Collaborative Multi-output Gaussian Processes: Supplementary Material

Trung V. Nguyen
ANU & NICTA
Canberra, Australia

Edwin V. Bonilla
NICTA & ANU
Sydney, Australia

1 Gaussian Identities

Let $\mathbf{g} = \{\mathbf{g}_j\}_{j=1}^q$, $\mathbf{h}$, and $\mathbf{y}$ be random variables with multivariate Gaussian distributions: $p(\mathbf{y}|\mathbf{g},\mathbf{h}) = \mathcal{N}(\mathbf{y}; \sum_{j=1}^Q \mathbf{W}_j \mathbf{g}_j + \mathbf{W}\mathbf{h}, \beta^{-1}\mathbf{I})$, $p(\mathbf{g}_j) = \mathcal{N}(\mathbf{g}_j; \mathbf{m}_j, \mathbf{S}_j)$, and $p(\mathbf{h}) = \mathcal{N}(\mathbf{h}; \mathbf{m}, \mathbf{S})$. The following identity is important in deriving the evidence lower bound:

$$\int \log p(\mathbf{y}|\mathbf{g},\mathbf{h}) \prod_{j=1}^q p(\mathbf{g}_j) p(\mathbf{h}) \mathrm{d}\mathbf{g}\mathrm{d}\mathbf{h}$$
$$= \log \mathcal{N}(\mathbf{y}; \sum_{j=1}^q \mathbf{W}_j \mathbf{m}_j + \mathbf{W}\mathbf{m}, \beta^{-1}\mathbf{I})$$
$$- \frac{1}{2}\beta \text{ tr } \mathbf{W}^T\mathbf{W}\mathbf{S} - \frac{1}{2}\beta \text{ tr } \sum_{j=1}^q \mathbf{W}_j^T \mathbf{W}_j \mathbf{S}_j. \quad (1)$$

The identity can be proved by using this fact:

$$\int (\mathbf{W}\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{W}\mathbf{x} - \boldsymbol{\mu}) \mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{S}) \mathrm{d}\mathbf{x}$$
$$= (\boldsymbol{\mu} - \mathbf{W}\mathbf{m})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{W}\mathbf{m}) + \text{ tr } \mathbf{W}^T \boldsymbol{\Sigma}^{-1} \mathbf{W}\mathbf{S}. \quad (2)$$

2 Derivation of the Variational Lower Bound

The variational lower bound of the log marginal (eq. 13 in the main text) is given by:

$$\log p(\mathbf{y}) \geq \int q(\mathbf{u},\mathbf{v}) \log \frac{p(\mathbf{y}|\mathbf{u},\mathbf{v}) p(\mathbf{u},\mathbf{v})}{q(\mathbf{u},\mathbf{v})} \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}$$
$$+ \int q(\mathbf{u},\mathbf{v}) \log \frac{p(\mathbf{u},\mathbf{v})}{q(\mathbf{u},\mathbf{v})} \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}$$
$$= \int q(\mathbf{u},\mathbf{v}) \log p(\mathbf{y}|\mathbf{u},\mathbf{v}) \mathrm{d}\mathbf{u}\mathrm{d}\mathbf{v}$$
$$- \sum_{j=1}^Q \mathrm{KL}[q(\mathbf{u}_j)||p(\mathbf{u}_j)] - \sum_{i=1}^P \mathrm{KL}[q(\mathbf{v}_i)||p(\mathbf{v}_i)],$$

Since the KL terms are analytically tractable, we compute the first term in the above sum. This is done first by deriving a lower bound to the likelihood $p(\mathbf{y}|\mathbf{u},\mathbf{v})$ (eq. 14 in the main text).

$$\log\ p(\mathbf{y}|\mathbf{u},\mathbf{v}) = \log \langle p(\mathbf{y}|\mathbf{g},\mathbf{h}) \rangle_{p(\mathbf{g},\mathbf{h}|\mathbf{u},\mathbf{v})}$$
$$\geq \langle \log p(\mathbf{y}|\mathbf{g},\mathbf{h}) \rangle_{p(\mathbf{g},\mathbf{h}|\mathbf{u},\mathbf{v})}$$
$$= \sum_{i=1}^P \sum_{n=1}^N \langle \log p(y_{in}|\mathbf{g}_n, h_{in}) \rangle_{p(\mathbf{g}|\mathbf{u}) p(\mathbf{h}_i|\mathbf{v}_i)}$$

Applying the identity in eq. 1, the expectation of an individual likelihood term with respect to the posterior distribution is given by:

$$l_{in} = \int \log p(y_{in}|\mathbf{g}_n, h_{in}) \prod_{j=1}^Q p(g_{jn}|\mathbf{u}_j) p(h_{in}|\mathbf{u}_i) \mathrm{d}\mathbf{g}_n \mathrm{d}h_{in}$$
$$= \log \mathcal{N}(y_{in}; \sum_{j=1}^Q w_{ij}\mu_{jn} + \mu_{in}^h, \beta_i^{-1})$$
$$- \frac{1}{2}\beta_i \sum_{j=1}^Q w_{ij}^2 \tilde{k}_{jnn} - \frac{1}{2}\beta_i \tilde{k}_{inn}^h, \quad (3)$$

where $\tilde{k}_{jnn} = (\tilde{\mathbf{K}}_j)_{nn}$, $\tilde{k}_{inn}^h = (\tilde{\mathbf{K}}_i^h)_{nn}$, $\mu_{jn} = (\boldsymbol{\mu}_j)_n$, and $\mu_{in}^h = (\boldsymbol{\mu}_i^h)_n$.

Substituting l_{in} into the expression for the lower bound of $\log p(\mathbf{y}|\mathbf{u},\mathbf{v})$ (again, this is eq. 14 in the main text), and applying the identity in eq. 1 to carry out the integral we obtain the lower bound as given in the main text.

3 Derivatives of the ELBO

For exposition, we derive the gradients of the lower bound for the case of a single GP (i.e. the bound in Hensman et al. [2013]). The derivatives of the ELBO of the collaborative multioutput GPs model are typically linear combination of the derivatives here. The lower bound as a function of all parameters is

$$\begin{aligned}\mathcal{L} =& \log \mathcal{N}(\mathbf{y}; \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{m}, \beta^{-1}\mathbf{I}) \\ & - \frac{1}{2}\beta \text{ tr } \tilde{\mathbf{K}} - \frac{1}{2}\beta \text{ tr } (\mathbf{S}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN}\mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}) \\ & - \frac{1}{2}\left(\log |\mathbf{K}_{MM}| + \text{ tr } (\mathbf{K}_{MM}^{-1}(\mathbf{m}\mathbf{m}^T + \mathbf{S}))\right) \\ =& \underbrace{\log \mathcal{N}(\mathbf{y}; \mathbf{A}\mathbf{m}, \beta^{-1}\mathbf{I})}_{\mathcal{L}_1} - \underbrace{\frac{1}{2}\beta \text{ tr } (\mathbf{K}_{NN} - \mathbf{A}\mathbf{K}_{MN})}_{\mathcal{L}_2} \\ & \underbrace{- \frac{1}{2}\left(\log |\mathbf{K}_{MM}| + \text{ tr } (\mathbf{K}_{MM}^{-1}(\mathbf{m}\mathbf{m}^T + \mathbf{S}))\right)}_{\mathcal{L}_4} \\ & \underbrace{- \frac{1}{2}\beta \text{ tr } (\mathbf{S}\mathbf{A}^T\mathbf{A})}_{\mathcal{L}_3}, \end{aligned} \tag{4}$$

where $\mathbf{A} = \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}$. Here $\mathbf{K}_{NM} = k(\mathbf{X}, \mathbf{Z})$ is the cross-covariance matrix between the observed inputs and the inducing inputs and $\mathbf{K}_{MM} = k(\mathbf{Z}, \mathbf{Z})$ is the auto-covariance matrix between the inducing inputs. Notice that we have re-written the sum of individual terms in matrix form which will make the derivation easier and also the computation faster via vectorization.

3.1 Derivative of the Noise Hyperparameter

The derivative of the noise hyperparameter β is easily computed as:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{m})^T(\mathbf{y} - \mathbf{A}\mathbf{m}) - \frac{\mathcal{L}_2}{\beta} - \frac{\mathcal{L}_3}{\beta}. \tag{5}$$

3.2 Derivatives of the Covariance Hyperparameters

To simplify the math, we utilize the matrix $\mathbf{A}$ defined above. Firstly, the derivative of $\mathbf{A}$ wrt a covariance hyperparameter t is given by:

$$\frac{\partial \mathbf{A}}{\partial t} = \left(\frac{\partial \mathbf{K}_{NM}}{\partial t} - \mathbf{A}\frac{\partial \mathbf{K}_{MM}}{\partial t}\right)\mathbf{K}_{MM}^{-1}. \tag{6}$$

The derivatives of $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$ and $\mathcal{L}_4$ are thus given by:

$$\frac{\partial \mathcal{L}_1}{\partial t} = \beta(\mathbf{y} - \mathbf{A}\mathbf{m})^T\frac{\partial \mathbf{A}}{\partial t}\mathbf{m} \tag{7}$$

$$\frac{\partial \mathcal{L}_2}{\partial t} = \frac{1}{2}\beta \text{ tr } \left(\frac{\partial \mathbf{K}_{NN}}{\partial t} - \mathbf{A}\frac{\partial \mathbf{K}_{MN}}{\partial t} - \frac{\partial \mathbf{A}}{\partial t}\mathbf{K}_{MN}\right) \tag{8}$$

$$\frac{\partial \mathcal{L}_3}{\partial t} = \beta \text{ tr } \left(\mathbf{A}\mathbf{S}\frac{\partial \mathbf{A}^T}{\partial t}\right) \tag{9}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_4}{\partial t} =& \frac{1}{2} \text{ tr } \left(\mathbf{K}_{MM}^{-1}\frac{\partial \mathbf{K}_{MM}}{\partial t}\right) \\ & - \frac{1}{2} \text{ tr } \left(\mathbf{K}_{MM}^{-1}\frac{\partial \mathbf{K}_{MM}}{\partial t}\mathbf{K}_{MM}^{-1}(\mathbf{m}\mathbf{m}^T + \mathbf{S})\right)\end{aligned} \tag{10}$$

The derivatives are then computed by taking the derivatives of the covariance matrices $\mathbf{K}_{NN}$(the diagonal only), $\mathbf{K}_{NM}$ and $\mathbf{K}_{MM}$, hence the covariance function, wrt the hyperparameters.

3.3 Derivatives of the Inducing Inputs

To compute the derivatives of $\mathcal{L}$ wrt the inducing inputs, first notice that $\mathbf{Z} = \{\mathbf{z}_m\}_{m=1}^M$ can also be viewed as parameters of the covariance matrices $\mathbf{K}_{NM}$ and $\mathbf{K}_{MM}$. Hence the derivative wrt a single dimension of an inducing input, i.e. z_{mj}, is the same as that of $\frac{\partial \mathcal{L}}{\partial t}$.

We re-write $\frac{\partial \mathcal{L}_1}{\partial t}$, $\frac{\partial \mathcal{L}_2}{\partial t}$, $\frac{\partial \mathcal{L}_3}{\partial t}$, $\frac{\partial \mathcal{L}_4}{\partial t}$ by expanding $\frac{\partial \mathbf{A}}{\partial t}$ (here $t = z_{mj}$):

$$\begin{aligned}\frac{\partial \mathcal{L}_1}{\partial t} =& \beta \text{ tr } (\mathbf{y} - \mathbf{A}\mathbf{m})^T\left(\frac{\partial \mathbf{K}_{NM}}{\partial t} - \mathbf{A}\frac{\partial \mathbf{K}_{MM}}{\partial t}\right)\mathbf{K}_{MM}^{-1}\mathbf{m} \\ =& \beta \text{ tr } \mathbf{K}_{MM}^{-1}\mathbf{m}(\mathbf{y} - \mathbf{A}\mathbf{m})^T\frac{\partial \mathbf{K}_{NM}}{\partial t} \\ & - \beta \text{ tr } \mathbf{K}_{MM}^{-1}\mathbf{m}(\mathbf{y} - \mathbf{A}\mathbf{m})^T\mathbf{A}\frac{\partial \mathbf{K}_{MM}}{\partial t}\end{aligned} \tag{11}$$

$$\frac{\partial \mathcal{L}_2}{\partial t} = -\beta \text{ tr } \mathbf{A}^T\frac{\partial \mathbf{K}_{NM}}{\partial t} + \frac{1}{2}\beta \text{ tr } \mathbf{A}^T\mathbf{A}\frac{\partial \mathbf{K}_{MM}}{\partial t} \tag{12}$$

$$\frac{\partial \mathcal{L}_3}{\partial t} = \beta \text{ tr } \mathbf{K}_{MM}^{-1}\mathbf{S}\mathbf{A}^T\frac{\partial \mathbf{K}_{NM}}{\partial t} - \beta \text{ tr } \mathbf{K}_{MM}^{-1}\mathbf{S}\mathbf{A}^T\mathbf{A}\frac{\partial \mathbf{K}_{MM}}{\partial t} \tag{13}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_4}{\partial t} =& \frac{1}{2} \text{ tr } \mathbf{K}_{MM}^{-1}\frac{\partial \mathbf{K}_{MM}}{\partial t} \\ & - \frac{1}{2} \text{ tr } \mathbf{K}_{MM}^{-1}(\mathbf{m}\mathbf{m}^T + \mathbf{S})\mathbf{K}_{MM}^{-1}\frac{\partial \mathbf{K}_{MM}}{\partial t}\end{aligned} \tag{14}$$

From the above 4 equations we get,

$$\frac{\partial \mathcal{L}}{\partial t} = \text{ tr } \mathbf{D}_1\frac{\partial \mathbf{K}_{NM}}{\partial t} + \text{ tr } \mathbf{D}_2\frac{\partial \mathbf{K}_{MM}}{\partial t}, \tag{15}$$

where

$$\mathbf{D}_1 = \beta\mathbf{K}_{MM}^{-1}\mathbf{m}(\mathbf{y} - \mathbf{A}\mathbf{m})^T + \beta\mathbf{A}^T - \beta\mathbf{K}_{MM}^{-1}\mathbf{S}\mathbf{A}^T \tag{16}$$

$$\begin{aligned}\mathbf{D}_2 =& -\beta \text{ tr } \mathbf{K}_{MM}^{-1}\mathbf{m}(\mathbf{y} - \mathbf{A}\mathbf{m})^T\mathbf{A} - \frac{1}{2}\beta\mathbf{A}^T\mathbf{A} - \frac{1}{2}\mathbf{K}_{MM}^{-1} \\ & + \beta\mathbf{K}_{MM}^{-1}\mathbf{S}\mathbf{A}^T\mathbf{A} + \frac{1}{2}\mathbf{K}_{MM}^{-1}(\mathbf{m}\mathbf{m}^T + \mathbf{S})\mathbf{K}_{MM}^{-1}\end{aligned} \tag{17}$$

Notice that $\mathbf{D}_1$ and $\mathbf{D}_2$ can be pre-computed with a cost of $\mathcal{O}(M^3)$ (or $\mathcal{O}(N_b M^2)$ if the minibatch size $N_b > M$). The computational cost of taking derivatives of MD inducing parameters is thus $\mathcal{O}(M^3 + MDM) = \mathcal{O}(M^3)$ as the cost of the two trace operators is $\mathcal{O}(M)$ due to the fact that only $\mathcal{O}(M)$ elements of $\frac{\partial \mathbf{K}_{MM}}{\partial t}$ or $\frac{\partial \mathbf{K}_{NM}}{\partial t}$ are non-zero.

For implementation with MATLAB, a loop over $M \times D$ parameters of the inducing inputs can be very slow for even moderate values of M and D. The aforementioned fact about $\frac{\partial \mathbf{K}_{MM}}{\partial t}$ and $\frac{\partial \mathbf{K}_{NM}}{\partial t}$ can be used to perform vectorized operations that compute the derivatives of all M parameters given a specific dimension at a cost of $\mathcal{O}(M^2)$. The loop is the executed over the input dimension D, leading to a complexity of still $\mathcal{O}(DM^2)$ only.

References

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, 2013.