An Improved Variational Approximate Posterior for the Deep Wishart Process (Supplementary Material)

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

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

A DERIVATION OF A-$\mathcal{GW}$ AND AB-$\mathcal{GW}$ DENSITIES

We briefly provide further background on the Wishart distribution, the Barlett decomposition, and discuss how to derive Jacobians for matrix transformations. Then we use this machinery to derive densities for the A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ distributions.

A.1 THE WISHART DISTRIBUTION

The Wishart distribution, $\mathcal{W}(\mathbf{S}, \nu)$, is a distribution over positive semi-definite $P \times P$ matrices, where $\mathbf{S} \in \mathbb{R}^{P\times P}$ is a positive definite covariance matrix, and $\nu > 0$ is an integer-valued degrees-of-freedom parameter. The Wishart distribution is most straightforwardly interpreted as a sum of outer products of multivariate Gaussian random variables. That is, if we define a random matrix $\mathbf{W}$ such that,

$$\mathbf{f}_\lambda \sim_{\text{iid}} \mathcal{N}(\mathbf{0}, \mathbf{S}),\ \lambda \in \{1, \dots, \nu\}, \tag{1}$$

$$\mathbf{W} = \sum_{\lambda=1}^{\nu} \mathbf{f}_\lambda \mathbf{f}_\lambda^T, \tag{2}$$

then we say that $\mathbf{W}$ is Wishart distributed, and write $\mathbf{W} \sim \mathcal{W}(\mathbf{S}, \nu)$. Equivalently, $\mathbf{W} = \mathbf{F}\mathbf{F}^T$, where $\mathbf{F} \in \mathbb{R}^{P\times P}$ is defined by stacking the vectors $\mathbf{f}_\lambda$, $\mathbf{F} = (\mathbf{f}_1 \quad \cdots \quad \mathbf{f}_\lambda)$. We say that $\mathbf{W}$ is standard Wishart distributed if $\mathbf{S} = \mathbf{I}$.

It is easy to generate Wishart random matrices from only standard Gaussian samples. Take $\mathbf{L} = \text{chol}(\mathbf{S})$ to be the Cholesky of $\mathbf{S}$ and $\boldsymbol{\xi}_\lambda \sim_{\text{iid}} \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $\mathbf{L}\boldsymbol{\xi}_\lambda \sim_{\text{iid}} \mathcal{N}(\mathbf{0}, \mathbf{S})$. It follows that,

$$\mathbf{L}\left(\sum_{\lambda=1}^{\nu} \boldsymbol{\xi}_\lambda \boldsymbol{\xi}_\lambda^T\right)\mathbf{L}^T = \mathbf{L}\boldsymbol{\Xi}\boldsymbol{\Xi}^T\mathbf{L}^T \sim \mathcal{W}(\mathbf{S}, \nu), \tag{3}$$

where $\boldsymbol{\Xi}$ is the matrix of stacked vectors $\boldsymbol{\xi}_\lambda$ such that $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1 \quad \cdots \quad \boldsymbol{\xi}_\nu)$. From (3), it can be observed that $\mathbb{E}[\mathbf{W}] = \mathbf{L}(\nu\mathbf{I})\mathbf{L}^T = \nu\mathbf{S}$. Additionally, $\mathbf{H} = \boldsymbol{\Xi}\boldsymbol{\Xi}^T$ is standard Wishart distributed, therefore (3) also gives us a way to transform a standard Wishart into a Wishart with covariance parameter $\mathbf{S}$: $\mathbf{H} \sim \mathcal{W}(\mathbf{I}, \nu) \implies \mathbf{L}\mathbf{H}\mathbf{L}^T \sim \mathcal{W}(\mathbf{S}, \nu)$.

Finally, note that the density of the Wishart distribution is given by,

$$\text{P}(\mathbf{W}) = \frac{\pi^{\nu(\tilde{\nu}-P)/2}}{2^{\nu P/2}\,|\mathbf{S}|^{\nu/2}\,\boldsymbol{\Gamma}_{\tilde{\nu}}\left(\frac{\nu}{2}\right)}\,|\mathbf{W}_{:\tilde{\nu},:\tilde{\nu}}|^{(\nu-P-1)/2}\,\text{etr}\left(-\mathbf{S}^{-1}\mathbf{W}/2\right), \tag{4}$$

where $\tilde{\nu} = \min(\nu, P)$, and $\Gamma_{\tilde{\nu}}$ is the multivariate gamma function [Srivastava, 2003].

[*]These authors contributed equally to this work.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

A.2 THE BARTLETT DECOMPOSITION AND SOME GENERALISATIONS

Suppose $\mathbf{W} \sim \mathcal{W}(\mathbf{I}, \nu)$, and $\nu \geq P$, then the Bartlett decomposition [Bartlett, 1933] allows for efficient sampling of $\mathbf{W}$ (the constraint $\nu \geq P$ refers to the fact that $\mathbf{W}$ almost surely has full rank). Rather than sampling νP^2 Gaussian random variables to construct $\mathbf{W}$ (which can become prohibitively costly when ν is large), the Bartlett decomposition allows us to sample only $P(P-1)/2$ Gaussian random variables, and P Gamma random variables. In particular, if $\mathbf{T}$ is a random matrix distributed according to,

$$\mathbf{T} = \begin{pmatrix} T_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ T_{P1} & \cdots & T_{PP} \end{pmatrix}, \tag{5a}$$

$$\mathrm{P}\left(T_{jj}^2\right) = \mathrm{Gamma}\left(T_{jj}^2;\, \tfrac{\nu-j+1}{2}, \tfrac{1}{2}\right), \tag{5b}$$

$$\mathrm{P}\left(T_{j>k}\right) = \mathcal{N}\left(T_{jk};\, 0, 1\right), \tag{5c}$$

then $\mathbf{T}\mathbf{T}^T \sim \mathcal{W}(\mathbf{I}, \nu)$. The utility of (5) can be extended in two ways. Firstly, we can use (5) to sample from non-standard Wisharts, since $\mathbf{L}(\mathbf{T}\mathbf{T}^T)\mathbf{L}^T \sim \mathcal{W}\left(\mathbf{L}\mathbf{L}^T, \nu\right)$. Secondly, Srivastava [2003] extends the Bartlett decomposition to allow for sampling of singular Wisharts. Suppose $\nu < P$, and take $\mathbf{T}$ to be distributed according to,

$$\mathbf{T} = \begin{pmatrix} T_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ T_{\nu 1} & \cdots & T_{\nu\nu} \\ \vdots & \ddots & \vdots \\ T_{P1} & \cdots & T_{P\nu} \end{pmatrix}, \tag{6a}$$

$$\mathrm{P}\left(T_{ii}^2\right) = \mathrm{Gamma}\left(T_{ii}^2;\, \tfrac{\nu-j+1}{2}, \tfrac{1}{2}\right),\ i \in \{1, \ldots, \nu\}, \tag{6b}$$

$$\mathrm{P}\left(T_{i>j}\right) = \mathcal{N}\left(T_{i>j};\, 0, 1\right), \tag{6c}$$

then $\mathbf{T}\mathbf{T}^T \sim \mathcal{W}(\mathbf{I}, \nu)$.

We arrive at the A- and AB-generalised (singular) Wishart distributions by generalising the (singular) Barlett decomposition in (6). Concretely, we borrow the form of (6), but allow the parameters of the Gaussian and gamma distributions to be arbitrary,

$$\mathbf{T} = \begin{pmatrix} T_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ T_{\nu 1} & \cdots & T_{\nu\nu} \\ \vdots & \ddots & \vdots \\ T_{P1} & \cdots & T_{P\nu} \end{pmatrix}, \tag{7a}$$

$$\mathrm{P}\left(T_{ii}^2\right) = \mathrm{Gamma}\left(T_{ii}^2;\, \alpha_i, \beta_i\right),\ i \in \{1, \ldots, \nu\}, \tag{7b}$$

$$\mathrm{P}\left(T_{i>j}\right) = \mathcal{N}\left(T_{i>j};\, \mu_{ij}, \sigma_{ij}^2\right). \tag{7c}$$

For any invertible matrix $\mathbf{A} \in \mathbb{R}^{P\times P}$ and any invertible lower triangular $\mathbf{B} \in \mathbb{R}^{\nu\times\nu}$, we write $\mathbf{A}\mathbf{T}\mathbf{T}^T\mathbf{A}^T \sim \text{A-}\mathcal{GW}(\mathbf{A}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ and $\mathbf{A}\mathbf{T}\mathbf{B}\mathbf{B}^T\mathbf{T}^T\mathbf{A}^T \sim \text{AB-}\mathcal{GW}(\mathbf{A}, \mathbf{B}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$. Given the necessary parameters, it is straightforward to sample matrices from the A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ families using (7). However it is non-trivial to write down the corresponding densities — the rest of this section is dedicated to this task.

A.3 JACOBIANS FOR MATRIX TRANSFORMATIONS

We want to obtain the densities of $\mathbf{W}_A := \mathbf{A}\mathbf{T}\mathbf{T}^T\mathbf{A}^T$ and $\mathbf{W}_{AB} := \mathbf{A}\mathbf{T}\mathbf{B}\mathbf{B}^T\mathbf{T}^T\mathbf{A}^T$, where we know the density of $\mathbf{T}$. Ultimately, we will use the change of variables formula,

$$\mathrm{Q}\left(\mathbf{W}\right) = \mathrm{Q}\left(\mathbf{T}\right)\left|\frac{\partial \mathbf{T}}{\partial \mathbf{W}}\right|, \tag{8}$$

where $|\partial \mathbf{T}/\partial \mathbf{W}|$ is the Jacobian determinant of the transformation.

For a vector-vector transformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$, the Jacobian $\partial \mathbf{y}/\partial \mathbf{x}$ can be calculated by evaluating $\partial y_i/\partial x_j$ for $i \in \{1, \ldots, m\}$, and $j \in \{1, \ldots, n\}$. It is less simple to calculate the Jacobian for matrix-matrix transformations, but it can be done by stacking the columns of our matrices into a long vector, and then calculating the associated vector-vector Jacobian. We demonstrate this with a simple example for 2×2 matrices. Consider,

$$\underbrace{\begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}}_{\mathbf{X}}. \tag{9}$$

We 'vectorise' $\mathbf{Y}$ and $\mathbf{X}$ to obtain,

$$\begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \end{pmatrix} = \underbrace{\begin{pmatrix} A_{11} & A_{12} & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 \\ 0 & 0 & A_{11} & A_{12} \\ 0 & 0 & A_{21} & A_{22} \end{pmatrix}}_{\mathbf{A}^*} \begin{pmatrix} X_{11} \\ X_{21} \\ X_{12} \\ X_{22} \end{pmatrix}. \tag{10}$$

The Jacobian of this transformation is clearly,

$$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}} = \mathbf{A}^*, \tag{11}$$

and the associated Jacobian determinant is therefore,

$$\left| \frac{\partial \mathbf{Y}}{\partial \mathbf{X}} \right| = |\mathbf{A}|^2. \tag{12}$$

We now consider how to calculate some Jacobian determinants that are relevant in calculating the A-$\mathcal{GW}$ and AB-$\mathcal{GW}$ densities.

A.4 JACOBIAN FOR THE PRODUCT OF A LOWER TRIANGULAR MATRIX WITH ITSELF

Consider the transformation $\mathbf{G} = \mathbf{\Lambda}\mathbf{\Lambda}^T$, where $\mathbf{\Lambda} \in \mathbb{R}^{P \times P}$, and $\mathbf{\Lambda}$ is lower triangular. Ober and Aitchison [2021a] showed that the Jacobian determinant is,

$$\left| \frac{\partial \mathbf{G}}{\partial \mathbf{\Lambda}} \right| = \prod_{i=1}^{P} 2\Lambda_{ii}^{P-i+1}. \tag{13}$$

They also showed that the same transformation, $\mathbf{G} = \mathbf{\Lambda}\mathbf{\Lambda}^T$, but in the case $\mathbf{\Lambda} \in \mathbb{R}^{P \times \nu}$ has Jacobian determinant,

$$\left| \frac{\partial \mathbf{G}}{\partial \mathbf{\Lambda}} \right| = \prod_{i=1}^{\tilde{\nu}} 2\Lambda_{ii}^{P-i+1}, \tag{14}$$

where $\tilde{\nu} = \min\{P, \nu\}$.

A.5 JACOBIAN FOR THE PRODUCT OF TWO DIFFERENT LOWER TRIANGULAR MATRICES

Consider the transformation $\mathbf{T} \mapsto \mathbf{\Lambda} = \mathbf{L}\mathbf{T}$, where $\mathbf{T} \in \mathbb{R}^{P \times \nu}$ and is lower triangular, and $\mathbf{L} \in \mathbb{R}^{P \times P}$ is also lower triangular. Ober and Aitchison [2021a] showed that the Jacobian determinant is,

$$\left| \frac{\partial \mathbf{\Lambda}}{\partial \mathbf{T}} \right| = \prod_{i=1}^{P} L_{ii}^{\min(i,\nu)}. \tag{15}$$

We also need the Jacobian determinant for a right linear transformation. Therefore, consider also the transformation $\mathbf{T} \mapsto \mathbf{\Lambda} = \mathbf{TB}$, where again $\mathbf{T} \in \mathbb{R}^{P\times\nu}$, but $\mathbf{B} \in \mathbb{R}^{\nu\times\nu}$ and is invertible lower triangular. It is helpful to write down the matrices explicitly,

$$\begin{pmatrix} \Lambda_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ \Lambda_{\nu 1} & \cdots & \Lambda_{\nu\nu} \\ \vdots & \vdots & \vdots \\ \Lambda_{P1} & \cdots & \Lambda_{P\nu} \end{pmatrix} = \begin{pmatrix} T_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ T_{\nu 1} & \cdots & T_{\nu\nu} \\ \vdots & \vdots & \vdots \\ T_{P1} & \cdots & T_{P\nu} \end{pmatrix} \begin{pmatrix} B_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ B_{\nu 1} & \cdots & B_{\nu\nu} \end{pmatrix}, \tag{16}$$

and consider the rows of $\mathbf{\Lambda}$. For the first row, we have

$$\begin{pmatrix} \Lambda_{11} \end{pmatrix} = \begin{pmatrix} T_{11} \end{pmatrix} \begin{pmatrix} B_{11} \end{pmatrix},$$

or equivalently,

$$\mathbf{\Lambda}_{1,:1} = \mathbf{T}_{1,:1}\mathbf{B}_{:1,:1}.$$

Similarly, for rows up to the ν^{th} row, i.e. for $i \leq \nu$, we have,

$$\begin{pmatrix} \Lambda_{i1} & \cdots & \Lambda_{ii} \end{pmatrix} = \begin{pmatrix} T_{i1} & \cdots & T_{ii} \end{pmatrix} \begin{pmatrix} B_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ B_{i1} & \cdots & B_{ii} \end{pmatrix},$$

which can be written as

$$\mathbf{\Lambda}_{i,:i} = \mathbf{T}_{i,:i}\mathbf{B}_{:i,:i}.$$

For rows beyond the ν^{th} row, i.e., $i > \nu$, the expression becomes,

$$\begin{pmatrix} \Lambda_{i1} & \cdots & \Lambda_{i\nu} \end{pmatrix} = \begin{pmatrix} T_{i1} & \cdots & T_{i\nu} \end{pmatrix} \begin{pmatrix} B_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ B_{\nu 1} & \cdots & B_{\nu\nu} \end{pmatrix},$$

which again can be written as,

$$\mathbf{\Lambda}_{i,:\nu} = \mathbf{T}_{i,:\nu}\mathbf{B}_{:\nu,:\nu} = \mathbf{T}_{i,:}\mathbf{B}.$$

To calculate the Jacobian, we proceed by taking the transpose of each of the rows and stacking them, giving,

$$\begin{pmatrix} \mathbf{\Lambda}_{1,:1}^\top \\ \mathbf{\Lambda}_{2,:2}^\top \\ \vdots \\ \mathbf{\Lambda}_{\nu,:\nu}^\top \\ \mathbf{\Lambda}_{\nu+1,:\nu}^\top \\ \vdots \\ \mathbf{\Lambda}_{P,:\nu}^\top \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{:1,:1}^\top & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{:2,:2}^\top & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}^\top & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}^\top & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{B}^\top \end{pmatrix} \begin{pmatrix} \mathbf{T}_{1,:1}^\top \\ \mathbf{T}_{2,:2}^\top \\ \vdots \\ \mathbf{T}_{\nu,:\nu}^\top \\ \mathbf{T}_{\nu+1,:\nu}^\top \\ \vdots \\ \mathbf{T}_{P,:\nu}^\top \end{pmatrix}.$$

Since the square matrix is upper triangular, its determinant is simply the product of the elements of its diagonal. This gives the Jacobian determinant,

$$\left|\frac{\partial\mathbf{\Lambda}}{\partial\mathbf{T}}\right| = \prod_{i=1}^{\tilde{\nu}} B_{ii}^{P-i+1}. \tag{17}$$

A.6 JACOBIAN FOR $\mathbf{C} = \mathbf{\Lambda}\mathbf{\Lambda}^\top \mapsto \mathbf{D} = \mathbf{A}\mathbf{C}\mathbf{A}^T$, WHERE A IS AN INVERTIBLE MATRIX

Now consider the transformation $\mathbf{C} = \mathbf{\Lambda}\mathbf{\Lambda}^\top \mapsto \mathbf{D} = \mathbf{A}\mathbf{C}\mathbf{A}^T$, where $\mathbf{\Lambda} \in \mathbb{R}^{P\times\nu}$ is lower triangular with rank ν, and $\mathbf{A} \in \mathbb{R}^{P\times P}$ is invertible. This Jacobian is difficult to derive from scratch; however, we can obtain it using the density of the singular Wishart. In particular, the probability density function of $\mathbf{D} \sim \mathcal{W}(\mathbf{S}, \nu)$ is given by,

$$\mathrm{P}_1(\mathbf{D}) = \frac{\pi^{\nu(\tilde{\nu}-P)/2}}{2^{\nu P/2}\left|\mathbf{S}\right|^{\nu/2}\Gamma_{\tilde{\nu}}\left(\frac{\nu}{2}\right)}\left|\mathbf{D}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}\operatorname{etr}\left(-\mathbf{S}^{-1}\mathbf{D}/2\right),$$

where $\tilde{\nu} = \min(\nu, P)$ as before. Note that $\mathbf{D}_{:\tilde{\nu},:\tilde{\nu}}$ is almost surely full rank. For $\mathbf{C} \sim \mathcal{W}(\mathbf{I}_P, \nu)$, this simplifies to,

$$\mathrm{P}_2(\mathbf{C}) = \frac{\pi^{\nu(\tilde{\nu}-P)/2}}{2^{\nu P/2}\Gamma_{\tilde{\nu}}\left(\frac{\nu}{2}\right)}\left|\mathbf{C}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}\operatorname{etr}\left(-\mathbf{C}/2\right).$$

Using these densities, we can use the identity,

$$\mathrm{P}_1(\mathbf{D}) = \mathrm{P}_2(\mathbf{C})\left|\frac{\partial\mathbf{C}}{\partial\mathbf{D}}\right|,$$

to obtain the desired Jacobian determinant,

$$\left|\frac{\partial\mathbf{D}}{\partial\mathbf{C}}\right| = \mathrm{P}_2(\mathbf{C}) / \mathrm{P}_1(\mathbf{D}) = \left|\mathbf{A}\mathbf{A}^T\right|^{\nu/2}\frac{\left|\mathbf{C}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}{\left|\mathbf{D}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}} = \left|\mathbf{A}\right|^{\nu}\frac{\left|\mathbf{C}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}{\left|\mathbf{D}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}. \tag{18}$$

We can now put these Jacobian determinant results together to derive the densities for the A- and AB-generalised (singular) Wishart distributions.

A.7 THE A-GENERALISED (SINGULAR) WISHART DENSITY

In Section 5.1 we said that $\mathbf{G} = \mathbf{A}\mathbf{T}(\mathbf{A}\mathbf{T})^T \sim \text{A-}\mathcal{GW}(\mathbf{A}, \nu, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma})$ if $\mathbf{A} \in \mathbb{R}^{P\times P}$ is invertible and $\mathbf{T} \in \mathbb{R}^{P\times\nu}$ is distributed according to (11). If we define $\mathbf{C}$ such that $\mathbf{G} = \mathbf{A}\mathbf{C}\mathbf{A}^T = \mathbf{A}\mathbf{T}\mathbf{T}^T\mathbf{A}^T$, then by the change of variables formula for probability densities,

$$\mathrm{Q}(\mathbf{G}) = \mathrm{Q}(\mathbf{T})\left|\frac{\partial\mathbf{T}}{\partial\mathbf{C}}\right|\left|\frac{\partial\mathbf{C}}{\partial\mathbf{G}}\right|. \tag{19}$$

By combining the density of $\mathbf{T}$,

$$\mathrm{Q}(\mathbf{T}) = 2^{\tilde{\nu}}\prod_{j=1}^{\tilde{\nu}} T_{jj}\,\mathrm{Gamma}\left(T_{jj}^2;\,\alpha_j,\beta_j\right)\prod_{i=j+1}^{P}\mathcal{N}\left(T_{ij};\mu_{ij},\sigma_{ij}^2\right), \tag{20}$$

the result from (14),

$$\left|\frac{\partial\mathbf{C}}{\partial\mathbf{T}}\right| = 2^{\tilde{\nu}}\prod_{j=1}^{\tilde{\nu}} T_{jj}^{P-j+1}, \tag{21}$$

and the result from (18),

$$\left|\frac{\partial\mathbf{G}}{\partial\mathbf{C}}\right| = \left|\mathbf{A}\right|^{\nu}\frac{\left|\mathbf{C}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}{\left|\mathbf{G}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}, \tag{22}$$

we obtain the A-generalised (singular) Wishart density,

$$\mathrm{Q}(\mathbf{G}) = \frac{\left|\mathbf{G}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}{\left|\mathbf{A}\right|^{\nu}\left|\mathbf{C}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}\prod_{j=1}^{\tilde{\nu}}\frac{\mathrm{Gamma}\left(T_{jj}^2;\,\alpha_j,\beta_j\right)}{T_{jj}^{P-j}}\prod_{i=j+1}^{P}\mathcal{N}\left(T_{ij};\mu_{ij},\sigma_{ij}^2\right). \tag{23}$$

A.8 THE AB-GENERALISED (SINGULAR) WISHART DENSITY

The derivation for the AB-generalised (singular) Wishart is similar to that of the A-generalised (singular) Wishart, with the addition of one extra step. Namely, as the AB-generalised (singular) Wishart defines $\mathbf{G} = \mathbf{ATB}(\mathbf{ATB})^T$, we define $\mathbf{\Lambda} = \mathbf{TB}$ and $\mathbf{C} = \mathbf{\Lambda}\mathbf{\Lambda}^T$, so that,

$$\mathrm{Q}\left(\mathbf{G}\right) = \mathrm{Q}\left(\mathbf{T}\right)\left|\frac{\partial \mathbf{T}}{\partial \mathbf{\Lambda}}\right|\left|\frac{\partial \mathbf{\Lambda}}{\partial \mathbf{C}}\right|\left|\frac{\partial \mathbf{C}}{\partial \mathbf{D}}\right|.$$

This first Jacobian determinant can be obtained using (17),

$$\left|\frac{\partial \mathbf{T}}{\partial \mathbf{\Lambda}}\right| = \prod_{i=1}^{\tilde{\nu}} \frac{1}{B_{ii}^{P-i+1}},$$

whereas the second,

$$\left|\frac{\partial \mathbf{\Lambda}}{\partial \mathbf{C}}\right| = \frac{1}{2^{\tilde{\nu}}}\prod_{i=1}^{\tilde{\nu}} \frac{1}{\Lambda_{ii}^{P-i+1}} = \frac{1}{2^{\tilde{\nu}}}\prod_{i=1}^{\tilde{\nu}} \frac{1}{T_{ii}^{P-i+1}B_{ii}^{P-i+1}},$$

arises from (14). The remaining Jacobians remain unchanged in form, so that our final density is given by,

$$\mathrm{Q}\left(\mathbf{G}\right) = \frac{\left|\mathbf{G}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}{\left|\mathbf{A}\right|^{\nu}\left|\mathbf{C}_{:\tilde{\nu},:\tilde{\nu}}\right|^{(\nu-P-1)/2}}\prod_{j=1}^{\tilde{\nu}}\frac{\mathrm{Gamma}\left(T_{jj}^2;\,\alpha_j,\beta_j\right)}{T_{jj}^{P-j}B_{jj}^{2(P-j+1)}}\prod_{i=j+1}^{P}\mathcal{N}\left(T_{ij};\mu_{ij},\sigma_{ij}^2\right). \tag{24}$$

B DETAILED EXPERIMENTAL RESULTS

All models were trained on the UCI splits from Gal and Ghahramani [2016], of which there are 20 for each dataset apart from PROTEIN. The datasets and the splits are available at `https://github.com/yaringal/DropoutUncertaintyExps/tree/master/UCI_Datasets`. Deep Wishart processes with the three kinds of approximate posterior ($\mathcal{GW}$, A-$\mathcal{GW}$, and AB-$\mathcal{GW}$) were trained, with number of layers $\ell \in \{2,\ldots,5\}$, and width ν_ℓ fixed to the number of input features. We applied the squared exponential kernel as a non-linearity at each layer, with automatic relevance determination (ARD, Williams and Rasmussen [2006]) in the first layer only. The DGPs trained reflected this architecture, with each GP layer returning features with dimension equal to the number of input features. In particular the DGPs were trained using global inducing point methods [Ober and Aitchison, 2021b]. The final layer of the DWP also uses a global inducing approximate posterior [Ober and Aitchison, 2021b].

All models were trained using the same scheme. $20\,000$ gradient steps were used to train each model, with the ADAM optimizer Kingma and Ba [2015]. We began with an initial learning rate of 10^{-2}, and then stepped the learning rate down to 10^{-3} after $10\,000$ gradient steps. The KL was annealed using a factor increasing linearly from 0 to 1 over the first $1\,000$ gradient steps. No pre-processing of the data was performed, other than normalizing inputs and outputs. To train, 10 samples were drawn from the approximate posterior, and to test 100 samples were drawn. For the smaller datasets (BOSTON, CONCRETE, ENERGY, WINE, YACHT), training was performed on a CPU (Intel Core i9-10900X), and for the other (larger) datasets, an internal cluster of machines was used, with NVIDIA GeForce 2080 Ti GPUs.

B.1 TABLES

Tables 1 to 4 report the ELBOs, test log likelihoods, and RMSEs from our UCI experiments respectively. In all cases, we give the mean of each metric (plus or minus one standard error), and highlight the model with the best mean value in bold for each configuration (unless all are equal).

Table 1: ELBOs per datapoint. We report mean plus or minus one standard error over the splits. Bold numbers correspond to the best models overall.

{Dataset}-{Depth}	DGP	DWP $\mathrm{Q}_{\mathcal{GW}}$	DWP $\mathrm{Q}_{\text{A-}\mathcal{GW}}$	DWP $\mathrm{Q}_{\text{AB-}\mathcal{GW}}$
BOSTON - 2	-0.38 ± 0.01	-0.33 ± 0.00	**-0.32 ± 0.01**	**-0.32 ± 0.00**
3	-0.40 ± 0.00	-0.34 ± 0.01	**-0.33 ± 0.00**	**-0.33 ± 0.01**
4	-0.43 ± 0.00	**-0.35 ± 0.00**	**-0.34 ± 0.01**	**-0.34 ± 0.01**
5	-0.45 ± 0.00	**-0.37 ± 0.01**	**-0.36 ± 0.00**	**-0.36 ± 0.00**
CONCRETE - 2	-0.45 ± 0.00	-0.42 ± 0.00	-0.40 ± 0.00	**-0.39 ± 0.00**
3	-0.47 ± 0.00	-0.43 ± 0.00	**-0.41 ± 0.00**	**-0.41 ± 0.00**
4	-0.49 ± 0.00	-0.46 ± 0.00	**-0.43 ± 0.00**	**-0.43 ± 0.00**
5	-0.50 ± 0.00	-0.49 ± 0.00	**-0.45 ± 0.00**	**-0.45 ± 0.00**
ENERGY - 2	1.43 ± 0.00	**1.46 ± 0.00**	**1.46 ± 0.00**	**1.46 ± 0.00**
3	1.42 ± 0.00	1.44 ± 0.00	**1.45 ± 0.00**	**1.45 ± 0.00**
4	1.40 ± 0.00	1.42 ± 0.00	**1.43 ± 0.00**	**1.43 ± 0.00**
5	1.38 ± 0.00	1.40 ± 0.00	**1.42 ± 0.00**	1.41 ± 0.00
KIN8NM - 2	-0.15 ± 0.00	-0.16 ± 0.00	**-0.14 ± 0.00**	**-0.14 ± 0.00**
3	-0.14 ± 0.00	-0.15 ± 0.00	**-0.13 ± 0.00**	**-0.13 ± 0.00**
4	-0.14 ± 0.00	-0.14 ± 0.00	**-0.11 ± 0.00**	**-0.11 ± 0.00**
5	-0.14 ± 0.00	-0.14 ± 0.00	**-0.11 ± 0.00**	**-0.11 ± 0.00**
NAVAL - 2	3.93 ± 0.05	3.82 ± 0.09	3.80 ± 0.13	3.84 ± 0.10
3	3.83 ± 0.06	3.71 ± 0.12	3.86 ± 0.06	**3.99 ± 0.04**
4	**3.91 ± 0.05**	3.66 ± 0.13	**3.75 ± 0.11**	**3.85 ± 0.09**
5	**3.92 ± 0.04**	3.59 ± 0.12	**3.97 ± 0.02**	3.63 ± 0.22
POWER - 2	0.03 ± 0.00	0.03 ± 0.00	**0.04 ± 0.00**	**0.04 ± 0.00**
3	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00
4	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.03 ± 0.00
5	**0.03 ± 0.00**	0.02 ± 0.00	**0.03 ± 0.00**	**0.03 ± 0.00**
PROTEIN - 2	**-1.06 ± 0.00**	-1.07 ± 0.00	**-1.06 ± 0.00**	**-1.06 ± 0.00**
3	-1.04 ± 0.00	-1.04 ± 0.00	**-1.03 ± 0.00**	**-1.03 ± 0.00**
4	-1.02 ± 0.00	-1.02 ± 0.00	**-1.00 ± 0.00**	-1.01 ± 0.00
5	**-1.00 ± 0.00**	-1.01 ± 0.00	**-1.00 ± 0.00**	**-1.00 ± 0.00**
WINE - 2	-1.18 ± 0.00	-1.18 ± 0.00	**-1.18 ± 0.00**	**-1.18 ± 0.00**
3	-1.19 ± 0.00	**-1.18 ± 0.00**	**-1.18 ± 0.00**	**-1.18 ± 0.00**
4	-1.19 ± 0.00	**-1.18 ± 0.00**	**-1.18 ± 0.00**	**-1.18 ± 0.00**
5	-1.19 ± 0.00	-1.19 ± 0.00	-1.19 ± 0.00	-1.19 ± 0.00
YACHT - 2	1.88 ± 0.03	2.02 ± 0.01	**2.07 ± 0.01**	**2.07 ± 0.01**
3	1.62 ± 0.01	1.86 ± 0.02	**2.02 ± 0.01**	**2.03 ± 0.01**
4	1.47 ± 0.02	1.73 ± 0.02	**1.93 ± 0.01**	1.91 ± 0.01
5	1.46 ± 0.02	1.59 ± 0.02	**1.79 ± 0.02**	**1.79 ± 0.02**

Table 2: ELBO differences per datapoint. We report mean differences plus or minus one standard error over the splits.

{Dataset}-{Depth}	$\mathrm{Q}_{\text{A-}\mathcal{GW}} - \mathrm{Q}_{\mathcal{GW}}$	$\mathrm{Q}_{\text{AB-}\mathcal{GW}} - \mathrm{Q}_{\mathcal{GW}}$	$\mathrm{Q}_{\text{A-}\mathcal{GW}} - \mathrm{Q}_{\text{AB-}\mathcal{GW}}$
BOSTON - 2	0.01 ± 0.01	0.01 ± 0.00	0.00 ± 0.01
3	0.01 ± 0.01	0.01 ± 0.01	0.00 ± 0.01
4	0.01 ± 0.01	0.01 ± 0.01	0.00 ± 0.01
5	0.01 ± 0.01	0.01 ± 0.01	0.00 ± 0.00
CONCRETE - 2	0.02 ± 0.00	0.03 ± 0.00	-0.01 ± 0.00
3	0.02 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
4	0.03 ± 0.00	0.03 ± 0.00	0.00 ± 0.00
5	0.04 ± 0.00	0.04 ± 0.00	0.00 ± 0.00
ENERGY - 2	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
3	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
4	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
5	0.02 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
KIN8NM - 2	0.02 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
3	0.02 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
4	0.03 ± 0.00	0.03 ± 0.00	0.00 ± 0.00
5	0.03 ± 0.00	0.03 ± 0.00	0.00 ± 0.00
NAVAL - 2	-0.02 ± 0.16	0.02 ± 0.13	-0.04 ± 0.16
3	0.15 ± 0.13	0.28 ± 0.13	-0.13 ± 0.07
4	0.09 ± 0.17	0.19 ± 0.16	-0.10 ± 0.14
5	0.38 ± 0.12	0.04 ± 0.25	0.34 ± 0.22
POWER - 2	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
3	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
4	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
5	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
PROTEIN - 2	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
3	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
4	0.02 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
5	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
WINE - 2	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
3	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
4	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
5	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
YACHT - 2	0.05 ± 0.01	0.05 ± 0.01	0.00 ± 0.01
3	0.16 ± 0.02	0.17 ± 0.02	-0.01 ± 0.01
4	0.20 ± 0.02	0.18 ± 0.02	0.02 ± 0.01
5	0.20 ± 0.03	0.20 ± 0.03	0.00 ± 0.03

Table 3: Average test log likelihoods. We report mean plus or minus one standard error over the splits. Bold numbers correspond to the best models overall.

{Dataset}-{Depth}	DGP	$Q_{\mathcal{GW}}$	DWP $Q_{\text{A-}\mathcal{GW}}$	$Q_{\text{AB-}\mathcal{GW}}$
BOSTON - 2	-2.43 ± 0.05	-2.40 ± 0.05	**-2.37 ± 0.05**	**-2.37 ± 0.05**
3	-2.39 ± 0.04	-2.38 ± 0.05	**-2.35 ± 0.04**	**-2.35 ± 0.04**
4	-2.41 ± 0.04	-2.38 ± 0.04	**-2.37 ± 0.04**	**-2.37 ± 0.04**
5	-2.43 ± 0.04	-2.38 ± 0.04	-2.39 ± 0.05	**-2.38 ± 0.04**
CONCRETE - 2	-3.10 ± 0.02	-3.12 ± 0.02	**-3.08 ± 0.02**	**-3.08 ± 0.02**
3	-3.08 ± 0.02	-3.10 ± 0.02	**-3.06 ± 0.02**	-3.07 ± 0.02
4	-3.13 ± 0.02	-3.12 ± 0.02	**-3.07 ± 0.02**	**-3.07 ± 0.02**
5	-3.13 ± 0.02	-3.13 ± 0.02	**-3.07 ± 0.02**	-3.08 ± 0.02
ENERGY - 2	-0.70 ± 0.03	-0.70 ± 0.03	-0.70 ± 0.03	-0.70 ± 0.03
3	-0.70 ± 0.03	-0.70 ± 0.03	-0.70 ± 0.03	-0.70 ± 0.03
4	-0.70 ± 0.03	-0.70 ± 0.03	-0.70 ± 0.03	-0.70 ± 0.03
5	-0.71 ± 0.03	-0.71 ± 0.03	**-0.70 ± 0.03**	**-0.70 ± 0.03**
KIN8NM - 2	1.35 ± 0.00	1.35 ± 0.00	**1.36 ± 0.00**	**1.36 ± 0.00**
3	1.37 ± 0.00	1.37 ± 0.00	**1.38 ± 0.00**	**1.38 ± 0.00**
4	1.38 ± 0.00	1.39 ± 0.01	**1.40 ± 0.00**	**1.40 ± 0.00**
5	1.38 ± 0.00	1.40 ± 0.01	**1.41 ± 0.01**	**1.41 ± 0.01**
NAVAL - 2	**8.24 ± 0.06**	8.23 ± 0.08	8.18 ± 0.11	8.18 ± 0.13
3	8.15 ± 0.06	8.18 ± 0.07	8.27 ± 0.05	**8.38 ± 0.03**
4	8.28 ± 0.04	8.17 ± 0.11	8.14 ± 0.13	**8.32 ± 0.06**
5	8.28 ± 0.04	8.17 ± 0.07	**8.40 ± 0.02**	8.10 ± 0.19
POWER - 2	-2.78 ± 0.01	-2.77 ± 0.01	**-2.76 ± 0.01**	**-2.76 ± 0.01**
3	-2.77 ± 0.01	**-2.76 ± 0.01**	**-2.76 ± 0.01**	**-2.76 ± 0.01**
4	-2.78 ± 0.01	-2.77 ± 0.01	**-2.75 ± 0.01**	**-2.75 ± 0.01**
5	-2.78 ± 0.01	-2.77 ± 0.01	**-2.76 ± 0.01**	**-2.76 ± 0.01**
PROTEIN - 2	-2.82 ± 0.00	**-2.81 ± 0.00**	**-2.81 ± 0.00**	**-2.81 ± 0.00**
3	-2.78 ± 0.00	-2.77 ± 0.00	**-2.76 ± 0.00**	**-2.76 ± 0.00**
4	-2.75 ± 0.00	-2.73 ± 0.00	**-2.72 ± 0.00**	-2.73 ± 0.01
5	-2.73 ± 0.01	-2.72 ± 0.01	-2.71 ± 0.01	**-2.70 ± 0.00**
WINE - 2	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01
3	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01
4	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01
5	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01	-0.96 ± 0.01
YACHT - 2	-0.29 ± 0.12	**-0.04 ± 0.10**	**-0.04 ± 0.08**	-0.08 ± 0.10
3	-0.63 ± 0.04	-0.13 ± 0.07	0.12 ± 0.07	**0.14 ± 0.06**
4	-0.77 ± 0.07	-0.26 ± 0.07	**-0.04 ± 0.09**	**-0.04 ± 0.09**
5	-0.73 ± 0.07	-0.58 ± 0.06	-0.22 ± 0.09	**-0.18 ± 0.07**

Table 4: Root mean square error. We report mean plus or minus one standard error over the splits. Bold numbers correspond to the best models overall.

{Dataset}-{Depth}	DGP	DWP $\mathrm{Q}_{\mathcal{GW}}$	DWP $\mathrm{Q}_{\text{A-}\mathcal{GW}}$	DWP $\mathrm{Q}_{\text{AB-}\mathcal{GW}}$
BOSTON - 2	2.72 ± 0.14	2.67 ± 0.14	2.60 ± 0.12	$\mathbf{2.59 \pm 0.13}$
3	2.73 ± 0.14	2.66 ± 0.13	$\mathbf{2.62 \pm 0.13}$	2.63 ± 0.13
4	2.76 ± 0.14	2.74 ± 0.15	2.71 ± 0.14	$\mathbf{2.68 \pm 0.14}$
5	2.81 ± 0.14	2.82 ± 0.17	$\mathbf{2.77 \pm 0.16}$	2.81 ± 0.17
CONCRETE - 2	5.41 ± 0.10	5.50 ± 0.12	$\mathbf{5.29 \pm 0.12}$	5.30 ± 0.12
3	5.31 ± 0.11	5.32 ± 0.10	$\mathbf{5.22 \pm 0.12}$	5.23 ± 0.12
4	5.54 ± 0.10	5.43 ± 0.11	5.24 ± 0.13	$\mathbf{5.22 \pm 0.13}$
5	5.49 ± 0.10	5.53 ± 0.10	5.26 ± 0.11	$\mathbf{5.24 \pm 0.11}$
ENERGY - 2	0.48 ± 0.01	0.48 ± 0.01	0.48 ± 0.01	0.48 ± 0.01
3	0.48 ± 0.01	0.48 ± 0.01	0.48 ± 0.01	0.48 ± 0.01
4	0.48 ± 0.01	0.48 ± 0.01	0.48 ± 0.01	0.48 ± 0.01
5	0.49 ± 0.01	$\mathbf{0.48 \pm 0.01}$	$\mathbf{0.48 \pm 0.01}$	$\mathbf{0.48 \pm 0.01}$
KIN8NM - 2	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.00	0.06 ± 0.00
3	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.00	0.06 ± 0.00
4	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.00	0.06 ± 0.00
5	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.00	0.06 ± 0.00
NAVAL - 2	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
3	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
4	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
5	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
POWER - 2	3.87 ± 0.04	3.83 ± 0.04	3.82 ± 0.04	$\mathbf{3.81 \pm 0.04}$
3	3.87 ± 0.03	3.82 ± 0.04	$\mathbf{3.81 \pm 0.04}$	$\mathbf{3.81 \pm 0.04}$
4	3.89 ± 0.04	3.84 ± 0.04	$\mathbf{3.78 \pm 0.04}$	$\mathbf{3.78 \pm 0.04}$
5	3.88 ± 0.04	3.84 ± 0.04	$\mathbf{3.80 \pm 0.04}$	$\mathbf{3.80 \pm 0.04}$
PROTEIN - 2	4.08 ± 0.01	4.06 ± 0.01	$\mathbf{4.05 \pm 0.02}$	$\mathbf{4.05 \pm 0.01}$
3	3.92 ± 0.02	3.90 ± 0.01	3.88 ± 0.01	$\mathbf{3.87 \pm 0.01}$
4	3.82 ± 0.01	3.79 ± 0.01	$\mathbf{3.75 \pm 0.01}$	3.79 ± 0.02
5	3.77 ± 0.02	3.76 ± 0.02	3.73 ± 0.02	$\mathbf{3.70 \pm 0.01}$
WINE - 2	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01
3	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01
4	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01
5	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01	0.63 ± 0.01
YACHT - 2	0.41 ± 0.04	$\mathbf{0.33 \pm 0.03}$	$\mathbf{0.33 \pm 0.03}$	$\mathbf{0.33 \pm 0.03}$
3	0.53 ± 0.03	0.35 ± 0.03	0.31 ± 0.03	$\mathbf{0.30 \pm 0.03}$
4	0.58 ± 0.05	0.41 ± 0.04	$\mathbf{0.33 \pm 0.03}$	$\mathbf{0.33 \pm 0.03}$
5	0.57 ± 0.05	0.50 ± 0.04	$\mathbf{0.37 \pm 0.03}$	0.38 ± 0.03

References

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

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.

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.

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.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*. MIT press Cambridge, MA, 2006.