Deep kernel processes

Laurence Aitchison [1] **Adam X. Yang** [1] **Sebastian W. Ober** [2]

Abstract

We define deep kernel processes in which positive definite Gram matrices are progressively transformed by nonlinear kernel functions and by sampling from (inverse) Wishart distributions. Remarkably, we find that deep Gaussian processes (DGPs), Bayesian neural networks (BNNs), infinite BNNs, and infinite BNNs with bottlenecks can all be written as deep kernel processes. For DGPs the equivalence arises because the Gram matrix formed by the inner product of features is Wishart distributed, and as we show, standard isotropic kernels can be written entirely in terms of this Gram matrix — we do not need knowledge of the underlying features. We define a tractable deep kernel process, the deep inverse Wishart process, and give a doubly-stochastic inducing-point variational inference scheme that operates on the Gram matrices, not on the features, as in DGPs. We show that the deep inverse Wishart process gives superior performance to DGPs and infinite BNNs on fully-connected baselines.[1]

1. Introduction

The deep learning revolution has shown us that effective performance on difficult tasks such as image classification (Krizhevsky et al., 2012) requires deep models with flexible lower-layers that learn task-dependent representations. Here, we consider whether these insights from the neural network literature can be applied to purely kernel-based methods. (Note that we do not consider deep Gaussian processes or DGPs to be "fully kernel-based" as they use a feature-based representation in intermediate layers).

Importantly, deep kernel methods (e.g. Cho & Saul, 2009) already exist. In these methods, which are closely related to infinite Bayesian neural networks (Lee et al., 2017; Matthews et al., 2018; Garriga-Alonso et al., 2018; Novak et al., 2018), we take an initial kernel (usually the dot product of the input features) and perform a series of deterministic, parameter-free transformations to obtain an output kernel that we use in e.g. a support vector machine or Gaussian process. However, the deterministic, parameter-free nature of the transformation from input to output kernel means that they lack the capability to learn a top-layer representation, which is believed to be crucial for the effectiveness of deep methods (Aitchison, 2019).

2. Contributions

1. We propose deep kernel processes (DKPs), which combine nonlinear transformations of the kernel, as in Cho & Saul (2009) with a flexible learned representation by exploiting a Wishart or inverse Wishart process (Dawid, 1981; Shah et al., 2014).

2. We show that models ranging from DGPs (Damianou & Lawrence, 2013; Salimbeni & Deisenroth, 2017) to Bayesian neural networks (BNNs; Blundell et al., 2015, App. C.1), infinite BNNs (App. C.2) and infinite BNNs with bottlenecks (App. C.3) can be written as DKPs.

3. We define a specific DKP, the deep inverse Wishart process (DIWP) which offers convenient variational approximate posteriors.

4. We develop a novel doubly-stochastic variational inducing-point inference scheme purely in the kernel domain (as opposed to Salimbeni & Deisenroth, 2017, who described DSVI for standard feature-based DGPs) for DIWPs.

5. We demonstrate improved performance of DIWPs on fully-connected benchmark datasets.

DKPs and specifically DIWPs offer two key advantages over feature-based methods such as DGPs and BNNs. First, DGPs and BNNs have complex approximate posteriors (Li et al., 2018), due in part to permuation/rotation symmetries in the posterior over weights/features (App. D.1 and D.2; MacKay, 1992; Moore, 2016; Pourzanjani et al., 2017).

[1]Department of Computer Science, Bristol, BS8 1UB, UK [2]Department of Engineering, Cambridge, CB2 1PZ, UK. Correspondence to: Laurence Aitchison <laurence.aitchison@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]Reference implementation at: `github.com/LaurenceA/bayesfunc`

This complexity means that common variational approximate posteriors can give a very poor approximation to the true posterior. In contrast, the Gram matrices in DKPs are invariant to permutations/rotations of the weights/features and thus have much simpler true posteriors which are more easily captured by variational approximate posteriors. Second, in DIWPs the "width" parameter is learnable, and in the limit of infinite width gives a series of deterministic kernel transformations, as in an infinite neural network. This gives DIWPs the ability to learn on a layer-by-layer basis where a deterministic kernel transformation is appropriate, or where more flexibility in the kernel is needed.

3. Background

We briefly revise Wishart and inverse Wishart distributions. The Wishart distribution is a generalization of the gamma distribution that is defined over positive semidefinite matrices. Suppose that we have a collection of P-dimensional random variables $\mathbf{x}_i$ with $i \in \{1, \dots, N\}$ such that

$$\mathbf{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{V}), \tag{1}$$

$$\textstyle\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T = \mathbf{S} \sim \mathcal{W}(\mathbf{V}, N) \tag{2}$$

has Wishart distribution with scale matrix $\mathbf{V}$ and N degrees of freedom. When $N > P - 1$, the density is,

$$\mathcal{W}(\mathbf{S}; \mathbf{V}, N) = \frac{|\mathbf{S}|^{(N-P-1)/2} e^{-\operatorname{Tr}\left(\mathbf{V}^{-1}\mathbf{S}/2\right)}}{2^{NP}|\mathbf{V}|\Gamma_P\left(\frac{N}{2}\right)} \tag{3}$$

where Γ_P is the multivariate gamma function. Further, the inverse, $\mathbf{S}^{-1}$ has inverse Wishart distribution, $\mathcal{W}^{-1}\left(\mathbf{V}^{-1}, N\right)$. The inverse Wishart is defined only for $N > P - 1$ and also has closed-form density. Finally, we note that the Wishart distribution has mean $N\mathbf{V}$ while the inverse Wishart has mean $\mathbf{V}^{-1}/(N-P-1)$ (for $N > P+1$).

4. Deep kernel processes

We define a kernel process to be a set of distributions over positive definite matrices of different sizes, that are consistent under marginalisation (Dawid, 1981; Shah et al., 2014). The two most common kernel processes are the Wishart process and inverse Wishart process, which we write in a slightly unusual form to ensure their expectation is $\mathbf{K}$. We take $\mathbf{G}$ and $\mathbf{G}'$ to be finite dimensional marginals of the underlying Wishart and inverse Wishart process,

$$\mathbf{G} \sim \mathcal{W}(\mathbf{K}/N, N), \tag{4a}$$

$$\mathbf{G}^* \sim \mathcal{W}(\mathbf{K}^*/N, N), \tag{4b}$$

$$\mathbf{G}' \sim \mathcal{W}^{-1}(\delta\mathbf{K}, \delta+(P+1)), \tag{4c}$$

$$\mathbf{G}'^* \sim \mathcal{W}^{-1}(\delta\mathbf{K}^*, \delta+(P^*+1)), \tag{4d}$$

and where we explicitly give the consistent marginal distributions over $\mathbf{K}^*$, $\mathbf{G}^*$ and $\mathbf{G}'^*$ which are $P^* \times P^*$ principal submatrices of the $P \times P$ matrices $\mathbf{K}$, $\mathbf{G}$ and $\mathbf{G}'$ dropping the same rows and columns. In the inverse-Wishart distribution, δ is a positive parameter that can be understood as controlling the degree of variability, with larger values for δ implying smaller variability in $\mathbf{G}'$.

We define a deep kernel process by analogy with a DGP, as a composition of kernel processes, and show in App. A that under sensible assumptions any such composition is itself a kernel process. [2]

4.1. DGPs with isotropic kernels are deep Wishart processes

We consider deep GPs of the form (Fig. 1 top) with $\mathbf{X} \in \mathbb{R}^{P \times N_0}$, where P is the number of input points and N_0 is the number of features in the input.

$$\mathbf{K}_\ell = \begin{cases} \frac{1}{N_0}\mathbf{X}\mathbf{X}^T & \text{for } \ell = 1, \\ \mathbf{K}(\mathbf{G}_{\ell-1}) & \text{for } \ell \in \{2, \dots, L+1\}, \end{cases} \tag{5a}$$

$$\mathrm{P}(\mathbf{F}_\ell | \mathbf{K}_\ell) = \textstyle\prod_{\lambda=1}^{N_\ell} \mathcal{N}\left(\mathbf{f}_\lambda^\ell; \mathbf{0}, \mathbf{K}_\ell\right), \tag{5b}$$

$$\mathbf{G}_\ell = \tfrac{1}{N_\ell}\mathbf{F}_\ell\mathbf{F}_\ell^T. \tag{5c}$$

Here, $\mathbf{F}_\ell \in \mathbb{R}^{P \times N_\ell}$ are the N_ℓ hidden features in layer ℓ; λ indexes hidden features so $\mathbf{f}_\lambda^\ell$ is a single column of $\mathbf{F}_\ell$, representing the value of the λth feature for all training inputs. Note that $\mathbf{K}(\cdot)$ is a function that takes a Gram matrix and returns a kernel matrix, whereas $\mathbf{K}_\ell$ is a (possibly random) variable representing the kernel matrix at layer ℓ. Note, we have restricted ourselves to kernels that can be written as functions of the Gram matrix, $\mathbf{G}_\ell$, and do not require the full set of activations, $\mathbf{F}_\ell$. As we describe later, this is not too restrictive, as it includes amongst others all isotropic kernels (i.e. those that can be written as a function of the distance between points Williams & Rasmussen, 2006). Note that we have a number of choices as to how to initialize the kernel in Eq. (5a). The current choice just uses a linear dot-product kernel, rather than immediately applying the kernel function $\mathbf{K}$. This is both to ensure exact equivalence with infinite NNs with bottlenecks (App. C.3) and also to highlight an interesting interpretation of this layer as Bayesian inference over generalised lengthscale hyperparameters in the squared-exponential kernel (App. B e.g. Lalchand & Rasmussen, 2020).

For DGP regression, the outputs, $\mathbf{Y}$, are most commonly given by a likelihood that can be written in terms of the output features, $\mathbf{F}_{L+1}$. For instance, for regression, the

[2] Note that we leave the question of the full Kolmogorov extension theorem (Kolmogorov, 1933) for matrices to future work: for our purposes, it is sufficient to work with very large but ultimately finite input spaces as in practice, the input vectors are represented by elements of the finite set of 32-bit or 64-bit floating-point numbers (Sterbenz, 1974).

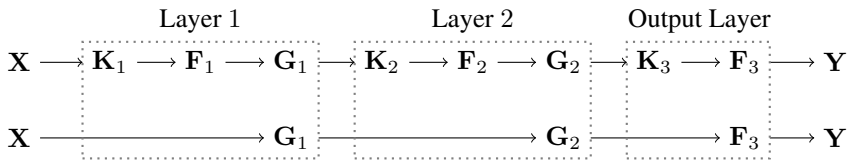

Figure 1. Generative models for two layer ($L = 2$) deep GPs. (**Top**) Generative model for a deep GP, with a kernel that depends on the Gram matrix, and with Gaussian-distributed features. (**Bottom**) Integrating out the features, the Gram matrices become Wishart distributed.

distribution of the λth output feature column could be

$$\mathrm{P}\left(\mathbf{y}_\lambda|\mathbf{F}_{L+1}\right) = \mathcal{N}\left(\mathbf{y}_\lambda; \mathbf{f}_\lambda^{L+1}, \sigma^2\mathbf{I}\right), \tag{6}$$

alternatively, we could use a classification likelihood,

$$\mathrm{P}\left(\mathbf{y}|\mathbf{F}_{L+1}\right) = \text{Categorical}\left(\mathbf{y}; \text{softmax}\left(\mathbf{F}_\lambda^{L+1}\right)\right). \tag{7}$$

Importantly, our methods can be used with any likelihood with a known probability density function.

The generative process for the Gram matrices, $\mathbf{G}_\ell$, consists of generating samples from a Gaussian distribution (Eq. 5b), and taking their product with themselves transposed (Eq. 5c). This exactly matches the generative process for a Wishart distribution (Eq. 1), so we can write the Gram matrices, $\mathbf{G}_\ell$, directly in terms of the kernel, without needing to sample features (Fig. 1 bottom),

$$\mathrm{P}\left(\mathbf{G}_1|\mathbf{X}\right) = \mathcal{W}\left(\tfrac{1}{N_1}\left(\tfrac{1}{N_0}\mathbf{X}\mathbf{X}^T\right), N_1\right), \tag{8a}$$

$$\mathrm{P}\left(\mathbf{G}_\ell|\mathbf{G}_{\ell-1}\right) = \mathcal{W}\left(\mathbf{K}\left(\mathbf{G}_{\ell-1}\right)/N_\ell, N_\ell\right), \tag{8b}$$

$$\mathrm{P}\left(\mathbf{F}_{L+1}|\mathbf{G}_L\right) = \textstyle\prod_{\lambda=1}^{N_{L+1}} \mathcal{N}\left(\mathbf{f}_\lambda^{L+1}; \mathbf{0}, \mathbf{K}\left(\mathbf{G}_L\right)\right). \tag{8c}$$

Except at the output, the model is phrased entirely in terms of positive-definite kernels and Gram matrices, and is consistent under marginalisation (assuming a valid kernel function) and is thus a DKP. At a high level, the model can be understood as alternatively sampling a Gram matrix (introducing flexibility in the representation), and nonlinearly transforming the Gram matrix using a kernel (Fig. 2).

This highlights a particularly simple interpretation of the DKP as an autoregressive process. In a standard autoregressive process, we might propagate the current vector, $\mathbf{x}_t$, through a deterministic function, $\mathbf{f}(\mathbf{x}_t)$, and add zero-mean Gaussian noise, $\boldsymbol{\xi}$,

$$\mathbf{x}_{t+1} = \mathbf{f}\left(\mathbf{x}_t\right) + \sigma^2\boldsymbol{\xi} \quad \text{such that} \quad \mathbb{E}\left[\mathbf{x}_{t+1}|\mathbf{x}_t\right] = \mathbf{f}\left(\mathbf{x}_t\right). \tag{9}$$

By analogy, the next Gram matrix has expectation centered on a deterministic transformation of the previous Gram matrix,

$$\mathbb{E}\left[\mathbf{G}_\ell|\mathbf{G}_{\ell-1}\right] = \mathbf{K}\left(\mathbf{G}_{\ell-1}\right), \tag{10}$$

so $\mathbf{G}_\ell$ can be written as this expectation plus a zero-mean random variable, $\boldsymbol{\Xi}_\ell$, that can be interpreted as noise,

$$\mathbf{G}_\ell = \mathbf{K}\left(\mathbf{G}_{\ell-1}\right) + \boldsymbol{\Xi}_\ell. \tag{11}$$

Note that $\boldsymbol{\Xi}_\ell$ is not in general positive definite, and may not have an analytically tractable distribution. This noise decreases as N_ℓ increases,

$$\begin{aligned}\mathbb{V}\left[G_{ij}^\ell\right] &= \mathbb{V}\left[\Xi_{ij}^\ell\right] \\ &= \tfrac{1}{N_\ell}\left(K_{ij}^2(\mathbf{G}_{\ell-1}) + K_{ii}^2(\mathbf{G}_{\ell-1})K_{jj}^2(\mathbf{G}_{\ell-1})\right).\end{aligned} \tag{12}$$

Notably, as N_ℓ tends to infinity, the Wishart samples converge on their expectation, and the noise disappears, leaving us with a series of deterministic transformations of the Gram matrix. Therefore, we can understand a deep kernel process as alternatively adding "noise" to the kernel by sampling e.g. a Wishart or inverse Wishart distribution ($\mathbf{G}_2$ and $\mathbf{G}_3$ in Fig. 2) and computing a nonlinear transformation of the kernel ($\mathbf{K}(\mathbf{G}_2)$ and $\mathbf{K}(\mathbf{G}_3)$ in Fig. 2)

Remember that we are restricted to kernels that can be written as a function of the Gram matrix,

$$\begin{aligned}\mathbf{K}_\ell &= \mathbf{K}\left(\mathbf{G}_\ell\right) = \mathbf{K}_{\text{features}}\left(\mathbf{F}_\ell\right), \\ K_{ij}^\ell &= k\left(\mathbf{F}_{i,:}^\ell, \mathbf{F}_{j,:}^\ell\right).\end{aligned} \tag{13}$$

where $\mathbf{K}_{\text{features}}(\cdot)$ takes a matrix of features, $\mathbf{F}_\ell$, and returns the kernel matrix, $\mathbf{K}_\ell$, and k is the usual kernel function, which takes two feature vectors (rows of $\mathbf{F}_\ell$) and returns an element of the kernel matrix. This does not include all possible kernels because it is not possible to recover the features from the Gram matrix. In particular, the Gram matrix is invariant to unitary transformations of the features: the Gram matrix is the same for $\mathbf{F}_\ell$ and $\mathbf{F}'_\ell = \mathbf{U}\mathbf{F}_\ell$ where $\mathbf{U}$ is a unitary matrix, such that $\mathbf{U}\mathbf{U}^T = \mathbf{I}$,

$$\mathbf{G}_\ell = \tfrac{1}{N_\ell}\mathbf{F}_\ell\mathbf{F}_\ell^T = \tfrac{1}{N_\ell}\mathbf{F}_\ell\mathbf{U}_\ell\mathbf{U}_\ell^T\mathbf{F}_\ell^T = \tfrac{1}{N_\ell}\mathbf{F}'_\ell\mathbf{F}'^T_\ell. \tag{14}$$

Superficially, this might seem very limiting — leaving us only with dot-product kernels (Williams & Rasmussen, 2006) such as,

$$k(\mathbf{f}, \mathbf{f}') = \mathbf{f}\cdot\mathbf{f}' + \sigma^2. \tag{15}$$

However, in reality, a far broader range of kernels fit within this class. Importantly, isotropic or radial basis function kernels including the squared exponential and Matern depend

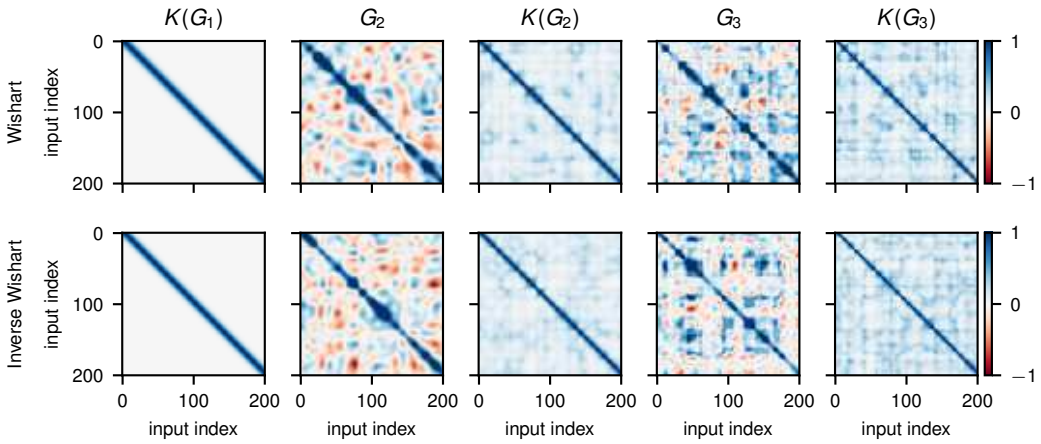

Figure 2. Visualisations of a single prior sample of the kernels and Gram matrices as they pass through the network. We use 1D, equally spaced inputs with a squared exponential kernel. As we transition $\mathbf{K}(\mathbf{G}_{\ell-1}) \to \mathbf{G}_\ell$, we add "noise" by sampling from a Wishart (top) or an inverse Wishart (bottom). As we transition from $\mathbf{G}_\ell$ to $\mathbf{K}(\mathbf{G}_\ell)$, we deterministically transform the Gram matrix using a squared-exponential kernel.

only on the squared distance between points, R, (Williams & Rasmussen, 2006)

$$k(\mathbf{f}, \mathbf{f}') = k\left(R\right), \qquad R = \left|\mathbf{f} - \mathbf{f}'\right|^2. \tag{16}$$

These kernels can be written as a function of $\mathbf{G}$, because the matrix of squared distances, $\mathbf{R}$, can be computed from $\mathbf{G}$,

$$\begin{aligned} R^\ell_{ij} &= \tfrac{1}{N_\ell}\textstyle\sum_{\lambda=1}^{N_\ell}\left(F^\ell_{i\lambda} - F^\ell_{j\lambda}\right)^2 \\ &= \tfrac{1}{N_\ell}\textstyle\sum_{\lambda=1}^{N_\ell}\left(\left(F^\ell_{i\lambda}\right)^2 - 2F^\ell_{i\lambda}F^\ell_{j\lambda} + \left(F^\ell_{j\lambda}\right)^2\right) \\ &= G^\ell_{ii} - 2G^\ell_{ij} + G^\ell_{jj}. \end{aligned} \tag{17}$$

5. Variational inference in deep kernel processes

A key part of the motivation for developing deep kernel processes was that the posteriors over weights in a BNN or over features in a deep GP are extremely complex and multimodal, with a large number of symmetries that are not captured by standard approximate posteriors (MacKay, 1992; Moore, 2016; Pourzanjani et al., 2017). For instance, in the Appendix we show that there are permutation symmetries in the prior and posteriors over weights in BNNs (App. D.1) and rotational symmetries in the prior and posterior over features in deep GPs with isotropic kernels (App. D.2). The inability to capture these symmetries in standard variational posteriors may introduce biases in the parameters inferred by variational inference, because the variational bound is not uniformly tight across the state-space (Turner & Sahani, 2011). Gram matrices are invariant to permutations or rotations of the features, so we can sidestep these complex posterior symmetries by working with the Gram matrices as the random variables in variational inference. However, variational inference in deep Wishart processes (equivalent to DGPs Sec. 4.1 and infinite NNs with bottlenecks App. C.3) is difficult because the approximate posterior we would like to use, the non-central Wishart (App. E), has a probability density function that is prohibitively costly and complex to evaluate in the inner loop of a deep learning model (Koev & Edelman, 2006). Instead, we consider an inverse Wishart process prior, for which the inverse Wishart itself makes a good choice of approximate posterior.

5.1. The deep inverse Wishart processes

By analogy with Eq. (8), we define a deep inverse Wishart processes (DIWPs). However, the inverse Wishart process introduces a new difficulty: that at the input layer, $\frac{1}{N_0}\mathbf{X}\mathbf{X}^T$ may be singular if there are more datapoints than features. Instead of attempting to use a singular Wishart distribution over $\mathbf{G}_1$, which would be complex and difficult to work with (Bodnar & Okhrin, 2008; Bodnar et al., 2016), we instead define an approximate posterior over the full-rank $N_0 \times N_0$ matrix, $\mathbf{\Omega}$, and use $\mathbf{G}_1 = \frac{1}{N_0}\mathbf{X}\mathbf{\Omega}\mathbf{X}^T \in \mathbb{R}^{P\times P}$.

$$\begin{aligned} \mathrm{P}\left(\mathbf{\Omega}\right) &= \mathcal{W}^{-1}(\delta_1\mathbf{I}, \delta_1{+}N_0{+}1)\,, \\ \left(\text{with } \mathbf{G}_1 &= \tfrac{1}{N_0}\mathbf{X}\mathbf{\Omega}\mathbf{X}^T,\right) \\ \mathrm{P}\left(\mathbf{G}_{\ell\in\{2\ldots L\}}|\mathbf{G}_{\ell-1}\right) &= \mathcal{W}^{-1}(\delta_\ell\mathbf{K}\left(\mathbf{G}_{\ell-1}\right), P{+}1{+}\delta_\ell)\,, \\ \mathrm{P}\left(\mathbf{F}_{L+1}|\mathbf{G}_L\right) &= \textstyle\prod_{\lambda=1}^{N_{L+1}}\mathcal{N}\left(\mathbf{f}^{L+1}_\lambda; \mathbf{0}, \mathbf{K}\left(\mathbf{G}_L\right)\right), \end{aligned} \tag{18}$$

remembering that $\mathbf{X} \in \mathbb{R}^{P\times N_0}$, $\mathbf{G}_\ell \in \mathbb{R}^{P\times P}$ and $\mathbf{F}_\ell \in \mathbb{R}^{P\times N_{L+1}}$.

Critically, the distributions in Eq. (18) are consistent under marginalisation as long as δ_ℓ is held constant (Dawid, 1981), with P taken to be the number of input points, or equivalently the size of $\mathbf{K}_{\ell-1}$. Further, the deep inverse Wishart process retains the interpretation as a deterministic trans-

formation of the kernel plus noise because the expectation is,

$$\mathbb{E}\left[\mathbf{G}_\ell|\mathbf{G}_{\ell-1}\right] = \frac{\delta_\ell \mathbf{K}\left(\mathbf{G}_{\ell-1}\right)}{(P+1+\delta_\ell)-(P+1)} = \mathbf{K}\left(\mathbf{G}_{\ell-1}\right). \tag{19}$$

The resulting inverse Wishart process does not have a direct interpretation as e.g. a deep GP, but does have more appealing properties for variational inference, as it is always full-rank and allows independent control over the approximate posterior mean and variance. Finally, it is important to note that Wishart and inverse Wishart distributions do not differ as much as one might expect; the standard Wishart and standard inverse Wishart distributions have isotropic distributions over the eigenvectors so they only differ in terms of their distributions over eigenvalues, and these are often quite similar, especially if we consider a Wishart model with ResNet-like structure (App. H).

5.2. An approximate posterior for the deep inverse Wishart process

Choosing an appropriate and effective form for variational approximate posteriors is usually a difficult research problem. Here, we take inspiration from Ober & Aitchison (2020) by exploiting the fact that the inverse-Wishart distribution is the conjugate prior for the covariance matrix of a multivariate Gaussian. In particular, if we consider an inverse-Wishart prior over $\mathbf{\Sigma} \in \mathbb{R}^{P\times P}$ with mean $\mathbf{\Sigma}_0$, which forms the covariance of Gaussian-distributed matrix, $\mathbf{V} \in \mathbb{R}^{P\times P}$, consisting of columns $\mathbf{v}_\lambda$, then the posterior over $\mathbf{\Sigma}$ is also inverse-Wishart,

$$\mathrm{P}\left(\mathbf{\Sigma}\right) = \mathcal{W}^{-1}(\mathbf{\Sigma}; \delta\mathbf{\Sigma}_0, P{+}1{+}\delta), \tag{20a}$$

$$\mathrm{P}\left(\mathbf{V}|\mathbf{\Sigma}\right) = \textstyle\prod_{\lambda=1}^{N_V}\mathcal{N}\left(\mathbf{v}_\lambda; \mathbf{0}, \mathbf{\Sigma}\right), \tag{20b}$$

$$\mathrm{P}\left(\mathbf{\Sigma}|\mathbf{V}\right) = \mathcal{W}^{-1}\left(\delta\mathbf{\Sigma}_0 + \mathbf{V}\mathbf{V}^T, P{+}1{+}\delta{+}N_V\right). \tag{20c}$$

Inspired by this exact posterior that is available in simple models, we choose the approximate posterior in our model to be,

$$\begin{aligned}
&\mathrm{Q}\left(\mathbf{\Omega}\right) = \mathcal{W}^{-1}\left(\delta_1\mathbf{I} + \mathbf{V}_1\mathbf{V}_1^T, \delta_1{+}\gamma_1{+}N_0{+}1\right),\\
&\quad(\text{with } \mathbf{G}_1 = \tfrac{1}{N_0}\mathbf{X}\mathbf{\Omega}\mathbf{X}^T,)\\
&\mathrm{Q}\left(\mathbf{G}_\ell|\mathbf{G}_{\ell-1}\right) =\\
&\quad\mathcal{W}^{-1}\left(\delta_\ell\mathbf{K}\left(\mathbf{G}_{\ell-1}\right){+}\mathbf{V}_\ell\mathbf{V}_\ell^T, \delta_\ell{+}\gamma_\ell{+}P{+}1\right),\\
&\mathrm{Q}\left(\mathbf{F}_{L+1}|\mathbf{G}_L\right) = \textstyle\prod_{\lambda=1}^{N_{L+1}}\mathcal{N}\left(\mathbf{f}_\lambda^{L+1}; \mathbf{\Sigma}_\lambda\mathbf{\Lambda}_\lambda\mathbf{v}_\lambda, \mathbf{\Sigma}_\lambda\right),\\
&\quad\text{where } \mathbf{\Sigma}_\lambda = \left(\mathbf{K}^{-1}\left(\mathbf{G}_L\right) + \mathbf{\Lambda}_\lambda\right)^{-1},
\end{aligned} \tag{21}$$

and where $\mathbf{V}_1$ is a learned $N_0 \times N_0$ matrix, $\{\mathbf{V}_\ell\}_{\ell=2}^L$ are $P \times P$ learned matrices and γ_ℓ are learned non-negative real numbers. For more details about the input layer, see App. F. At the output layer, we use the global inducing approximate posterior for DGPs from Ober & Aitchison (2020), with learned parameters being vectors, $\mathbf{v}_\lambda$, and positive definite matrices, $\mathbf{\Lambda}_\lambda$ (see App. G).

In summary, the prior has parameters δ_ℓ (which also appears in the approximate posterior), and the posterior has parameters $\mathbf{V}_\ell$ and γ_ℓ for the inverse-Wishart hidden layers, and $\{\mathbf{v}_\lambda\}_{\lambda=1}^{N_{L+1}}$ and $\{\mathbf{\Lambda}_\lambda\}_{\lambda=1}^{N_{L+1}}$ at the output. In all our experiments, we optimize all five parameters $\{\delta_\ell, \mathbf{V}_\ell, \gamma_\ell\}_{\ell=1}^L$ and $(\{\mathbf{v}_\lambda, \mathbf{\Lambda}_\lambda\}_{\lambda=1}^{N_{L+1}})$, and in addition, for inducing-point methods, we also optimize a single set of "global" inducing inputs, $\mathbf{X}_\mathrm{i} \in \mathbb{R}^{P_\mathrm{i}\times N_0}$, which are defined only at the input layer.

5.3. Doubly stochastic inducing-point variational inference in deep inverse Wishart processes

For efficient inference in high-dimensional problems, we take inspiration from the DGP literature (Salimbeni & Deisenroth, 2017) by considering doubly-stochastic inducing-point deep inverse Wishart processes. We begin by decomposing all variables into inducing and training (or test) points $\mathbf{X}_t \in \mathbb{R}^{P_t\times N_0}$,

$$\mathbf{X} = \begin{pmatrix}\mathbf{X}_\mathrm{i}\\ \mathbf{X}_\mathrm{t}\end{pmatrix} \quad \mathbf{F}_{L+1} = \begin{pmatrix}\mathbf{F}_\mathrm{i}^{L+1}\\ \mathbf{F}_\mathrm{t}^{L+1}\end{pmatrix} \quad \mathbf{G}_\ell = \begin{pmatrix}\mathbf{G}_\mathrm{ii}^\ell & \mathbf{G}_\mathrm{it}^\ell\\ \mathbf{G}_\mathrm{ti}^\ell & \mathbf{G}_\mathrm{tt}^\ell\end{pmatrix} \tag{22}$$

where e.g. $\mathbf{G}_\mathrm{ii}^\ell$ is $P_\mathrm{i} \times P_\mathrm{i}$ and $\mathbf{G}_\mathrm{it}^\ell$ is $P_\mathrm{i} \times P_\mathrm{t}$ where P_i is the number of inducing points, and P_t is the number of testing/training points. Note that $\mathbf{\Omega}$ does not decompose as it is $N_0 \times N_0$. The full ELBO including latent variables for all the inducing and training points is,

$$\mathcal{L} = \mathbb{E}\left[\log \mathrm{P}\left(\mathbf{Y}|\mathbf{F}_{L+1}\right) + \log\frac{\mathrm{P}\left(\mathbf{\Omega}, \{\mathbf{G}_\ell\}_{\ell=2}^L, \mathbf{F}_{L+1}|\mathbf{X}\right)}{\mathrm{Q}\left(\mathbf{\Omega}, \{\mathbf{G}_\ell\}_{\ell=2}^L, \mathbf{F}_{L+1}|\mathbf{X}\right)}\right] \tag{23}$$

where the expectation is taken over $\mathrm{Q}\left(\mathbf{\Omega}, \{\mathbf{G}_\ell\}_{\ell=2}^L, \mathbf{F}_{L+1}|\mathbf{X}\right)$. The prior is given by combining all terms in Eq. (18) for both inducing and test/train inputs,

$$\begin{aligned}
&\mathrm{P}\left(\mathbf{\Omega}, \{\mathbf{G}_\ell\}_{\ell=2}^L, \mathbf{F}_{L+1}|\mathbf{X}\right) =\\
&\quad\mathrm{P}\left(\mathbf{\Omega}\right)\left[\textstyle\prod_{\ell=2}^L \mathrm{P}\left(\mathbf{G}_\ell|\mathbf{G}_{\ell-1}\right)\right]\mathrm{P}\left(\mathbf{F}_{L+1}|\mathbf{G}_L\right),
\end{aligned} \tag{24}$$

where the $\mathbf{X}$-dependence enters on the right because $\mathbf{G}_1 = \frac{1}{N_0}\mathbf{X}\mathbf{\Omega}\mathbf{X}^T$. Taking inspiration from Salimbeni & Deisenroth (2017), the full approximate posterior is the product of an approximate posterior over inducing points and the conditional prior for train/test points,

$$\begin{aligned}
&\mathrm{Q}\left(\mathbf{\Omega}, \{\mathbf{G}_\ell\}_{\ell=2}^L, \mathbf{F}_{L+1}|\mathbf{X}\right) =\\
&\quad\mathrm{Q}\left(\mathbf{\Omega}, \{\mathbf{G}_\mathrm{ii}^\ell\}_{\ell=2}^L, \mathbf{F}_\mathrm{i}^{L+1}|\mathbf{X}_\mathrm{i}\right)\\
&\quad\mathrm{P}\left(\{\mathbf{G}_\mathrm{it}^\ell\}_{\ell=2}^L, \{\mathbf{G}_\mathrm{tt}^\ell\}_{\ell=2}^L, \mathbf{F}_\mathrm{t}^{L+1}|\mathbf{\Omega}, \{\mathbf{G}_\mathrm{ii}^\ell\}_{\ell=2}^L, \mathbf{F}_\mathrm{i}^{L+1}, \mathbf{X}\right)
\end{aligned} \tag{25}$$

and the prior can be written in the same form,

$$\mathrm{P}\left(\boldsymbol{\Omega}, \{\mathbf{G}_\ell\}_{\ell=2}^L, \mathbf{F}_{L+1}|\mathbf{X}\right) = \mathrm{P}\left(\boldsymbol{\Omega}, \{\mathbf{G}_{\mathrm{ii}}^\ell\}_{\ell=2}^L, \mathbf{F}_{\mathrm{i}}^{L+1}|\mathbf{X}_{\mathrm{i}}\right) \mathrm{P}\left(\{\mathbf{G}_{\mathrm{it}}^\ell\}_{\ell=2}^L, \{\mathbf{G}_{\mathrm{tt}}^\ell\}_{\ell=2}^L, \mathbf{F}_{\mathrm{t}}^{L+1}|\boldsymbol{\Omega}, \{\mathbf{G}_{\mathrm{ii}}^\ell\}_{\ell=2}^L, \mathbf{F}_{\mathrm{i}}^{L+1}, \mathbf{X}\right) \tag{26}$$

To obtain the full ELBO, we substitute Eqs. (25) and (26) into Eq.(23), the conditional prior terms cancel,

$$\mathcal{L} = \mathbb{E}\left[\log \mathrm{P}\left(\mathbf{Y}|\mathbf{F}_{L+1}\right) + \log \frac{\mathrm{P}\left(\boldsymbol{\Omega}, \{\mathbf{G}_{\mathrm{ii}}^\ell\}_{\ell=2}^L, \mathbf{F}_{L+1}|\mathbf{X}\right)}{\mathrm{Q}\left(\boldsymbol{\Omega}, \{\mathbf{G}_{\mathrm{ii}}^\ell\}_{\ell=2}^L, \mathbf{F}_{L+1}|\mathbf{X}\right)}\right] \tag{27}$$

where,

$$\mathrm{P}\left(\boldsymbol{\Omega}, \{\mathbf{G}_{\mathrm{ii}}^\ell\}_{\ell=2}^L, \mathbf{F}_{\mathrm{i}}^{L+1}|\mathbf{X}_{\mathrm{i}}\right) = \mathrm{P}\left(\boldsymbol{\Omega}\right)\left[\prod_{\ell=2}^L \mathrm{P}\left(\mathbf{G}_{\mathrm{ii}}^\ell|\mathbf{G}_{\mathrm{ii}}^{\ell-1}\right)\right] \mathrm{P}\left(\mathbf{F}_{\mathrm{i}}^{L+1}|\mathbf{G}_{\mathrm{ii}}^L\right), \tag{28}$$

$$\mathrm{Q}\left(\boldsymbol{\Omega}, \{\mathbf{G}_{\mathrm{ii}}^\ell\}_{\ell=2}^L, \mathbf{F}_{\mathrm{i}}^{L+1}|\mathbf{X}_{\mathrm{i}}\right) = \mathrm{Q}\left(\boldsymbol{\Omega}\right)\left[\prod_{\ell=2}^L \mathrm{Q}\left(\mathbf{G}_{\mathrm{ii}}^\ell|\mathbf{G}_{\mathrm{ii}}^{\ell-1}\right)\right] \mathrm{Q}\left(\mathbf{F}_{\mathrm{i}}^{L+1}|\mathbf{G}_{\mathrm{ii}}^L\right). \tag{29}$$

Importantly, the first term in the ELBO (Eq. 27 is a summation across test/train datapoints, and the second term depends only on the inducing points, so as in Salimbeni & Deisenroth (2017) we can compute unbiased estimates of the expectation by taking only a minibatch of datapoints, and we never need to compute the density of the conditional prior (Eq. 30), we only need to be able to sample it.

Finally, to sample the test/training points, conditioned on the inducing points, we need to sample,

$$\mathrm{P}\left(\{\mathbf{G}_{\mathrm{it}}^\ell, \mathbf{G}_{\mathrm{tt}}^\ell\}_{\ell=2}^L, \mathbf{F}_{\mathrm{t}}^{L+1}|\boldsymbol{\Omega}, \{\mathbf{G}_{\mathrm{ii}}^\ell\}_{\ell=2}^L, \mathbf{F}_{\mathrm{i}}^{L+1}, \mathbf{X}\right) = \mathrm{P}\left(\mathbf{F}_{\mathrm{t}}^{L+1}|\mathbf{F}_{\mathrm{i}}^{L+1}, \mathbf{G}_L\right) \prod_{\ell=2}^L \mathrm{P}\left(\mathbf{G}_{\mathrm{it}}^\ell, \mathbf{G}_{\mathrm{tt}}^\ell|\mathbf{G}_{\mathrm{ii}}^\ell, \mathbf{G}_{\ell-1}\right). \tag{30}$$

The first distribution, $\mathrm{P}\left(\mathbf{F}_{\mathrm{t}}^{L+1}|\mathbf{F}_{\mathrm{i}}^{L+1}, \mathbf{G}_L\right)$, is a multivariate Gaussian, and can be evaluated using methods from the GP literature (Williams & Rasmussen, 2006; Salimbeni & Deisenroth, 2017). The difficulties arise for the inverse Wishart terms, $\mathrm{P}\left(\mathbf{G}_{\mathrm{it}}^\ell, \mathbf{G}_{\mathrm{tt}}^\ell|\mathbf{G}_{\mathrm{ii}}^\ell, \mathbf{G}_{\ell-1}\right)$. To sample this distribution, note that samples from the joint over inducing and train/test locations can be written,

$$\begin{pmatrix}\mathbf{G}_{\mathrm{ii}}^\ell & \mathbf{G}_{\mathrm{it}}^\ell \\ \mathbf{G}_{\mathrm{ti}}^\ell & \mathbf{G}_{\mathrm{tt}}^\ell\end{pmatrix} \sim \mathcal{W}^{-1}\left(\begin{pmatrix}\boldsymbol{\Psi}_{\mathrm{ii}} & \boldsymbol{\Psi}_{\mathrm{it}} \\ \boldsymbol{\Psi}_{\mathrm{ti}} & \boldsymbol{\Psi}_{\mathrm{tt}}\end{pmatrix}, \delta_\ell + P_{\mathrm{i}} + P_{\mathrm{t}} + 1\right), \quad \begin{pmatrix}\boldsymbol{\Psi}_{\mathrm{ii}} & \boldsymbol{\Psi}_{\mathrm{it}} \\ \boldsymbol{\Psi}_{\mathrm{ti}} & \boldsymbol{\Psi}_{\mathrm{tt}}\end{pmatrix} = \delta_\ell \mathbf{K}\left(\mathbf{G}_{\ell-1}\right), \tag{31}$$

and where P_{i} is the number of inducing inputs, and P_t is the number of train/test inputs. Defining the Schur complements,

$$\mathbf{G}_{\mathrm{tt}\cdot\mathrm{i}}^\ell = \mathbf{G}_{\mathrm{tt}}^\ell - \mathbf{G}_{\mathrm{ti}}^\ell \left(\mathbf{G}_{\mathrm{ii}}^\ell\right)^{-1} \mathbf{G}_{\mathrm{it}}^\ell, \tag{32}$$

$$\boldsymbol{\Psi}_{\mathrm{tt}\cdot\mathrm{i}} = \boldsymbol{\Psi}_{\mathrm{tt}} - \boldsymbol{\Psi}_{\mathrm{ti}} \boldsymbol{\Psi}_{\mathrm{ii}}^{-1} \boldsymbol{\Psi}_{\mathrm{it}}. \tag{33}$$

We know that $\mathbf{G}_{\mathrm{tt}\cdot\mathrm{i}}^\ell$ and $\left(\mathbf{G}_{\mathrm{ii}}^\ell\right)^{-1} \mathbf{G}_{\mathrm{it}}^\ell$ have distribution, (Eaton, 1983)

$$\mathbf{G}_{\mathrm{tt}\cdot\mathrm{i}}^\ell \big| \mathbf{G}_{\mathrm{ii}}^\ell, \mathbf{G}_{\ell-1} \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}_{\mathrm{tt}\cdot\mathrm{i}}, \delta_\ell + P_{\mathrm{i}} + P_{\mathrm{t}} + 1), \tag{34a}$$

$$\mathbf{G}_{\mathrm{it}}^\ell \big| \mathbf{G}_{\mathrm{tt}\cdot\mathrm{i}}^\ell, \mathbf{G}_{\mathrm{ii}}^\ell, \mathbf{G}_{\ell-1} \sim \mathcal{MN}\left(\mathbf{G}_{\mathrm{ii}}^\ell \boldsymbol{\Psi}_{\mathrm{ii}}^{-1} \boldsymbol{\Psi}_{\mathrm{it}}, \mathbf{G}_{\mathrm{ii}}^\ell \boldsymbol{\Psi}_{\mathrm{ii}}^{-1} \mathbf{G}_{\mathrm{ii}}^\ell, \mathbf{G}_{\mathrm{tt}\cdot\mathrm{i}}^\ell\right), \tag{34b}$$

where $\mathcal{MN}$ is the matrix normal. Now, $\mathbf{G}_{\mathrm{it}}^\ell$ and $\mathbf{G}_{\mathrm{tt}}^\ell$, can be recovered by algebraic manipulation. Finally, because of the doubly stochastic form for the objective, we do not need to sample multiple of jointly consistent samples for test points; instead, (and as in DGPs Salimbeni & Deisenroth, 2017) we can independently sample each test point (App. I), which dramatically reduces computational complexity.

The full algorithm is given in Alg. 1, where the P and Q distributions for $\boldsymbol{\Omega}$ and for inducing points are given by Eq. (18) and (21). We optimize using standard reparameterised variational inference (Kingma & Welling, 2013; Rezende et al., 2014) (for details on how to reparameterise samples from the Wishart, see Ober & Aitchison, 2020).

6. Computational complexity

As in non-deep GPs, the complexity is $\mathcal{O}(P^3)$ for time and $\mathcal{O}(P^2)$ for space for standard DKPs (the $\mathcal{O}(P^3)$ time dependencies emerge e.g. because of inverses and determinants required for the inverse Wishart distributions). For DSVI, there is a P_{i}^3 time and P_{i}^2 space term for the inducing points, because the computations for inducing points are exactly the same as in the non-DSVI case. As we can treat each test/train point independently (App. I), the complexity for test/training points must scale linearly with P_{t}, and this term has P_{i}^2 time scaling, e.g. due to the matrix products in Eq. (32). Thus, the overall complexity for DSVI is $\mathcal{O}(P_{\mathrm{i}}^3 + P_{\mathrm{i}}^2 P_{\mathrm{t}})$ for time and $\mathcal{O}(P_{\mathrm{i}}^2 + P_{\mathrm{i}} P_{\mathrm{t}})$ for space which is exactly the same as non-deep inducing GPs. Thus, and exactly as in non-deep inducing-GPs, by using a small number of inducing points, we are able to convert a cubic dependence on the number of input points into a linear dependence, which gives considerably better scaling.

Surprisingly, this is substantially better than standard DGPs. In standard DGPs, we allow the approximate posterior covariance for each feature to differ (Salimbeni & Deisenroth, 2017), in which case, we are in essence doing standard inducing-GP inference over N hidden features, which gives complexity of $\mathcal{O}(NP_{\mathrm{i}}^3 + NP_{\mathrm{i}}^2 P_{\mathrm{t}})$ for time and $\mathcal{O}(NP_{\mathrm{i}}^2 + NP_{\mathrm{i}} P_{\mathrm{t}})$ for space (Salimbeni & Deisenroth, 2017). It is possible to improve this complexity by restricting the approximate posterior to have the same covariance for each point (but this restriction harms performance).

Algorithm 1 Computing predictions/ELBO for one batch

P parameters: $\{\delta_\ell\}_{\ell=1}^L$.
Q parameters: $\{\mathbf{V}_\ell, \gamma_\ell\}_{\ell=1}^L, (\{\mathbf{v}_\lambda, \mathbf{\Lambda}_\lambda\}_{\lambda=1}^{N_{L+1}}), \mathbf{X}_\mathrm{i}$.
Inputs: $\mathbf{X}_\mathrm{t}$; **Targets:** $\mathbf{Y}$
combine inducing and test/train inputs
$\mathbf{X} = \begin{pmatrix}\mathbf{X}_\mathrm{i} & \mathbf{X}_\mathrm{t}\end{pmatrix}$
sample first Gram matrix and update ELBO
$\mathbf{\Omega} \sim \mathrm{Q}\left(\mathbf{\Omega}\right)$
$\mathcal{L} = \log \mathrm{P}\left(\mathbf{\Omega}\right) - \log \mathrm{Q}\left(\mathbf{\Omega}\right)$
$\mathbf{G}_1 = \frac{1}{N_0}\mathbf{X}\mathbf{\Omega}\mathbf{X}^T$
for ℓ **in** $\{2, \ldots, L\}$ **do**
  sample inducing Gram matrix and update ELBO
  $\mathbf{G}_{\mathrm{ii}}^\ell \sim \mathrm{Q}\left(\mathbf{G}_{\mathrm{ii}}^\ell | \mathbf{G}_{\mathrm{ii}}^{\ell-1}\right)$
  $\mathcal{L} \leftarrow \mathcal{L} + \log \mathrm{P}\left(\mathbf{G}_{\mathrm{ii}}^\ell | \mathbf{G}_{\mathrm{ii}}^{\ell-1}\right) - \log \mathrm{Q}\left(\mathbf{G}_{\mathrm{ii}}^\ell | \mathbf{G}_{\mathrm{ii}}^{\ell-1}\right)$
  sample full Gram matrix from conditional prior
  $\mathbf{\Psi} = \delta_\ell \mathbf{K}(\mathbf{G}_{\ell-1})$
  $\mathbf{\Psi}_{\mathrm{tt\cdot i}} = \mathbf{\Psi}_{\mathrm{tt}} - \mathbf{\Psi}_{\mathrm{ti}}\mathbf{\Psi}_{\mathrm{ii}}^{-1}\mathbf{\Psi}_{\mathrm{it}}$
  $\mathbf{G}_{\mathrm{tt\cdot i}}^\ell \sim \mathcal{W}\left(\mathbf{\Psi}_{\mathrm{tt\cdot i}}^\ell, \delta_\ell + P_\mathrm{i} + P_\mathrm{t} + 1\right)$
  $\mathbf{G}_{\mathrm{it}}^\ell \sim \mathcal{MN}\left(\mathbf{G}_{\mathrm{ii}}^\ell \mathbf{\Psi}_{\mathrm{ii}}^{-1}\mathbf{\Psi}_{\mathrm{it}}, \mathbf{G}_{\mathrm{ii}}^\ell \mathbf{\Psi}_{\mathrm{ii}}^{-1}\mathbf{G}_{\mathrm{ii}}^\ell, \mathbf{G}_{\mathrm{tt\cdot i}}^\ell\right)$
  $\mathbf{G}_{\mathrm{tt}}^\ell = \mathbf{G}_{\mathrm{tt\cdot i}}^\ell + \mathbf{\Psi}_{\mathrm{ti}}\mathbf{\Psi}_{\mathrm{ii}}^{-1}\mathbf{\Psi}_{\mathrm{it}}$.
  $\mathbf{G}_\ell = \begin{pmatrix}\mathbf{G}_{\mathrm{ii}}^\ell & \mathbf{G}_{\mathrm{it}}^\ell \\ \mathbf{G}_{\mathrm{ti}}^\ell & \mathbf{G}_{\mathrm{tt}}^\ell\end{pmatrix}$
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
$\mathcal{L} \leftarrow \mathcal{L} + \log \mathrm{P}\left(\mathbf{Y} | \mathbf{F}_\mathrm{t}^{L+1}\right)$

7. Results

We began by comparing the performance of our deep inverse Wishart process (DIWP) against infinite Bayesian neural networks (known as the neural network Gaussian process or NNGP) and DGPs. To ensure sensible comparisons against the NNGP, we used a ReLU kernel in all models (Cho & Saul, 2009). For all models, we used three layers (two hidden layers and one output layer), with three applications of the kernel. In each case, we used a learned bias and scale for each input feature, and trained for 8000 gradient steps with the Adam optimizer with 100 inducing points, a learning rate of 10^{-2} for the first 4000 steps and 10^{-3} for the final 4000 steps. For evaluation, we used approximate posterior 100 samples, and for each training step we used 10 approximate posterior samples in the smaller datasets (boston, concrete, energy, wine, yacht), and 1 in the larger datasets.

We found that DIWP usually gives better predictive performance and (and when it does not, the differences are very small; Table 1). We expected DIWP to be better than (or the same as) the NNGP as the NNGP was a special case of our DIWP (sending $\delta_\ell \to \infty$ sends the variance of the inverse Wishart to zero, so the model becomes equivalent to the NNGP). We found that the DGP performs poorly in comparison to DIWP and NNGPs, and even to past baselines on all datasets except protein (which is by far the largest). This is because we used a plain feedforward architecture for all models. In contrast, Salimbeni & Deisenroth (2017) found that good performance (or even convergence) with DGPs on UCI datasets required a complex GP-prior inspired by skip connections. Here, we used simple feedforward architectures, both to ensure a fair comparison to the other models, and to avoid the need for an architecture search. In addition, the inverse Wishart process is implicitly able to learn the network "width", δ_ℓ, whereas in the DGPs, the width is fixed to be equal to the number of input features, following standard practice in the literature (e.g. Salimbeni & Deisenroth, 2017).

Next, we considered fully-connected networks for small image classification datasets (MNIST and CIFAR-10; Table 2). We used the same models as in the previous section, with the omission of learned bias and scaling of the inputs. Note that we do not expect these methods to perform well relative to standard methods (e.g. CNNs) for these datasets, as we are using fully-connected networks with only 100 inducing points (whereas e.g. work in the NNGP literature uses the full $60,000 \times 60,000$ covariance matrix). Nonetheless, as the architectures are carefully matched, it provides another opportunity to compare the performance of DIWPs, NNGPs and DGPs. Again, we found that DIWP usually gave statistically significant gains in predictive performance (except for CIFAR-10 test-log-likelihood, where DIWP lagged by only 0.01). Importantly, DIWP gives very large improvements in the ELBO, with gains of 0.09 against DGPs for MNIST and 0.08 for CIFAR-10 (Table 2). For MNIST, remember that the ELBO must be negative (because both the log-likelihood for classification and the KL-divergence term give negative contributions), so the change from -0.301 to -0.214 represents a dramatic improvement.

8. Related work

Our first contribution was the observation that DGPs with isotropic kernels can be written as deep Wishart processes as the kernel depends only on the Gram matrix. We then gave similar observations for neural networks (App. C.1), infinite neural networks (App. C.2) and infinite network with bottlenecks (App. C.3, also see Aitchison, 2019). These observations motivated us to consider the deep inverse Wishart process prior, which is a novel combination of two pre-existing elements: nonlinear transformations of the kernel (e.g. Cho & Saul, 2009) and inverse Wishart priors over kernels (e.g. Shah et al., 2014). Deep nonlinear transformations

Table 1. Performance measured as predictive log-likelihood for a three-layer (two hidden layer) DGP, NNGP and DIWP on UCI benchmark tasks. We have consider relu and squared exponential kernels. Errors are quoted as two standard errors in the *difference* between that method and the best performing method, as in a paired t-test. This is to account for the shared variability that arises due to the use of different test/train splits in the data (20 splits for all but protein, where 5 splits are used Gal & Ghahramani, 2015) some splits are harder for all models, and some splits are easier. Because we consider these differences, errors for the best measure are implicitly included in errors for other measures, and we cannot provide a comparable error for the best method itself.

kernel	dataset	DGP	NNGP	DIWP
relu	boston	-3.44 ± 0.13	-2.46 ± 0.03	$\mathbf{-2.40}$
	concrete	-3.20 ± 0.03	-3.13 ± 0.03	$\mathbf{-3.04}$
	energy	-0.90 ± 0.05	$\mathbf{-0.71}$	$\mathbf{-0.71 \pm 0.01}$
	kin8nm	1.05 ± 0.01	$\mathbf{1.10}$	1.09 ± 0.00
	naval	2.80 ± 0.14	5.74 ± 0.14	$\mathbf{5.95}$
	power	-2.85 ± 0.01	-2.83 ± 0.00	$\mathbf{-2.81}$
	protein	$\mathbf{-2.80}$	-2.88 ± 0.01	-2.81 ± 0.01
	wine	-1.18 ± 0.03	-0.96 ± 0.01	$\mathbf{-0.95}$
	yacht	-2.45 ± 0.49	-0.77 ± 0.07	$\mathbf{-0.64}$
sq. exp.	boston	-3.63 ± 0.09	-2.48 ± 0.04	$\mathbf{-2.40}$
	concrete	-4.22 ± 0.05	-3.13 ± 0.03	$\mathbf{-3.08}$
	energy	-3.73 ± 0.06	$\mathbf{-0.70}$	$\mathbf{-0.70 \pm 0.01}$
	kin8nm	-0.09 ± 0.01	$\mathbf{1.04}$	1.01 ± 0.02
	naval	2.80 ± 0.12	$\mathbf{5.96}$	$\mathbf{5.92 \pm 0.27}$
	power	-2.84 ± 0.01	-2.80 ± 0.01	$\mathbf{-2.78}$
	protein	-3.23 ± 0.01	-2.88 ± 0.01	$\mathbf{-2.74}$
	wine	-1.22 ± 0.02	$\mathbf{-0.96}$	-1.00 ± 0.01
	yacht	-4.12 ± 0.14	$\mathbf{-0.50 \pm 0.13}$	$\mathbf{-0.39}$

Table 2. Performance in terms of ELBO test log-likelihood and test accuracy for fully-connected three-layer (two hidden layer) DGPs, NNGP and DIWP on MNIST and CIFAR-10.

metric	dataset	DGP	NNGP	DIWP
ELBO	MNIST	-0.301 ± 0.001	-0.268 ± 0.001	$\mathbf{-0.198 \pm 0.000}$
	CIFAR-10	-1.735 ± 0.002	-1.719 ± 0.001	$\mathbf{-1.606 \pm 0.001}$
test LL	MNIST	-0.130 ± 0.001	-0.134 ± 0.002	$\mathbf{-0.089 \pm 0.001}$
	CIFAR-10	-1.516 ± 0.002	-1.539 ± 0.002	$\mathbf{-1.433 \pm 0.004}$
test acc.	MNIST	$96.5 \pm 0.1\%$	$96.5 \pm 0.0\%$	$\mathbf{97.7 \pm 0.0\%}$
	CIFAR-10	$46.8 \pm 0.1\%$	$47.4 \pm 0.1\%$	$\mathbf{50.5 \pm 0.1\%}$

of the kernel have been used in the infinite neural network literature (Lee et al., 2017; Matthews et al., 2018) where they form deterministic, parameter-free kernels that do not have any flexibility in the lower-layers (Aitchison, 2019). Likewise, inverse-Wishart distributions have been suggested as priors over covariance matrices (Shah et al., 2014), but they considered a model without nonlinear transformations of the kernel. Surprisingly, without these nonlinear transformations, the inverse Wishart prior becomes equivalent to simply scaling the covariance with a scalar random variable (App. L; Shah et al., 2014).

In addition, there are *generalised* Wishart processes (Wilson & Ghahramani, 2010, contrasting with our *deep* Wishart processes). While the term "generalised Wishart process" is not yet in widespread use, it allows us to make a distinction that is very useful in our context. In particular, a generalised Wishart process is a distribution over infinitely many finite-dimensional marginally Wishart matrices. For instance, these might represent the noise in a dynamical system. In that case, there would in principle be infinitely covariance matrices, one for each state-space location or time-point (Wilson & Ghahramani, 2010; Heaukulani & van der Wilk, 2019; Jorgensen et al., 2020). In contrast, kernel processes (Dawid, 1981; Bru, 1991) are distributions over a single infinite dimensional matrix. We stack these kernel process to form a (non-genearlised) deep kernel pro-

cess. Importantly, generalised Wishart priors are actually quite inflexible. They are not capable of capturing a DKP prior because in a generalised Wishart process, the Wishart matrices are generated from underlying features, and these features are jointly multivariate Gaussian at all locations (Sec. 4 in Wilson & Ghahramani, 2010) and therefore lack the required nonlinearities between layers. In addition, inference is also very different. In particular, inference for the generalised Wishart is generally performed on the underlying multivariate Gaussian feature vectors (Eq. 1 e.g. Eq. 15-18 in Wilson & Ghahramani 2010, Eq. 12 in Heaukulani & van der Wilk 2019 or Eq. 24 in Jorgensen et al. 2020). Unfortunately, variational approximate posteriors defined over multivariate Gaussian feature vectors fail to capture symmetries in the true posterior (Eq. 14). In contrast, we define approximate posteriors directly over the symmetric positive semi-definite Gram matrices themselves, which required us to develop new, more flexible distributions over these matrices.

Further linear (inverse) Wishart processes have been used in the financial domain to model how the volatility of asset prices changes over time (Philipov & Glickman, 2006b;a; Asai & McAleer, 2009; Gourieroux & Sufana, 2010; Wilson & Ghahramani, 2010; Heaukulani & van der Wilk, 2019). Importantly, inference in these dynamical (inverse) Wishart processes is often performed by assuming fixed, integer degrees of freedom, and working with underlying Gaussian distributed features. This approach allows one to leverage standard GP techniques (e.g. Kandemir & Hamprecht, 2015; Heaukulani & van der Wilk, 2019), but it is not possible to optimize the degrees of freedom and the posterior over these features usually has rotational symmetries (App. D.2) that are not captured by standard variational posteriors. In contrast, we give a novel doubly-stochastic variational inducing point inference method that operates purely on Gram matrices and thus avoids needing to capture these symmetries.

9. Conclusions

We proposed deep kernel processes which combine nonlinear transformations of the Gram matrix with sampling from matrix-variate distributions such as the inverse Wishart. We showed that DGPs, BNNs (App. C.1), infinite BNNs (App. C.2) and infinite BNNs with bottlenecks (App. C.3) are all instances of DKPs. We defined a new family of deep inverse Wishart processes, and give a novel doubly-stochastic inducing point variational inference scheme that works purely in the space of Gram matrices. DIWP performed better than fully connected NNGPs and DGPs on UCI, MNIST and CIFAR-10 benchmarks.

Acknowledgments

SWO acknowledges the support of the Gates Cambridge Trust for funding his doctoral studies. We would like to thank the University of Bristol's Advanced Computing Research Centre (ACRC) for computational resources.

References

Aitchison, L. Why bigger is not always better: on finite and infinite neural networks. *arXiv preprint arXiv:1910.08013*, 2019.

Asai, M. and McAleer, M. The structure of dynamic correlations in multivariate stochastic volatility models. *Journal of Econometrics*, 150(2):182–192, 2009.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Bodnar, T. and Okhrin, Y. Properties of the singular, inverse and generalized inverse partitioned Wishart distributions. *Journal of Multivariate Analysis*, 99(10):2389–2405, 2008.

Bodnar, T., Mazur, S., and Podgórski, K. Singular inverse Wishart distribution and its application to portfolio theory. *Journal of Multivariate Analysis*, 143:314–326, 2016.

Bru, M.-F. Wishart processes. *Journal of Theoretical Probability*, 4(4):725–751, 1991.

Cho, Y. and Saul, L. K. Kernel methods for deep learning. In *Advances in neural information processing systems*, pp. 342–350, 2009.

Damianou, A. and Lawrence, N. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pp. 207–215, 2013.

Dawid, A. P. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 1981.

Eaton, M. L. *Multivariate Statistics. A Vector Space Approach.-A Volume in the Wiley Series in Probability and Mathematical Statistics*. Wiley-Interscience, 1983.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *arXiv:1506.02142*, 2015.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. Deep convolutional networks as shallow Gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pp. 1884–1892, 2016.

Gourieroux, C. and Sufana, R. Derivative pricing with wishart multivariate stochastic volatility. *Journal of Business & Economic Statistics*, 28(3):438–451, 2010.

Heaukulani, C. and van der Wilk, M. Scalable Bayesian dynamic covariance modeling with variational Wishart and inverse Wishart processes. In *Advances in Neural Information Processing Systems*, pp. 4582–4592, 2019.

Hofmann, T., Schölkopf, B., and Smola, A. J. Kernel methods in machine learning. *The annals of statistics*, pp. 1171–1220, 2008.

Jorgensen, M., Deisenroth, M., and Salimbeni, H. Stochastic differential equations with variational Wishart diffusions. In *International Conference on Machine Learning*, pp. 4974–4983. PMLR, 2020.

Kandemir, M. and Hamprecht, F. A. The deep feed-forward Gaussian process: An effective generalization to covariance priors. In *Feature Extraction: Modern Questions and Challenges*, pp. 145–159, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Koev, P. and Edelman, A. The efficient evaluation of the hypergeometric function of a matrix argument. *Mathematics of Computation*, 75(254):833–846, 2006.

Kolmogorov, A. Grundbegriffe der Wahrscheinlichkeitreichnung. *Ergebnisse der Mathematik*, 1933.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Lalchand, V. and Rasmussen, C. E. Approximate inference for fully Bayesian Gaussian process regression. In *Symposium on Advances in Approximate Bayesian Inference*, pp. 1–12. PMLR, 2020.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as Gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *Advances in neural information processing systems*, pp. 6389–6399, 2018.

MacKay, D. J. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Moore, D. A. Symmetrized variational inference. In *NIPS Workshop on Advances in Approximate Bayesian Inferece*, volume 4, pp. 31, 2016.

Novak, R., Xiao, L., Lee, J., Bahri, Y., Yang, G., Hron, J., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Bayesian deep convolutional networks with many channels are Gaussian processes. *arXiv preprint arXiv:1810.05148*, 2018.

Ober, S. W. and Aitchison, L. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. *arXiv preprint arXiv:2005.08140*, 2020.

Philipov, A. and Glickman, M. E. Factor multivariate stochastic volatility via Wishart processes. *Econometric Reviews*, 25(2-3):311–334, 2006a.

Philipov, A. and Glickman, M. E. Multivariate stochastic volatility via Wishart processes. *Journal of Business & Economic Statistics*, 24(3):313–328, 2006b.

Pourzanjani, A. A., Jiang, R. M., and Petzold, L. R. Improving the identifiability of neural networks for Bayesian inference. In *NIPS Workshop on Bayesian Deep Learning*, volume 4, pp. 31, 2017.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Rivasplata, O., Tankasali, V. M., and Szepesvari, C. PAC-Bayes with backprop. *arXiv preprint arXiv:1908.07380*, 2019.

Salimbeni, H. and Deisenroth, M. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 4588–4599, 2017.

Shah, A., Wilson, A., and Ghahramani, Z. Student-t processes as alternatives to Gaussian processes. In *Artificial intelligence and statistics*, pp. 877–885, 2014.

Srivastava, M. S. et al. Singular Wishart and multivariate beta distributions. *The Annals of Statistics*, 31(5):1537–1560, 2003.

Sterbenz, P. H. *Floating-point computation*. Prentice-Hall, 1974.

Turner, R. E. and Sahani, M. Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, T., and Chiappa, S. (eds.), *Inference and Learning in Dynamic Models*. Cambridge University Press, 2011.

Uhlig, H. On singular Wishart and singular multivariate beta distributions. *The Annals of Statistics*, pp. 395–405, 1994.

Williams, C. K. and Rasmussen, C. E. *Gaussian Processes for Machine Learning*. MIT press Cambridge, MA, 2006.

Wilson, A. G. and Ghahramani, Z. Generalised Wishart processes. *arXiv preprint arXiv:1101.0240*, 2010.