# Supplementary material for: Variational Heteroscedastic Gaussian Process Regression

In this extra material, we briefly mention additional details about the MV bound that were not included in the main text due to lack of space.

## A Relations that hold at maxima of the MV bound

Recall that the MV bound can be expressed as

$$F(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = F(q(\mathbf{g})) = \log Z(q(\mathbf{g})) - \mathrm{KL}(q(\mathbf{g})||p(\mathbf{g}))$$

where $q(\mathbf{g}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. At a local maximum of the MV bound, derivatives w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ vanish. Letting $\frac{\partial \log Z(q(\mathbf{g}))}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\boldsymbol{\Lambda}$ we have

$$\frac{\partial F(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\boldsymbol{\Lambda} + \frac{1}{2}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\mathbf{K}_g^{-1} = 0,$$

so that we know that, at the maximum of $F(\mu, \boldsymbol{\Sigma})$,

$$\boldsymbol{\Sigma} = (\mathbf{K}_g^{-1} + \boldsymbol{\Lambda})^{-1}$$

for some $\boldsymbol{\Lambda}$.

Now we will see which properties hold for $\boldsymbol{\Lambda}$. Using the definition of $Z(q(\mathbf{g}))$ and (4) from the main paper, we can express

$$Z(q(\mathbf{g})) = \int e^{\log \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{R}) - \frac{1}{4}\operatorname{tr}(\boldsymbol{\Sigma})} \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_f) \mathrm{d}\mathbf{f}$$

where $\mathbf{R}$ is a diagonal matrix with elements $[\mathbf{R}]_{ii} = e^{[\boldsymbol{\mu}]_i - [\boldsymbol{\Sigma}]_{ii}/2}$.

We can draw a first conclusion here: Since $Z(q(\mathbf{g}))$ only depends on the diagonal terms of $\boldsymbol{\Sigma}$, the off-diagonal terms of $\frac{\partial \log Z(q(\mathbf{g}))}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\boldsymbol{\Lambda}$ must be zero, and therefore $\boldsymbol{\Lambda}$ must be diagonal.

We will now show that all the elements of $\boldsymbol{\Lambda}$ are non-negative, i.e., that $-2\frac{\partial \log Z(q(\mathbf{g}))}{\partial [\boldsymbol{\Sigma}]_{ii}}$ is non-negative. Expanding this derivative, we have

$$-2\frac{\partial \log Z(q(\mathbf{g}))}{\partial [\boldsymbol{\Sigma}]_{ii}} = \frac{\int -2\frac{\partial (\log \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{R}) - \frac{1}{4}\operatorname{tr}(\boldsymbol{\Sigma}))}{\partial [\boldsymbol{\Sigma}]_{ii}} e^{\log \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{R}) - \frac{1}{4}\operatorname{tr}(\boldsymbol{\Sigma})} \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_f) \mathrm{d}\mathbf{f}}{Z(q(\mathbf{g}))}$$

The denominator is positive by definition, and if $-2\frac{\partial (\log \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{R}) - \frac{1}{4}\operatorname{tr}(\boldsymbol{\Sigma}))}{\partial [\boldsymbol{\Sigma}]_{ii}}$ is non-negative, the numerator is the expectation of a non-negative value, hence the whole expression is non-negative. We expand the critical term as

$$-2\frac{\partial (\log \mathcal{N}(\mathbf{y}|\mathbf{f}, \mathbf{R}) - \frac{1}{4}\operatorname{tr}(\boldsymbol{\Sigma}))}{\partial [\boldsymbol{\Sigma}]_{ii}} = \frac{1}{2}([\mathbf{y}]_i - [\mathbf{f}]_i)^2 e^{-[\boldsymbol{\mu}]_i + [\boldsymbol{\Sigma}]_{ii}/2},$$

which is obviously non-negative. Thus we have proved $\boldsymbol{\Lambda}$ must be both diagonal and non-negative.

The derivative of the MV bound w.r.t. $\boldsymbol{\mu}$ can be expressed in terms of its derivative w.r.t. $\boldsymbol{\Sigma}$. Specifically

$$\frac{\partial F(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = (\boldsymbol{\Lambda} - \tfrac{1}{2}\mathbf{I})\mathbf{1} - \mathbf{K}_g^{-1}(\boldsymbol{\mu} - \mu_0 \mathbf{1}) = 0$$

so that we know that, at the maximum of $F(\mu, \boldsymbol{\Sigma})$,

$$\boldsymbol{\mu} = \mathbf{K}_g(\boldsymbol{\Lambda} - \tfrac{1}{2}\mathbf{I})\mathbf{1} + \mu_0 \mathbf{1}$$

for some $\boldsymbol{\Lambda}$. Thus, we can define both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in terms of the $n$ non-negative elements of diagonal matrix $\boldsymbol{\Lambda}$ and express the MV bound as $F(\boldsymbol{\Lambda})$.

## B Derivatives of the MV bound

First, let us introduce the following definitions

$$\begin{aligned}
\boldsymbol{\alpha} &= (\mathbf{K}_f + \mathbf{R})^{-1}\mathbf{y} \\
\boldsymbol{\beta} &= (\boldsymbol{\Lambda} - \tfrac{1}{2}\mathbf{I})\mathbf{1} \\
[\overline{\boldsymbol{\Lambda}}]_{ii} &= \frac{1}{2}[(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - (\mathbf{K}_f + \mathbf{R})^{-1})\mathbf{R}]_{ii} + \frac{1}{2} \\
\overline{\boldsymbol{\beta}} &= (\overline{\boldsymbol{\Lambda}} - \tfrac{1}{2}\mathbf{I})\mathbf{1},
\end{aligned}$$

where $\overline{\boldsymbol{\Lambda}}$ is a diagonal matrix with the specified elements. Then, making explicit the dependence of the bound $F$ on the free variational parameters $\boldsymbol{\Lambda}$ and hyperparameters $\boldsymbol{\theta}$, derivatives can be computed analytically in terms of the selected covariance functions as:

$$\begin{aligned}
\frac{\partial F(\boldsymbol{\Lambda}, \boldsymbol{\theta})}{\partial [\boldsymbol{\Lambda}]_{ii}} &= [\mathbf{K}_g(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})]_i + \tfrac{1}{2}[\boldsymbol{\Sigma}(\overline{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda})\boldsymbol{\Sigma}]_{ii} \\
\frac{\partial F(\boldsymbol{\Lambda}, \boldsymbol{\theta})}{\partial \mu_0} &= \overline{\boldsymbol{\beta}}^\top \mathbf{1} \\
\frac{\partial F(\boldsymbol{\Lambda}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_f} &= \frac{1}{2}\operatorname{tr}\left(\frac{\partial \mathbf{K}_f}{\partial \boldsymbol{\theta}_f}[\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - (\mathbf{K}_f + \mathbf{R})^{-1}]\right) \\
\frac{\partial F(\boldsymbol{\Lambda}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_g} &= \frac{1}{2}\operatorname{tr}\Bigg(\frac{\partial \mathbf{K}_g}{\partial \boldsymbol{\theta}_g}[\boldsymbol{\beta}\boldsymbol{\beta}^\top - (\mathbf{K}_g + \boldsymbol{\Lambda}^{-1})^{-1} \\
&\qquad - \mathbf{K}_g^{-1}\boldsymbol{\Sigma}(\overline{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda})\boldsymbol{\Sigma}\mathbf{K}_g^{-1} + 2(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{\beta}^\top]\Bigg).
\end{aligned}$$

It is therefore possible to use gradient-based procedures, such as conjugate gradient, to jointly optimize the variational parameters and hyperparameters. The variational bound and its gradient can be computed in $\mathcal{O}(n^3)$ time.