Supplementary material for: Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning

In this extra material, we provide more details about the variational EM algorithm for multi-task and multiple kernel learning (Section 1) as well as the updates for the paired Gibbs sampler (Section 2).

1 Variational EM algorithm for multi-task and multiple kernel learning

The joint probability density function is

$$\begin{aligned} p(\mathbf{Y}, \widetilde{\mathbf{W}}, \mathbf{S}, \mathbf{\Phi}) &= \prod_{q=1}^{Q} \mathcal{N}(\mathbf{y}_q | \sum_{m=1}^{M} s_{qm} \widetilde{w}_{qm} \boldsymbol{\phi}_m, \sigma_q^2) \\ &\times \prod_{q=1}^{Q} \prod_{m=1}^{M} \left[\mathcal{N}(\widetilde{w}_{qm} | 0, \sigma_w^2) \pi^{s_{qm}} (1-\pi)^{s_{qm}} \right] \prod_{m=1}^{M} \mathcal{N}(\boldsymbol{\phi}_m | \mathbf{0}, \mathbf{K}_m), \end{aligned}$$

where the GP latent vector $\boldsymbol{\phi}_m \in \mathbb{R}^N$ and where we assumed zero-mean GPs for simplicity. The logarithm of the marginal likelihood is

$$\log p(\mathbf{Y}) = \log \sum_{\mathbf{S}} \int_{\mathbf{W}, \mathbf{\Phi}} p(\mathbf{Y}, \widetilde{\mathbf{W}}, \mathbf{S}, \mathbf{\Phi}) \, d\mathbf{W} \, d\mathbf{\Phi}.$$

The variational Bayesian method maximizes the following Jensen's lower bound on the above log marginal likelihood

$$\mathcal{F} = \sum_{\mathbf{S}} \int_{\mathbf{W}, \mathbf{\Phi}} q(\widetilde{\mathbf{W}}, \mathbf{S}, \mathbf{\Phi}) \log \frac{p(\mathbf{Y}, \widetilde{\mathbf{W}}, \mathbf{S}, \mathbf{\Phi})}{q(\widetilde{\mathbf{W}}, \mathbf{S}, \mathbf{\Phi})} \, d\mathbf{W} \, d\mathbf{\Phi},$$

where the variational distribution is assumed to factorize as follows

$$q(\widetilde{\mathbf{W}}, \mathbf{S}, \mathbf{\Phi}) = \prod_{q=1}^{Q} \prod_{m=1}^{M} q(\widetilde{w}_{qm}, s_{qm}) \prod_{m=1}^{M} q(\boldsymbol{\phi}_m).$$

In the next two sections we present a variational EM algorithm for the maximization of this lower bound. Section 1.1 describes the E-step updates and section 1.2 describes the M-step updates. The whole algorithm is a standard variational EM and all its updates are used by our implementation together with a specialized update presented in section 1.3. More precisely, as mentioned in the main paper separately updating the factor $q(\boldsymbol{\phi}_m)$ of the GP latent vector and the hyperparameters $\boldsymbol{\theta}_m$ of the covariance function of the same GP exhibits slow convergence. This is because of the strong dependence of the hyperparameters $\boldsymbol{\theta}_m$ on posterior $q(\boldsymbol{\phi}_m)$. Notice that an analogous problem arises when applying MCMC to GP models [1]. Section 1.3 shows how this problem can be solved by performing a joint update of $(q(\boldsymbol{\phi}_m), \boldsymbol{\theta}_m)$. Note that for clarity reasons we have made the choice to firstly present the regular EM updates and then the specialized step in order to gain a better understanding about the whole issue.

1.1 E-Step

The update for the factor $q(\widetilde{w}_{qm}, s_{qm})$ is such that $q(\widetilde{w}_{qm}, s_{qm}) = q(\widetilde{w}_{qm} | s_{qm}) q(s_{qm})$ where

$$\gamma_{qm} = q(s_{qm} = 1) = \frac{1}{1 + e^{-u_{qm}}}$$

$$u_{qm} = \log \frac{\pi}{1-\pi} + \frac{1}{2} \log \frac{\sigma_q^2}{\sigma_w^2} - \frac{1}{2} \log \left(\langle \boldsymbol{\phi}_m^{\mathrm{T}} \boldsymbol{\phi}_m \rangle + \frac{\sigma_q^2}{\sigma_w^2} \right) + \frac{1}{2\sigma_q^2} \frac{\left(\mathbf{y}_q^{\mathrm{T}} \langle \boldsymbol{\phi}_m \rangle - \sum_{k \neq m} \langle s_{qk} w_{qk} \rangle \langle \boldsymbol{\phi}_m^{\mathrm{T}} \rangle \langle \boldsymbol{\phi}_k \rangle \right)^2}{\left(\langle \boldsymbol{\phi}_m^{\mathrm{T}} \boldsymbol{\phi}_m \rangle + \frac{\sigma_q^2}{\sigma_w^2} \right)}$$

$$q(\widetilde{w}_{qm}|s_{qm}=0) = \mathcal{N}(\widetilde{w}_{qm}|0,\sigma_w^2)$$

$$\begin{aligned} q(\widetilde{w}_{qm}|s_{qm}=1) &= \mathcal{N}\left(\widetilde{w}_{qm}|\frac{\langle \boldsymbol{\phi}_m^{\mathrm{T}}\rangle \mathbf{y}_q - \sum_{k\neq m}\langle s_{qk}w_{qk}\rangle\langle \boldsymbol{\phi}_m^{\mathrm{T}}\rangle\langle \boldsymbol{\phi}_k\rangle}{\langle \boldsymbol{\phi}_m^{\mathrm{T}}\boldsymbol{\phi}_m\rangle + \frac{\sigma_q^2}{\sigma_w^2}}, \frac{\sigma_q^2}{\langle \boldsymbol{\phi}_m^{\mathrm{T}}\boldsymbol{\phi}_m\rangle + \frac{\sigma_q^2}{\sigma_w^2}}\right) \\ &= \mathcal{N}(\widetilde{w}_{qm}|\mu_{w_{qm}},\sigma^2_{w_{qm}}) \end{aligned} \quad (1)$$

So overall an update of $q(\widetilde{w}_{qm}, s_{qm})$ reduces to an update of the variational parameters $(\mu_{w_{qm}}, \sigma^2_{w_{qm}}, \gamma_{qm})$. In summary, $q(\widetilde{w}_{qm}, s_{qm})$ could be written as

$$q(\widetilde{w}_{qm}|s_{qm}) \times q(s_{qm}) = \mathcal{N}(\widetilde{w}_{qm}|s_{qm}\mu_{w_{qm}}, s_{qm}\sigma^2_{w_{qm}} + (1-s_{qm})\sigma_w^2) \times \gamma_{qm}^{s_{qm}}(1-\gamma_{qm})^{1-s_{qm}}.$$

Finally, note that under the distribution $q(\widetilde{w}_{qm}, s_{qm})$, the expectation $\langle s_{qm}w_{qm}\rangle = \gamma_{qm}\mu_{w_{qm}}$.

The variational update for each factor $q(\boldsymbol{\phi}_m)$ can be computed as

$$q(\boldsymbol{\phi}_m) = \mathcal{N}(\boldsymbol{\phi}_m|\boldsymbol{\mu}_{\phi_m}, \boldsymbol{\Sigma}_{\phi_m})$$

where

$$\boldsymbol{\Sigma}_{\phi_m} = \left(\sum_{q=1}^{Q}\frac{\langle s_{qm}\widetilde{w}_{qm}^2\rangle}{\sigma_q^2}\mathbf{I} + \mathbf{K}_m^{-1}\right)^{-1}$$

and

$$\boldsymbol{\mu}_{\phi_m} = \boldsymbol{\Sigma}_{\phi_m}\sum_{q=1}^{Q}\frac{\langle s_{qm}\widetilde{w}_{qm}\rangle}{\sigma_q^2}\left(\mathbf{y}_q - \sum_{k\neq m}\langle s_{qk}\widetilde{w}_{qk}\rangle\langle \boldsymbol{\phi}_k\rangle\right)$$

where $\langle s_{qm}\widetilde{w}_{qm}^2\rangle = \gamma_{qm}(\mu^2_{w_{qm}} + \sigma^2_{w_{qm}})$. Also the expectation $\langle \boldsymbol{\phi}_m^{\mathrm{T}}\boldsymbol{\phi}_m\rangle = \boldsymbol{\mu}_{\phi_m}^{\mathrm{T}}\boldsymbol{\mu}_{\phi_m} + \mathrm{tr}(\boldsymbol{\Sigma}_{\phi_m})$.

Notice that the update for $\boldsymbol{\Sigma}_{\phi_m}$ depends on the inverse $\mathbf{K}_m^{-1}$ which is not numerically stable as $\mathbf{K}_m$ in computer precision might not be invertible. This, however, is easily resolved by re-writing $\boldsymbol{\Sigma}_{\phi_m}$ as

$$\Sigma_{\phi_m} = \mathbf{K}_m(\alpha_m\mathbf{K}_m + I)^{-1},$$

where $\alpha_m = \sum_{q=1}^{Q}\frac{\langle s_{qm}\widetilde{w}_{qm}^2\rangle}{\sigma_q^2}$ is just a scalar. This now can be implemented in a symmetric and numerically stable way through the use of the Cholesky decomposition (and inverse Cholesky) of $(\alpha_m\mathbf{K}_m + I)$.

1.2 M-step

In the M-step, the bound is maximized w.r.t. hyperparameters $\{\{\sigma_q^2\}_{q=1}^{Q}, \sigma_w^2, \pi\}$ and the kernel hyperparameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_m\}_{m=1}^{M}$. The first set of hyperparameters is maximized using analytical updates. On the other hand, kernel hyperparameters require nonlinear gradient-based optimization.

The explicit form of the variational lower bound is

$$
\begin{aligned}
\mathcal{F} &= -\frac{QN}{2}\log(2\pi) - \frac{N}{2}\sum_{q=1}^{Q}\log(\sigma_q^2) - \frac{1}{2}\sum_{q=1}^{Q}\frac{\mathbf{y}_q^{\mathrm{T}}\mathbf{y}_q}{\sigma_q^2} && \%\boldsymbol{\mathcal{F}_1}\\
&+ \sum_{m=1}^{M}\left(\sum_{q=1}^{Q}\frac{\langle s_{qm}\widetilde{w}_{qm}\rangle}{\sigma_q^2}\mathbf{y}_q\right)^{\mathrm{T}}\langle\boldsymbol{\phi}_m\rangle && \%\boldsymbol{\mathcal{F}_2}\\
&- \frac{1}{2}\sum_{m=1}^{M}\left(\sum_{q=1}^{Q}\frac{\langle s_{qm}\widetilde{w}_{qm}^2\rangle}{\sigma_q^2}\right)\langle\boldsymbol{\phi}_m\boldsymbol{\phi}_m^{\mathrm{T}}\rangle && \%\boldsymbol{\mathcal{F}_3}\\
&- \sum_{m=1}^{M}\left(\sum_{m'=m+1}^{M}\left(\sum_{q=1}^{Q}\frac{\langle s_{qm}\widetilde{w}_{qm}\rangle\langle s_{qm'}\widetilde{w}_{qm'}\rangle}{\sigma_q^2}\right)\langle\boldsymbol{\phi}_{m'}\rangle\right)^{\mathrm{T}}\langle\boldsymbol{\phi}_m\rangle && \%\boldsymbol{\mathcal{F}_4}\\
&- \frac{MQ}{2}\log(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2}\sum_{q=1}^{Q}\sum_{m=1}^{M}\langle\widetilde{w}_{qm}^2\rangle && \%\boldsymbol{\mathcal{F}_5}\\
&+ \log(\pi)\sum_{q=1}^{Q}\sum_{m=1}^{M}\langle s_{qm}\rangle + \log(1-\pi)\sum_{q=1}^{Q}\sum_{m=1}^{M}\langle 1-s_{qm}\rangle && \%\boldsymbol{\mathcal{F}_6}\\
&- \frac{MN}{2}\log(2\pi) - \frac{1}{2}\sum_{m=1}^{M}\left(\log|\mathbf{K}_m| + \mathrm{tr}[\mathbf{K}_m^{-1}\langle\boldsymbol{\phi}_m\boldsymbol{\phi}_m^{\mathrm{T}}\rangle]\right) && \%\boldsymbol{\mathcal{F}_7}\\
&+ \frac{MQ}{2}\log(2e\pi\sigma_w^2) - \frac{1}{2}\log\sigma_w^2\sum_{q=1}^{Q}\sum_{m=1}^{M}\langle s_{qm}\rangle + \frac{1}{2}\sum_{q=1}^{Q}\sum_{m=1}^{M}\langle s_{qm}\rangle\log\sigma_{w_{qm}}^2 && \%\boldsymbol{\mathcal{E}_1}\\
&- \sum_{q=1}^{Q}\sum_{m=1}^{M}\left[\langle 1-s_{qm}\rangle\log\langle 1-s_{qm}\rangle - \langle s_{qm}\rangle\log\langle s_{qm}\rangle\right] && \%\boldsymbol{\mathcal{E}_2}\\
&+ \frac{MN}{2}\log(2\pi) + \frac{MN}{2} + \frac{1}{2}\sum_{m=1}^{M}\log|\boldsymbol{\Sigma}_{\phi_m}| && \%\boldsymbol{\mathcal{E}_3}
\end{aligned} \tag{2}
$$

The $\mathcal{F}_5$ term can be further simplified by using the fact that $\langle\widetilde{w}_{qm}^2\rangle = \gamma_{qm}(\mu_{w_{qm}}^2 + \sigma_{w_{qm}}^2) + (1-\gamma_{qm})\sigma_w^2$. Also some terms above cancel out such as the term $\frac{MQ}{2}\log(2\pi\sigma_w^2)$.

Finally, the updates for the hyperparameters are as follows

$$\sigma_q^2 = \frac{1}{N}\mathrm{tr}[\mathbf{y}_q\mathbf{y}_q^{\mathrm{T}} - \mathbf{y}_q\sum_{m=1}^{M}\langle s_{qm}\widetilde{w}_{qm}\rangle\langle\boldsymbol{\phi}_m\rangle^{\mathrm{T}} + \sum_{m=1}^{M}\langle s_{qm}\widetilde{w}_{qm}^2\rangle\langle\boldsymbol{\phi}_m\boldsymbol{\phi}_m^{\mathrm{T}}\rangle + 2\sum_{m>m'}\langle s_{qm}\widetilde{w}_{qm}\rangle\langle s_{qm'}\widetilde{w}_{qm'}\rangle\langle\boldsymbol{\phi}_m\rangle\langle\boldsymbol{\phi}_{m'}\rangle]$$

$$\sigma_w^2 = \frac{\sum_{q=1}^{Q}\sum_{m=1}^{M}\gamma_{qm}(\mu_{w_{qm}}^2 + \sigma_{w_{qm}}^2)}{\sum_{q=1}^{Q}\sum_{m=1}^{M}\gamma_{qm}}$$

$$\pi = \frac{1}{MQ}\sum_{q=1}^{Q}\sum_{m=1}^{M}\langle s_{qm}\rangle$$

$$\boldsymbol{\theta}_m = \arg\max_{\boldsymbol{\theta}_m}\left[-\frac{1}{2}\log|\mathbf{K}_m| - \frac{1}{2}\mathrm{tr}[\mathbf{K}_m^{-1}\langle\boldsymbol{\phi}_m\boldsymbol{\phi}_m^{\mathrm{T}}\rangle]\right]$$

where anything in brackets $\langle\cdot\rangle$ is computed under the current value of the variational distribution and is assumed to be fixed (given from the E-step).

1.3 A joint update for $q(\boldsymbol{\phi}_m)$ and $\boldsymbol{\theta}_m$

Notice that the update for the hyperparameter $\boldsymbol{\theta}_m$, which parameterize $\mathbf{K}_m$, is problematic for two reasons. Firstly, it requires the inverse of $\mathbf{K}_m$ and this is numerically unstable as in (computer precision) such an inverse might not exist. Of course, such a problem can be partially overcome by adding a small amount of "jitter" into the diagonal of $\mathbf{K}_m$, which however is not ideal. Secondly, the update of the hyperparameters $\boldsymbol{\theta}_m$ strongly depends on the statistic $\langle \boldsymbol{\phi}_m \boldsymbol{\phi}_m^{\mathrm{T}} \rangle$ computed under the factor $q(\boldsymbol{\phi}_m)$ which is fixed. The update of $\boldsymbol{\theta}_m$ can be "slow" because $\langle \boldsymbol{\phi}_m \boldsymbol{\phi}_m^{\mathrm{T}} \rangle$ depends on the kernel matrix $\mathbf{K}_m^{old}$ evaluated at the old values of the hyperparameter $\boldsymbol{\theta}_m^{old}$. To resolve this, we would like to update simultaneously somehow $\boldsymbol{\theta}_m$ and the statistic $\langle \boldsymbol{\phi}_m \boldsymbol{\phi}_m^{\mathrm{T}} \rangle$, i.e. the factor $q(\boldsymbol{\phi}_m)$. This can be done in an elegant and efficient way using a Marginalized Variational step [2]. Next we describe the whole idea.

We would like to perform a joint optimization update for $(q(\boldsymbol{\phi}_m), \boldsymbol{\theta}_m)$ in a way that the factor $q(\boldsymbol{\phi}_m)$ is marginalized/removed optimally from the optimization problem. We write the variational lower bound as follows

$$\mathcal{F}(\boldsymbol{\theta}_m) = \int q(\boldsymbol{\phi}_m) q(\Theta) \log \frac{p(\mathbf{Y}, \boldsymbol{\phi}_m, \Theta) p(\Theta) \mathcal{N}(\boldsymbol{\phi}_m | \mathbf{0}, \mathbf{K}_m)}{q(\boldsymbol{\phi}_m) q(\Theta)} \, d\boldsymbol{\phi}_m \, d\Theta,$$

where Θ are all random variables excluding $\boldsymbol{\phi}_m$ and $q(\Theta)$ their variational distribution. Given that we wish to update the factor $q(\boldsymbol{\phi}_m)$ and the kernel matrix $\mathbf{K}_m$ while the rest are just constants, the above is written as

$$\mathcal{F}(\boldsymbol{\theta}_m) = \int q(\boldsymbol{\phi}_m) q(\Theta) \log \frac{p(\mathbf{Y}, \boldsymbol{\phi}_m, \Theta) \mathcal{N}(\boldsymbol{\phi}_m | \mathbf{0}, \mathbf{K}_m)}{q(\boldsymbol{\phi}_m)} \, d\boldsymbol{\phi}_m \, d\Theta + const.$$

Now the optimal $q(\boldsymbol{\phi}_m)$ is

$$q(\boldsymbol{\phi}_m) = \frac{\exp^{\langle \log p(\mathbf{Y}, \boldsymbol{\phi}_m, \Theta) \rangle_{q(\Theta)}} \mathcal{N}(\boldsymbol{\phi}_m | \mathbf{0}, \mathbf{K}_m)}{\int \exp^{\langle \log p(\mathbf{Y}, \boldsymbol{\phi}_m, \Theta) \rangle_{q(\Theta)}} \mathcal{N}(\boldsymbol{\phi}_m | \mathbf{0}, \mathbf{K}_m) \, d\boldsymbol{\phi}_m}$$

Substituting this optimal $q(\boldsymbol{\phi}_m)$ back into the bound we obtain

$$\mathcal{F}(\boldsymbol{\theta}_m) = \log \int \exp^{\langle \log p(\mathbf{Y}, \boldsymbol{\phi}_m, \Theta) \rangle_{q(\Theta)}} \mathcal{N}(\boldsymbol{\phi}_m | \mathbf{0}, \mathbf{K}_m) \, d\boldsymbol{\phi}_m + const.$$

This now is analytically tractable and can neatly be written as the marginal likelihood of a standard GP regression model:

$$\mathcal{F}(\boldsymbol{\theta}_m) = \log \mathcal{N}(\bar{\mathbf{y}} | \mathbf{0}, \mathbf{K}_m + \alpha_m^{-1} I) + const$$

where

$$\bar{\mathbf{y}} = \frac{1}{\alpha_m} \sum_{q=1}^{Q} \frac{\langle s_{qm} \widetilde{w}_{qm} \rangle}{\sigma_q^2} \left(\mathbf{y}_q - \sum_{k \neq m} \langle s_{qk} \widetilde{w}_{qk} \rangle \langle \boldsymbol{\phi}_k \rangle \right)$$

are like fixed pseudo-data and

$$\alpha_m = \sum_{q=1}^{Q} \frac{\langle s_{qm} \widetilde{w}_{qm}^2 \rangle}{\sigma_q^2}$$

is a fixed inverse noise variance parameter. The above now is optimized wrt $\boldsymbol{\theta}_m$ and this can be done by using any standard GP implementation for maximizing the marginal likelihood of a GP standard regression model (we will only need to keep fixed the noise variance α_m^{-1}). Notice that the optimization requires the inverse of $\mathbf{K}_m + \alpha_m^{-1} I$ which often will be numerically stable due to the addition of α_m^{-1} in the diagonal of $\mathbf{K}_m$.

Once the optimization is completed, we evaluate the final value of the factor $q(\boldsymbol{\phi}_m)$ and then continue with other variational EM updates.

2 Paired Gibbs sampling for spike and slab linear regression

Consider a single-output regression model:

$$p(\mathbf{y}, \widetilde{\mathbf{w}}, \mathbf{s}) = \mathcal{N}(\mathbf{y} | \sum_{m=1}^{M} s_m \widetilde{w}_m \mathbf{x}_m, \sigma^2 \mathbf{I}) \prod_{m=1}^{M} \left[\mathcal{N}(\widetilde{w}_m | 0, \sigma_w^2) [\pi^{s_m} (1 - \pi)^{1 - s_m}] \right]$$

The paired Gibbs sampler iteratively samples from the following conditional

$$p(\widetilde{w}_m, s_m|\widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y}) = p(\widetilde{w}_m|s_m, \widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y})p(s_m|\widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y}).$$

$p(s_m = 1|\widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y})$ is obtained analytically to be

$$\begin{aligned} p(s_m = 1|\widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y}) &= \frac{\pi\mathcal{N}(\mathbf{y}|\sum_{k\neq m} s_k\widetilde{w}_k\mathbf{x}_k, \sigma^2\mathbf{I} + \sigma_w^2\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}})}{\pi\mathcal{N}(\mathbf{y}|\sum_{k\neq m} s_k\widetilde{w}_k\mathbf{x}_k, \sigma^2\mathbf{I} + \sigma_w^2\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}) + (1-\pi)\mathcal{N}(\mathbf{y}|\sum_{k\neq m} s_k\widetilde{w}_k\mathbf{x}_k, \sigma^2\mathbf{I})} \\ &= \frac{\pi\mathcal{N}(\mathbf{y}|\mathbf{b}_m, \sigma^2\mathbf{I} + \sigma_w^2\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}})}{\pi\mathcal{N}(\mathbf{y}|\mathbf{b}_m, \sigma^2\mathbf{I} + \sigma_w^2\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}}) + (1-\pi)\mathcal{N}(\mathbf{y}|\mathbf{b}_m, \sigma^2\mathbf{I})} \end{aligned}$$

where

$$\mathbf{b}_m = \sum_{k\neq m} s_k\widetilde{w}_k\mathbf{x}_k$$

A computationally more efficient expression can be obtained by applying matrix inversion lemma:

$$p(s_m = 1|\widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y}) = \sigma(u_m)$$

where $\sigma(u_m) = \frac{1}{1+e^{-u_m}}$ and

$$\begin{aligned} u_m &= \log\frac{\pi}{1-\pi} + \frac{1}{2}\log\frac{\sigma^2}{\sigma_w^2} - \frac{1}{2}\log\left(\mathbf{x}_m^{\mathrm{T}}\mathbf{x}_m + \frac{\sigma^2}{\sigma_w^2}\right) + \frac{1}{2\sigma^2}\frac{\left(\mathbf{x}_m^{\mathrm{T}}\mathbf{y} - \sum_{k\neq m} s_k\widetilde{w}_k\mathbf{x}_m^{\mathrm{T}}\mathbf{x}_k\right)^2}{\left(\mathbf{x}_m^{\mathrm{T}}\mathbf{x}_m + \frac{\sigma^2}{\sigma_w^2}\right)} \\ &= \log\frac{\pi}{1-\pi} + \frac{1}{2}\log\frac{\sigma^2}{\sigma_w^2} - \frac{1}{2}\log\left(\mathbf{x}_m^{\mathrm{T}}\mathbf{x}_m + \frac{\sigma^2}{\sigma_w^2}\right) + \frac{1}{2\sigma^2}\frac{\left(\mathbf{x}_m^{\mathrm{T}}\mathbf{y} - \mathbf{x}_m^{\mathrm{T}}\mathbf{b}_k\right)^2}{\left(\mathbf{x}_m^{\mathrm{T}}\mathbf{x}_m + \frac{\sigma^2}{\sigma_w^2}\right)} \end{aligned}$$

Also, $p(\widetilde{w}_m|s_m = 0, \widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y}) = \mathcal{N}(\widetilde{w}_m|0, \sigma_w^2)$ and $p(\widetilde{w}_m|s_m = 1, \widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y})$ is

$$p(\widetilde{w}_m|s_m = 1, \widetilde{w}_{\backslash m}, s_{\backslash m}, \mathbf{y}) = \mathcal{N}\left(\widetilde{w}_m \left| \frac{\mathbf{x}_m^{\mathrm{T}}\mathbf{y} - \sum_{k\neq m} s_k\widetilde{w}_k\mathbf{x}_m^{\mathrm{T}}\mathbf{x}_k}{\mathbf{x}_m^{\mathrm{T}}\mathbf{x}_m + \frac{\sigma^2}{\sigma_w^2}}, \frac{\sigma^2}{\mathbf{x}_m^{\mathrm{T}}\mathbf{x}_m + \frac{\sigma^2}{\sigma_w^2}}\right.\right)$$

References

[1] Iain Murray and Ryan Prescott Adams. Slice sampling covariance hyperparameters of latent Gaussian models. In J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1723–1731. 2010.

[2] M. Lázaro-Gredilla and M. Titsias. Variational heteroscedastic Gaussian process regression. In *28th International Conference on Machine Learning (ICML-11)*, pages 841–848, New York, NY, USA, June 2011. ACM.