approximate posterior used by Ober and Aitchison [2021a] was,

$$\begin{aligned}&\mathrm{Q}_{\mathcal{GW}}\left(\mathbf{G}_\ell \mid \mathbf{G}_{\ell-1}\right)\\&\quad= \mathcal{GW}\big(\mathbf{G}_\ell;\, (1-q_\ell)\tfrac{1}{\nu_\ell}\mathbf{K}\left(\mathbf{G}_{\ell-1}\right) + q_\ell \mathbf{V}_\ell \mathbf{V}_\ell^T,\\&\qquad\qquad \nu_\ell,\, \boldsymbol{\alpha}_\ell,\, \boldsymbol{\beta}_\ell,\, \boldsymbol{\mu}_\ell,\, \boldsymbol{\sigma}_\ell\big), \end{aligned}\tag{15}$$

where $\{\mathbf{V}_\ell, \boldsymbol{\alpha}_\ell, \boldsymbol{\beta}_\ell, \boldsymbol{\mu}_\ell, \boldsymbol{\sigma}_\ell, q_\ell\}_{\ell=1}^L$ are the learned variational parameters. Notice that $\mathbf{V}_\ell$ provides flexibility, since q_ℓ allows us to control the relative influence of the kernel from the previous layer, $\mathbf{K}(\mathbf{G}_{\ell-1})$, and an *arbitrary, learnable* positive (semi-) definite matrix, $\mathbf{V}_\ell\mathbf{V}_\ell^T$.

To use the A and AB generalised (singular) Wishart distributions as approximate posteriors, we obtain $\mathbf{A}_\ell$ by combining an arbitrary invertible matrix of parameters, $\mathbf{A}'_\ell$, with the Cholesky of $(1-q_\ell)\frac{1}{\nu_\ell}\mathbf{K}\left(\mathbf{G}_{\ell-1}\right) + q_\ell\mathbf{V}_\ell\mathbf{V}_\ell^T$ to give across-layer dependencies similar to those in the previous $\mathcal{GW}$ approximate posterior (Eq. 15),

$$\mathbf{A}_\ell = \mathrm{chol}\left((1-q_\ell)\tfrac{1}{\nu_\ell}\mathbf{K}\left(\mathbf{G}_{\ell-1}\right) + q_\ell\mathbf{V}_\ell\mathbf{V}_\ell^T\right)\mathbf{A}'_\ell \tag{16}$$

where $\mathrm{chol}(\cdot)$ returns the lower triangular Cholesky factor. The A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ approximate posteriors are then written in terms of this $\mathbf{A}_\ell$:

$$\mathrm{Q}_{\text{A-}\mathcal{GW}}\left(\mathbf{G}_\ell \mid \mathbf{G}_{\ell-1}\right) \tag{17a}$$
$$= \text{A-}\mathcal{GW}\big(\mathbf{G}_\ell;\, \mathbf{A}_\ell,\, \nu_\ell,\, \boldsymbol{\alpha}_\ell,\, \boldsymbol{\beta}_\ell,\, \boldsymbol{\mu}_\ell,\, \boldsymbol{\sigma}_\ell\big),$$
$$\mathrm{Q}_{\text{AB-}\mathcal{GW}}\left(\mathbf{G}_\ell \mid \mathbf{G}_{\ell-1}\right) \tag{17b}$$
$$= \text{AB-}\mathcal{GW}\big(\mathbf{G}_\ell;\, \mathbf{A}_\ell,\, \mathbf{B}_\ell,\, \nu_\ell,\, \boldsymbol{\alpha}_\ell,\, \boldsymbol{\beta}_\ell,\, \boldsymbol{\mu}_\ell,\, \boldsymbol{\sigma}_\ell\big).$$

Here, the variational parameters are $\{\mathbf{A}'_\ell, \mathbf{V}_\ell, \boldsymbol{\alpha}_\ell, \boldsymbol{\beta}_\ell, \boldsymbol{\mu}_\ell, \boldsymbol{\sigma}_\ell, q_\ell\}_{\ell=1}^L$ for the A-$\mathcal{GW}$ approximate posterior, and $\{\mathbf{A}'_\ell, \mathbf{B}_\ell, \mathbf{V}_\ell, \boldsymbol{\alpha}_\ell, \boldsymbol{\beta}_\ell, \boldsymbol{\mu}_\ell, \boldsymbol{\sigma}_\ell, q_\ell\}_{\ell=1}^L$ for the AB-$\mathcal{GW}$ approximate posterior.

5.3 A-$\mathcal{GW}$ IS MORE FLEXIBLE THAN $\mathcal{GW}$

To further motivate the utility of the proposed distributions, we now demonstrate that the A-$\mathcal{GW}$ family of distributions (Eq. 13) is a proper superset of the $\mathcal{GW}$ family (Eq. 12). The fact that it is a superset can be seen by noting that if we take $\mathbf{A}$ to be lower triangular, and use $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T$, then the A-$\mathcal{GW}$ distribution reduces to the $\mathcal{GW}$ distribution. Note that the AB-$\mathcal{GW}$ distribution also reduces to the previous $\mathcal{GW}$ distribution if we take $\mathbf{A}$ to be lower triangular and $\mathbf{B} = \mathbf{I}$. We do not make the claim that the AB-$\mathcal{GW}$ distribution is strictly more flexible than the A-$\mathcal{GW}$ distribution (though it clearly contains the A-$\mathcal{GW}$ by just setting $\mathbf{B} = \mathbf{I}$), and instead leave this to future work.

In order to show that the A-$\mathcal{GW}$ family contains a proper superset of $\mathcal{GW}$, we must show it contains distributions which the $\mathcal{GW}$ cannot capture. To this end, consider a toy setting where $P = 2$ and $\nu = 1$, so, $\mathbf{T} \in \mathbb{R}^{2\times 1}$. We choose

Table 1: ELBOs, test log-likelihoods, and test root mean square error for UCI datasets from [Gal and Ghahramani, 2016] for a five-layer network. All metrics are quoted as the mean, plus or minus one standard error, over the splits. Better results are highlighted; see Appendix B for other depths and additional information.

			DWP		
	Dataset	DGP	$\mathrm{Q}_{\mathcal{GW}}$	$\mathrm{Q}_{\text{A-}\mathcal{GW}}$	$\mathrm{Q}_{\text{AB-}\mathcal{GW}}$
ELBO	Boston	-0.45 ± 0.00	-0.37 ± 0.01	**-0.36 ± 0.00**	**-0.36 ± 0.00**
	Concrete	-0.50 ± 0.00	-0.49 ± 0.00	**-0.45 ± 0.00**	**-0.45 ± 0.00**
	Energy	1.38 ± 0.00	1.40 ± 0.00	**1.42 ± 0.00**	**1.41 ± 0.00**
	Kin8nm	-0.14 ± 0.00	-0.14 ± 0.00	**-0.11 ± 0.00**	**-0.11 ± 0.00**
	Naval	3.92 ± 0.04	3.59 ± 0.12	**3.97 ± 0.02**	3.63 ± 0.22
	Power	**0.03 ± 0.00**	0.02 ± 0.00	**0.03 ± 0.00**	**0.03 ± 0.00**
	Protein	**-1.00 ± 0.00**	-1.01 ± 0.00	**-1.00 ± 0.00**	**-1.00 ± 0.00**
	Wine	-1.19 ± 0.00	-1.19 ± 0.00	-1.19 ± 0.00	-1.19 ± 0.00
	Yacht	1.46 ± 0.02	1.59 ± 0.02	**1.79 ± 0.02**	**1.79 ± 0.02**
LL	Boston	-2.43 ± 0.04	**-2.38 ± 0.04**	-2.39 ± 0.05	**-2.38 ± 0.04**
	Concrete	-3.13 ± 0.02	-3.13 ± 0.02	**-3.07 ± 0.02**	-3.08 ± 0.02
	Energy	-0.71 ± 0.03	-0.71 ± 0.03	**-0.70 ± 0.03**	**-0.70 ± 0.03**
	Kin8nm	1.38 ± 0.00	1.40 ± 0.01	**1.41 ± 0.01**	**1.41 ± 0.01**
	Naval	8.28 ± 0.04	8.17 ± 0.07	**8.40 ± 0.02**	8.10 ± 0.19
	Power	-2.78 ± 0.01	-2.77 ± 0.01	**-2.76 ± 0.01**	**-2.76 ± 0.01**
	Protein	-2.73 ± 0.01	-2.72 ± 0.01	-2.71 ± 0.01	**-2.70 ± 0.00**
	Wine	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01
	Yacht	-0.73 ± 0.07	-0.58 ± 0.06	-0.22 ± 0.09	**-0.18 ± 0.07**
RMSE	Boston	2.81 ± 0.14	2.82 ± 0.17	**2.77 ± 0.16**	2.81 ± 0.17
	Concrete	5.49 ± 0.10	5.53 ± 0.10	5.26 ± 0.11	**5.24 ± 0.11**
	Energy	0.49 ± 0.01	**0.48 ± 0.01**	**0.48 ± 0.01**	**0.48 ± 0.01**
	Kin8nm	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.00	0.06 ± 0.00
	Naval	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	Power	3.88 ± 0.04	3.84 ± 0.04	**3.80 ± 0.04**	**3.80 ± 0.04**
	Protein	3.77 ± 0.02	3.76 ± 0.02	3.73 ± 0.02	**3.70 ± 0.01**
	Wine	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01
	Yacht	0.57 ± 0.05	0.50 ± 0.04	**0.37 ± 0.03**	0.38 ± 0.03

this matrix to be,

$$\mathbf{T} := \begin{pmatrix} g \\ n \end{pmatrix} \sim \begin{pmatrix} \mathcal{N}(0,1) \\ \mathcal{N}\left(\mu, \sigma^2\right) \end{pmatrix} \tag{18}$$

so that $g^2 \sim \chi^2(1) = \text{Gamma}\left(\frac{1}{2}, 2\right)$. Then we have,

$$\mathbf{W} = \mathbf{AT}(\mathbf{AT})^T \sim \text{A-}\mathcal{GW}\left(\mathbf{A}, 1, \frac{1}{2}, 2, \mu, \sigma^2\right) \tag{19}$$

where $\mathbf{W} \in \mathbb{R}^{2\times 2}$ and $\mathbf{A} \in \mathbb{R}^{2\times 2}$ is any invertible matrix. We shall show that, for certain choices of $\mathbf{A}$, $\mathbf{W}$ has a distribution that cannot be captured by the $\mathcal{GW}$ distribution. Firstly, we have,

$$\mathbf{TT}^T = \begin{pmatrix} g^2 & gn \\ gn & n^2 \end{pmatrix}, \tag{20}$$

and taking $\mathbf{A}$ to have concrete value,

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \tag{21}$$

we obtain,

$$(\mathbf{AT})(\mathbf{AT})^T = \begin{pmatrix} g^2 + 2gn + n^2 & gn + n^2 \\ gn + n^2 & n^2 \end{pmatrix}. \tag{22}$$

We need only consider the distribution of the top-left element. We have $g^2 + 2gn + n^2 = (g+n)^2 = X^2$ where $X \sim \mathcal{N}\left(\mu, \sigma^2 + 1\right)$, since g and n are normally distributed. Hence if we choose $\mu \neq 0$, the top-left element, X^2, is noncentral chi-squared distributed.

To conclude the proof, we show that the top-left element of a $\mathcal{GW}$-distributed matrix is restricted to a Gamma distribution, which does not contain noncentral chi-squared distributions. In particular, taking $\mathbf{\Sigma} = \mathbf{LL}^T$ to be the cholesky decomposition of $\mathbf{\Sigma}$, any $\mathcal{GW}$-distributed matrix,

$$\mathbf{W}' = (\mathbf{LT}')(\mathbf{LT}')^T \sim \mathcal{GW}\left(\mathbf{\Sigma}, \nu', \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\mu}', \boldsymbol{\sigma}'^2\right), \tag{23}$$

can be written according to the generalised Bartlett decomposition in Eq. 11. To obtain the top-left element of $\mathbf{W}'$, we

Table 2: Average runtime (seconds) for an epoch of BOSTON and PROTEIN. Error bars were negligible and are excluded.

{dataset} - {depth}	DGP	$\mathrm{Q}_{\mathcal{GW}}$	$\mathrm{Q}_{\text{A-}\mathcal{GW}}$	$\mathrm{Q}_{\text{AB-}\mathcal{GW}}$
BOSTON - 2	0.463	0.200	0.203	0.202
5	1.292	0.358	0.373	0.370
PROTEIN - 2	0.903	0.843	0.854	0.869
5	2.012	1.806	1.846	1.839

first explicitly write down the first element of $\mathbf{LT}'$,

$$\mathbf{LT}' = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} T'_{11} \\ T'_{21} \end{pmatrix} = \begin{pmatrix} L_{11}T'_{11} \\ \dots \end{pmatrix}, \tag{24}$$

where we avoid writing down the second element because it will not be needed. Then, we can explicitly write down the top-left element of the $\mathcal{GW}$ distributed $\mathbf{W}'$,

$$\begin{aligned} \mathbf{W}' = \mathbf{LT}'(\mathbf{LT}')^T &= \begin{pmatrix} L_{11}T'_{11} \\ \dots \end{pmatrix} \begin{pmatrix} L_{11}T'_{11} \\ \dots \end{pmatrix}^T \\ &= \begin{pmatrix} L_{11}^2 (T'_{11})^2 & \dots \\ \dots & \dots \end{pmatrix}. \end{aligned} \tag{25}$$

By Eq. 11b, we have $(T'_{11})^2 \sim \mathrm{Gamma}\,(\alpha'_1, \beta'_1)$, and so the top-left element of $\mathbf{W}$ has distribution

$$W'_{11} = L_{11}^2 (T'_{11})^2 \sim \mathrm{Gamma}\left(\alpha'_1, \frac{\beta'_1}{L_{11}^2}\right). \tag{26}$$

Since the Gamma distribution is not capable of capturing noncentral chi-square distributions, we conclude that the A-$\mathcal{GW}$ family is strictly larger than the $\mathcal{GW}$ family.

Figure 1 shows a probability plot to empirically demonstrate that the $\mathcal{GW}$ distribution is not capable of capturing the A-$\mathcal{GW}$ distribution. Specifically, we consider the same A-$\mathcal{GW}$ top-left element as in our counter-example above, using $\mu = 3, \sigma^2 = 1$ in Eq. 18, and take samples from it. We then fit a Gamma distribution to these samples, and show that there is a clear mismatch.

6 RESULTS

To compare our A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ approximate posteriors to the $\mathcal{GW}$ approximate posterior from Ober and Aitchison [2021a], we trained multiple DWPs on UCI datasets [Gal and Ghahramani, 2016] using the same architectures, only varying the approximate posterior. The algorithm for a DWP with an AB-$\mathcal{GW}$ approximate posterior is shown in Algorithm 1. The algorithm for a DWP with an A-$\mathcal{GW}$ approximate posterior is recovered by fixing $\mathbf{B}_\ell = \mathbf{I}$. The algorithm is similar to Algorithm 1 from Ober and Aitchison [2021a], but the step for sampling the inducing Gram matrix has changed (since we are using a different approximate posterior).

We also trained DGPs with the same architectures, where we used global inducing point methods from Ober and Aitchison [2021b]. All models were trained with $20\,000$ gradient steps using the ADAM optimizer [Kingma and Ba, 2015], with no pre-processing of the data other than normalizing inputs and outputs. An initial learning rate of 10^{-2} was used, and after $10\,000$ steps it was set to 10^{-3}. RMSE, ELBO and log likelihood are all reported plus or minus one standard error, calculated over 20 splits (apart from the PROTEIN dataset, where 5 splits were used). Results are shown for a 5-layer architecture in Table 1, and results for 2, 3, and 4 layers can be found in Appendix B. Layer widths ν_l in all layers were set to the number of features in the input data.

Taking the standard errors into account, we see that A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ approximate posteriors are uniformly as good or better than $\mathcal{GW}$ approximate posteriors across all metrics in the 5-layer case, and this is the case for almost all the experiments we ran (see Appendix B). Notably, A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ approximate posteriors are able to achieve higher ELBO and log likelihoods (this is expected since they provide more flexible approximate posteriors). The largest improvements in ELBO are for YACHT and NAVAL, and in the case of YACHT this leads to a large gain in RMSE.

Note that the dataset size varies, with the smallest being YACHT with 308 observations, and the largest being PROTEIN with 45730 observations. Training times per epoch for a large and a small dataset, PROTEIN and BOSTON (506 observations) can be found in Table 2. The results show that training time for the models with A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ approximate posteriors is very similar to that of previous DWPs with $\mathcal{GW}$ approximate posteriors in Ober and Aitchison [2021a], so the additional computational cost incurred by adding the new parameters $\mathbf{A}$ and $\mathbf{B}$ is negligible. All DWP models tested trained faster than the equivalent DGP models.

7 CONCLUSION

We extended the generalised (singular) Wishart distribution, $\mathcal{GW}$, introduced by Ober and Aitchison [2021a] to the A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ distributions, which we proved (both analytically and empirically) to be strictly more flexible than the $\mathcal{GW}$ distribution. These A- and AB-generalisations of the Wishart distribution are effective when used as approx-

Algorithm 1 Computing predictions/ELBO for a DWP with AB-$\mathcal{GW}$ variational posterior.

```
**Hyperparameters:** $\{\nu_\ell\}_{\ell=1}^L$
**Learned** $\mathrm{Q}$ **parameters:**
  $\{\mathbf{A}'_\ell, \mathbf{B}_\ell, \mathbf{V}_\ell, \boldsymbol{\alpha}_\ell, \boldsymbol{\beta}_\ell, \boldsymbol{\mu}_\ell, \boldsymbol{\sigma}_\ell, q_\ell\}_{\ell=1}^L$
**Other learned parameters:** $\mathbf{X}_\mathrm{i}$
**Inputs:** $\mathbf{X}_\mathrm{t}$; **Targets:** $\mathbf{Y}$

$\mathbf{X} = \begin{pmatrix}\mathbf{X}_\mathrm{i} & \mathbf{X}_\mathrm{t}\end{pmatrix}$
$\mathbf{G}_0 = \frac{1}{\nu_0}\mathbf{X}\mathbf{X}^T$
**for** $\ell$ **in** $\{1, \dots, L\}$ **do**
  $\mathbf{S} \leftarrow \frac{1}{\nu_\ell}\mathbf{K}(\mathbf{G}_{\ell-1})$
  apply Eq. (16)
  $\mathbf{A}_l \leftarrow \text{chol}\left((1 - q_\ell)\mathbf{S}_\mathrm{ii} + q_\ell \mathbf{V}_\ell \mathbf{V}_\ell^T\right) \mathbf{A}_\ell{}'$
  sample inducing Gram matrix
  $(\mathbf{A}_\ell \mathbf{T}_\ell \mathbf{B}_\ell)(\mathbf{A}_\ell \mathbf{T}_\ell \mathbf{B}_\ell)^T = \mathbf{G}_\mathrm{ii}^\ell \sim \mathrm{Q}\left(\mathbf{G}_\mathrm{ii}^\ell | \mathbf{G}_\mathrm{ii}^{\ell-1}\right)$
  update ELBO
  $\mathcal{L} \leftarrow \mathcal{L} + \log \mathrm{P}\left(\mathbf{G}_\mathrm{ii}^\ell | \mathbf{G}_\mathrm{ii}^{\ell-1}\right) - \log \mathrm{Q}\left(\mathbf{G}_\mathrm{ii}^\ell | \mathbf{G}_\mathrm{ii}^{\ell-1}\right)$
  sample full Gram matrix from conditional prior
  $\mathbf{S}_{\mathrm{tt}\cdot\mathrm{i}} \leftarrow \mathbf{S}_\mathrm{tt} - \mathbf{S}_\mathrm{it}^T \mathbf{S}_\mathrm{ii}^{-1} \mathbf{S}_\mathrm{it}$
  $\mathbf{F}_\mathrm{i}^\ell \leftarrow \mathbf{A}_\ell \mathbf{T}_\ell \mathbf{B}_\ell$
  $\mathbf{F}_\mathrm{t}^\ell \sim \mathcal{MN}\left(\mathbf{S}_\mathrm{ti}^T \mathbf{S}_\mathrm{ii}^{-1} \mathbf{F}_\mathrm{i}, \mathbf{S}_{\mathrm{tt}\cdot\mathrm{i}}, \mathbf{I}\right)$
  $\mathbf{G}_\ell = \begin{pmatrix} \mathbf{G}_\mathrm{ii}^\ell & \mathbf{F}_\mathrm{i}^\ell (\mathbf{F}_\mathrm{t}^\ell)^T \\ \mathbf{F}_\mathrm{t}^\ell (\mathbf{F}_\mathrm{i}^\ell)^T & \mathbf{F}_\mathrm{t}^\ell (\mathbf{F}_\mathrm{t}^\ell)^T \end{pmatrix}$
**end for**
sample GP inducing outputs and update ELBO
$\mathbf{F}_\mathrm{i}^{L+1} \sim \mathrm{Q}\left(\mathbf{F}_\mathrm{i}^{L+1} | \mathbf{G}_\mathrm{ii}^L\right)$
$\mathcal{L} \leftarrow \mathcal{L} + \log \mathrm{P}\left(\mathbf{F}_\mathrm{i}^{L+1} | \mathbf{G}_\mathrm{ii}^L\right) - \log \mathrm{Q}\left(\mathbf{F}_\mathrm{i}^{L+1} | \mathbf{G}_\mathrm{ii}^L\right)$
sample GP predictions conditioned on inducing points
$\mathbf{F}_\mathrm{t}^{L+1} \sim \mathrm{Q}\left(\mathbf{F}_\mathrm{t}^{L+1} | \mathbf{G}^L, \mathbf{F}_\mathrm{i}^{L+1}\right)$
add likelihood to ELBO
$\mathcal{L} \leftarrow \mathcal{L} + \log \mathrm{P}\left(\mathbf{Y} | \mathbf{F}_\mathrm{t}^{L+1}\right)$
```

imate posteriors for DWPs, as shown by the near-universal improvement in predictive performance on UCI datasets, both over similar DGP models, and over DWP models that use the less flexible $\mathcal{GW}$ distribution for their approximate posteriors. Furthermore, we showed that this increased flexibility comes at a negligible additional cost in computation. As a result, this is the first DWP work to achieve equal-or-better predictive performance than comparable DGPs on UCI datasets (and it is also cheaper to train). This is significant as DGP priors are equivalent to DWP priors [Aitchison et al., 2021], but DWP posteriors are invariant to certain types of posterior symmetries that affect DGPs, meaning they should in theory be easier to capture under variational inference, but until this work practical results had not shown this to be the case.

References

Laurence Aitchison. Why bigger is not always better: on finite and infinite neural networks. In *International Conference on Machine Learning*, pages 156–164. PMLR, 2020.

Laurence Aitchison, Adam Yang, and Sebastian W Ober. Deep kernel processes. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 130–140. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/aitchison21a.html.

Maurice Stevenson Bartlett. On the Theory of Statistical Regression. *Proceedings of the Royal Society of Edinburgh*, 53:260–283, 1933.

Andreas Damianou and Neil Lawrence. Deep Gaussian Processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059. JMLR.org, 2016.

Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Edward Milsom, Ben Anson, and Laurence Aitchison. Convolutional deep kernel machines. 2023.

Roman Novak, Lechao Xiao, Jaehoon Lee, Yasaman Bahri, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Bayesian deep convolutional networks with many channels are gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

Sebastian Ober and Laurence Aitchison. A variational approximate posterior for the deep Wishart process. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6567–6579. Curran Associates, Inc., 2021a. URL `https://proceedings.neurips.cc/paper/2021/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf`.

Sebastian W Ober and Laurence Aitchison. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8248–8259. PMLR, 18–24 Jul 2021b. URL `https://proceedings.mlr.press/v139/ober21a.html`.

M.S. Srivastava. Singular Wishart and multivariate beta distributions. *The Annals of Statistics*, 31(5):1537 – 1560, 2003. doi: 10.1214/aos/1065705118. URL `https://doi.org/10.1214/aos/1065705118`.

Adam X Yang, Maxime Robeyns, Edward Milsom, Ben Anson, Nandi Schoots, and Laurence Aitchison. A theory of representation learning in deep neural networks gives a deep generalisation of kernel methods. *International Conference on Machine Learning*, 2023.

Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.