S1 Variational Inference: Regression Model

The variational approach introduced by Titsias [2009], considers an augmented probability model for the marginal likelihood, based on a set of inducing inputs $\mathbf{Z}$. The exact marginal likelihood is then computed by

$$p(\mathbf{y}|\mathbf{X}) = \iint p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u},\mathbf{X},\mathbf{Z})p(\mathbf{u}|\mathbf{Z})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{f}, \quad \text{(S.1)}$$

where $\mathbf{u}$ is the Gaussian process evaluated at $\mathbf{Z}$. The inducing inputs are turned into variational parameters by introducing the variational distribution

$$q(\mathbf{f},\mathbf{u}|\mathbf{X},\mathbf{Z}) = p(\mathbf{f}|\mathbf{u},\mathbf{X},\mathbf{Z})\phi(\mathbf{u}). \quad \text{(S.2)}$$

The variational terms $\mathbf{Z}$ and $\phi(\cdot)$ are then optimized by minimizing the the Kullback-Leibler divergence $\mathrm{KL}\left(q(\mathbf{f},\mathbf{u}|\mathbf{X},\mathbf{Z}) \,\|\, p(\mathbf{f},\mathbf{u}|\mathbf{y},\mathbf{X},\mathbf{Z})\right)$. We refer the reader to Titsias [2009] for a detailed explanation.

S2 Variational Inference: Joint Model

The joint probability model for the GP-LVM is given by

$$p(\mathbf{Y},\mathbf{X}) = p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}), \quad \text{(S.3)}$$

where the dimensions of $\mathbf{Y}$ are considered independent conditioned on the features. Hence, $p(\mathbf{Y}|\mathbf{X})$ can be factorized as

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{j=1}^{p} p(\mathbf{y}_j|\mathbf{X}), \quad \text{(S.4)}$$

where $\{\mathbf{y}_j\}_{j=1}^{p}$ represent the columns of $\mathbf{Y}$. Notice that $\mathbf{X}$ is non-linear inside $p(\mathbf{y}_j|\mathbf{X})$. The exact marginal likelihood is given by

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}. \quad \text{(S.5)}$$

Thus, computing the marginal distribution of the joint model involves computing the expected value of (S.1) under the distribution $p(\mathbf{X})$. This does not allow for the variational distribution $\phi(\cdot)$ to be optimized in the same way as in the regression case. For this reason, Titsias and Lawrence [2010] introduced a factorized variational distribution

$$q(\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_i,\mathbf{S}_i), \quad \text{(S.6)}$$

where $\{\mathbf{S}_i\}_{i=1}^{n}$ are defined as diagonal matrices. A lower bound on $\log p(\mathbf{Y})$ can now be defined using (S.6) and $\phi(\cdot)$ can determined through a mean field approach. We refer the reader to Titsias and Lawrence [2010] for a detailed explanation.

S3 EP-DTC Derivation

Consider a Gaussian process $\mathbf{f}$ with covariance $\mathbf{K_{ff}}$. In the regression case with Gaussian likelihoods, the probabilistic variational sparse GP approximation to the marginal likelihood, proposed by Titsias [2009], is formulated as a lower bound. This guarantees the parameters learning to be more rigorous. The lower bound is defined as the DTC approximation to the marginal likelihood, but corrected with a trace term, which guarantees consistency when modelling the training and test sets. We are interested in extending the variational sparse GP approximation for the case of non-Gaussian likelihoods. As a first step, we propose a derivation of the EP algorithm based on the DTC approximation.

Let the dependence of $\mathbf{f}$ on the inducing inputs $\mathbf{u}$ be defined deterministically as in the DTC approximation

$$q(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u},\mathbf{0}), \quad \text{(S.7)}$$

where $\mathbf{K_{fu}}$ is the covariance function computed across the training data and the inducing inputs, and $\mathbf{K_{uu}}$ is the covariance function computed between the inducing inputs. Then, in can be shown that the marginal distribution of $\mathbf{f}$ is given by

$$q(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{Q_{ff}}), \quad \text{(S.8)}$$

where $\mathbf{Q_{ff}} = \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}}$.

Let the EP site approximations $\{t_i(f_i) \approx p(y_i|f_i)\}_{i=1}^{n}$ be un-normalized Gaussians with moment parameters $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$. The overall likelihood approximation is then be given by

$$p(\mathbf{y}|\mathbf{f}) \approx \tilde{\mathbf{Z}} \times \mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}), \quad \text{(S.9)}$$

for some constant $\tilde{\mathbf{Z}}$. Notice that $p(\mathbf{y}|\mathbf{f})$ normalizes over $\mathbf{y}$, whilst the Gaussian distribution in the r.h.s. of (S.9) normalizes over $\mathbf{f}$.

The combination of the prior distribution in (S.8) with the likelihood in (S.9) yields a posterior distribution of $\mathbf{f}$ with parameters

$$\boldsymbol{\Sigma} = \left(\mathbf{Q_{ff}^{-1}} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1}, \quad \text{(S.10)}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}\right). \quad \text{(S.11)}$$

The gain of the sparse approximation depends on formulating the computation of the posterior parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, in an efficient way. In this case, this is achievable at a computational complexity of $\mathcal{O}(m^2 n)$ and with storage demands of $\mathcal{O}(mn)$. By applying the

matrix inversion lemma, the posterior variance can be computed as

$$\begin{aligned}\boldsymbol{\Sigma} &= \left(\left(\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}}\right)^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1}\right)^{-1} \\ &= \tilde{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}}\left(\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}} + \tilde{\boldsymbol{\Sigma}}\right)^{-1}\tilde{\boldsymbol{\Sigma}} \\ &= \mathbf{K_{fu}}\left(\mathbf{K_{uu}} + \mathbf{K_{uf}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K_{fu}}\right)^{-1}\mathbf{K_{uf}} \\ &= \mathbf{K_{fu}}(\mathbf{L}\mathbf{L}^\top)^{-1}\mathbf{K_{uf}},\end{aligned} \tag{S.12}$$

where $\mathbf{L}$ is the Cholesky decomposition of $(\mathbf{K_{uu}} + \mathbf{K_{uf}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K_{fu}})$.

EP is an iterative algorithm in which the site approximations are updated one at a time, until convergence is achieved (see Williams and Rasmussen [2006] for a detailed explanation). In our sparse formulation, the procedure for updating the parameters of the site approximations remains the same as in standard EP. What changes is the computation of the posterior parameters, which now depends on factorization of the covariance matrix given in (S.12). We will explain these updates based on the natural parameters $\{\tilde{\tau}_i\}_{i=1}^n$ and $\{\tilde{\nu}_i\}_{i=1}^n$, rather than the moment paramenters $\{\tilde{\mu}_i\}_{i=1}^n$ and $\{\tilde{\sigma}_i^2\}_{i=1}^n$, as this simplifies the notation. Suppose that, after updating the i-th site approximation, we change its natural parameters by $\Delta\tilde{\tau}_i$ and $\Delta\tilde{\nu}_i$. Let

$$\mathbf{E} = \tilde{\boldsymbol{\Sigma}}^{-1} + \Delta\tilde{\tau}_i \mathbf{e}_i \mathbf{e}_i^\top, \tag{S.13}$$

$$\mathbf{E}^{-1} = \tilde{\boldsymbol{\Sigma}} - \frac{\tilde{\tau}_i^2 \Delta\tilde{\tau}_i}{1 + \tilde{\tau}_i \Delta\tilde{\tau}_i}\mathbf{e}_i\mathbf{e}_i^\top, \tag{S.14}$$

where $\mathbf{e}_i$ is the i-th canonical basis vector of $\mathbb{R}^n$. Then, the updates of the posterior variance can be computed as

$$\begin{aligned}\boldsymbol{\Sigma}^{\text{new}} &= \left(\left(\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}}\right)^{-1} + \mathbf{E}\right)^{-1} \\ &= \mathbf{K_{fu}}\left(\mathbf{K_{uu}} + \mathbf{K_{uf}}\mathbf{E}\mathbf{K_{fu}}\right)^{-1}\mathbf{K_{uf}} \\ &= \mathbf{K_{fu}}(\mathbf{L}\mathbf{L}^\top + \mathbf{k}_i\Delta\tilde{\tau}_i\mathbf{k}_i^\top\mathbf{K_{fu}})^{-1}\mathbf{K_{uf}} \\ &= \mathbf{K_{fu}}(\mathbf{L}^{\text{new}}\mathbf{L}^{\text{new}\top})^{-1}\mathbf{K_{uf}},\end{aligned} \tag{S.15}$$

where $\mathbf{k}_i$ is the i-th column of $\mathbf{K_{uf}}$ and $\mathbf{L}^{\text{new}}$ is the Cholesky decomposition of $(\mathbf{L}\mathbf{L}^\top + \mathbf{k}_i\Delta\tilde{\tau}_i\mathbf{k}_i^\top\mathbf{K_{fu}})$.

Finally, the update of $\boldsymbol{\mu}$ can be computed as

$$\begin{aligned}\boldsymbol{\mu}^{\text{new}} &= \boldsymbol{\Sigma}^{\text{new}}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \Delta\tilde{\nu}_i\mathbf{e}_i\right) \\ &= \boldsymbol{\Sigma}^{\text{new}}\left(\left(\boldsymbol{\Sigma}^{\text{new}-1} - \Delta\tilde{\tau}_i\mathbf{e}_i\mathbf{e}_i^\top\right)\boldsymbol{\mu} + \Delta\tilde{\nu}_i\right) \\ &= \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\text{new}}\left(\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\mu_i\right)\mathbf{e}_i \\ &= \boldsymbol{\mu} + \left(\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\mu_i\right)\mathbf{s}_i^{\text{new}},\end{aligned} \tag{S.16}$$

where $\mathbf{s}_i^{\text{new}}$ is the i-th column of $\boldsymbol{\Sigma}^{\text{new}}$.

S4 Variational inference and EP-DTC

Assume we already have an optimal EP-DTC approximation of the form of (S.9). Following Titsias [2009] in using Jensen's inequality to define a lower bound on the logarithm of (S.1), we see that

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &= \log \iint p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u},\mathbf{X},\mathbf{Z}) \\ &\quad \times p(\mathbf{u}|\mathbf{Z})\frac{\phi(\mathbf{u})}{\phi(\mathbf{u})}\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u} \\ &\geq \iint p(\mathbf{f}|\mathbf{u},\mathbf{X},\mathbf{Z})\phi(\mathbf{u}) \\ &\quad \times \log\frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{u}|\mathbf{Z})}{\phi(\mathbf{u})}\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u}.\end{aligned} \tag{S.17}$$

By replacing $p(\mathbf{y}|\mathbf{f})$ in (S.17) with the EP-DTC approximation, we obtain

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}) &\gtrsim \iint p(\mathbf{f}|\mathbf{u},\mathbf{X},\mathbf{Z})\phi(\mathbf{u}) \\ &\quad \times \log\frac{\tilde{\mathbf{Z}}\mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})p(\mathbf{u}|\mathbf{Z})}{\phi(\mathbf{u})}\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u} \\ &\gtrsim \int \phi(\mathbf{u})\left(H + \log\frac{\tilde{\mathbf{Z}}p(\mathbf{u}|\mathbf{Z})}{\phi(\mathbf{u})}\right)\mathrm{d}\mathbf{u},\end{aligned} \tag{S.18}$$

where

$$H = \int p(\mathbf{f}|\mathbf{u},\mathbf{X},\mathbf{Z})\log\mathcal{N}(\mathbf{f}|\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})\mathrm{d}\mathbf{f}. \tag{S.19}$$

Let $\boldsymbol{\alpha} = \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}$, we can re-express H as

$$\begin{aligned}H &= -\frac{n}{2}\log 2\pi - \frac{1}{2}|\tilde{\boldsymbol{\Sigma}}| \\ &\quad - \int p(\mathbf{f}|\mathbf{u},\mathbf{X},\mathbf{Z})(\mathbf{f}-\tilde{\boldsymbol{\mu}})^\top\tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{f}-\tilde{\boldsymbol{\mu}})\mathrm{d}\mathbf{f} \\ &= -\frac{N}{2}\log 2\pi - \frac{1}{2}|\tilde{\boldsymbol{\Sigma}}| \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - 2\tilde{\boldsymbol{\mu}}\boldsymbol{\alpha}^\top + \tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}^\top\right)\tilde{\boldsymbol{\Sigma}}^{-1}\right) \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\left(\mathbf{K_{ff}} - \mathbf{Q_{ff}}\right)\tilde{\boldsymbol{\Sigma}}^{-1}\right) \\ &= \log\mathcal{N}(\tilde{\boldsymbol{\mu}}|\boldsymbol{\alpha},\tilde{\boldsymbol{\Sigma}}) - \frac{1}{2}\mathrm{tr}\left(\left(\mathbf{K_{ff}} - \mathbf{Q_{ff}}\right)\tilde{\boldsymbol{\Sigma}}^{-1}\right).\end{aligned} \tag{S.20}$$

Using (S.20) in (S.18) and reversing Jensen's inequality in (S.17) leads to the definition of the lower bound on

$\log p(\mathbf{y}|\mathbf{X})$

$$\begin{aligned}\mathcal{L}_E &= \log \int \mathcal{N}(\tilde{\boldsymbol{\mu}}|\boldsymbol{\alpha}, \tilde{\boldsymbol{\Sigma}}) p(\mathbf{u}|\mathbf{X}) \mathrm{d}\mathbf{u} \\ &\quad - \frac{1}{2}\mathrm{tr}\left((\mathbf{K_{ff}} - \mathbf{Q_{ff}})\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \tilde{\mathbf{Z}} \\ &= \log \mathcal{N}(\tilde{\boldsymbol{\mu}}|\mathbf{0}, \mathbf{Q_{ff}} + \tilde{\boldsymbol{\Sigma}}) \\ &\quad - \frac{1}{2}\mathrm{tr}\left((\mathbf{K_{ff}} - \mathbf{Q_{ff}})\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \tilde{\mathbf{Z}} \\ &\lesssim \log p(\mathbf{y}|\mathbf{X}).\end{aligned} \tag{S.21}$$

S5 Sparse EP For Uncertain Inputs

Consider a posterior covariance and a posterior mean given by

$$\boldsymbol{\Sigma} = \left(\tilde{\boldsymbol{\Sigma}}^{-1} + (\hat{\boldsymbol{\Psi}}^\top \mathbf{R}^\top \mathbf{R}\hat{\boldsymbol{\Psi}} + \hat{\boldsymbol{\Lambda}})^{-1}\right)^{-1}, \tag{S.22}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}\right), \tag{S.23}$$

where $\hat{\boldsymbol{\Psi}} \in \mathbb{R}^{m\times n}$, $\hat{\boldsymbol{\Lambda}} \in \mathbb{R}^{n\times n}$ is a diagonal matrix, and $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ are the mean and covariance of an EP likelihood approximation. By applying the matrix inversion lemma, $\boldsymbol{\Sigma}$ can be re-expressed as follows:

$$\begin{aligned}\boldsymbol{\Sigma} = \Big(&\tilde{\boldsymbol{\Sigma}}^{-1} + \hat{\boldsymbol{\Lambda}}^{-1} \\ &-\hat{\boldsymbol{\Lambda}}^{-1}\hat{\boldsymbol{\Psi}}\mathbf{R}^\top(\mathbf{R}\hat{\boldsymbol{\Psi}}^\top\hat{\boldsymbol{\Lambda}}^{-1}\hat{\boldsymbol{\Psi}}\mathbf{R}^\top + \mathbf{I})^{-1}\mathbf{R}\hat{\boldsymbol{\Psi}}^\top\hat{\boldsymbol{\Lambda}}^{-1}\Big)^{-1}.\end{aligned} \tag{S.24}$$

After applying a second time the matrix inversion lemma, to get rid of the negative exponent in (S.24), we get that

$$\begin{aligned}\boldsymbol{\Sigma} =& \hat{\boldsymbol{\Psi}}^{\text{new}}(\mathbf{R}^{\text{new}})^\top\mathbf{R}^{\text{new}}(\hat{\boldsymbol{\Psi}}^{\text{new}})^\top \\ &+ (\hat{\boldsymbol{\Lambda}}^{\text{new}})^\top,\end{aligned} \tag{S.25}$$

for some suitable $\hat{\boldsymbol{\Psi}}^{\text{new}}$, $\mathbf{R}^{\text{new}}$ and $\hat{\boldsymbol{\Lambda}}^{\text{new}}$.

As in the case of the EP-DTC formulation, the updates in this new setting will depend on the covariance factorization, given by (S.25) in this case. Suppose that, after updating the i-th site approximation, we change its natural parameters by $\Delta\tilde{\tau}_i$ and $\Delta\tilde{\nu}_i$. An efficient way of defining the updates according to these changes is given by the following equations:

$$\hat{\boldsymbol{\Lambda}}^{\text{new}} = \hat{\boldsymbol{\Lambda}} - \frac{\Delta\tilde{\tau}_i\hat{\lambda}_{ii}^2}{1+\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i\mathbf{e}_i^\intercal, \tag{S.26}$$

$$\hat{\boldsymbol{\Psi}}^{\text{new}} = \hat{\boldsymbol{\Psi}} - \frac{\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}{1+\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i\hat{\boldsymbol{\psi}}_i, \tag{S.27}$$

$$\delta_i = \frac{\Delta\tilde{\tau}_i}{1+\Delta\tilde{\tau}_i s_{ii}}, \tag{S.28}$$

$$\mathbf{R}^{\text{new}} = \text{Cholesky}\left(\mathbf{R}^\intercal\left(\mathbf{I} - \mathbf{R}\hat{\boldsymbol{\psi}}_i\delta_i\hat{\boldsymbol{\psi}}_i^\intercal\mathbf{R}^\intercal\right)\mathbf{R}\right). \tag{S.29}$$

Let $\boldsymbol{\mu}$ be re-expressed as

$$\boldsymbol{\mu} = \boldsymbol{\omega} + \hat{\boldsymbol{\Psi}}\boldsymbol{\gamma}, \tag{S.30}$$

for some $\boldsymbol{\omega} \in \mathbb{R}^n$ and $\boldsymbol{\gamma} \in \mathbb{R}^m$. Then, the corresponding updates are given by

$$\boldsymbol{\omega}^{\text{new}} = \boldsymbol{\omega} + \frac{(\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\omega_i)\hat{\lambda}_{ii}}{1+\Delta\tilde{\tau}_i\hat{\lambda}_{ii}}\mathbf{e}_i, \tag{S.31}$$

$$\begin{aligned}\boldsymbol{\gamma}^{\text{new}} =& \hat{\boldsymbol{\Psi}}^{\text{new}}\boldsymbol{\gamma} \\ &+\hat{\boldsymbol{\Psi}}^{\text{new}}\left((\Delta\tilde{\nu}_i - \Delta\tilde{\tau}_i\tilde{\mu}_i)\mathbf{R}^{\text{new}\intercal}\mathbf{R}^{\text{new}}\hat{\boldsymbol{\psi}}_i^{\text{new}}\right).\end{aligned} \tag{S.32}$$

S6 Uncertain Inputs with EP

Following Titsias and Lawrence [2010] and putting together (S.5) and (S.6), we get the lower bound

$$\log p(\mathbf{Y}) \geq \sum_{j=1}^{q}\langle p(\mathbf{y}_j|\mathbf{X})\rangle_{q(\mathbf{X})} - \mathrm{KL}\left(q(\mathbf{X})\,\|\,p(\mathbf{X})\right). \tag{S.33}$$

If we consider an EP approximation, as in (S.9), and follow a similar approach as the one from (S.17) to (S.21), we get that

$$\begin{aligned}\langle p(\mathbf{y}_j|\mathbf{X})\rangle_{q(\mathbf{X})} \gtrsim& \log\langle\exp(\mathcal{N}(\tilde{\boldsymbol{\mu}}_j|\boldsymbol{\alpha}_j, \tilde{\boldsymbol{\Sigma}}_j))\rangle_{p(\mathbf{u}_j|\mathbf{X})} \\ &- \frac{1}{2}\mathrm{tr}\left(\langle\mathbf{K_{ff}} - \mathbf{Q_{ff}}\rangle_{q(\mathbf{X})}\tilde{\boldsymbol{\Sigma}}^{-1}\right) + \tilde{\mathbf{Z}}.\end{aligned} \tag{S.34}$$

We can now compute a lower bound on the log-marginal likelihood as

$$\begin{aligned}\mathcal{L}_H &= \log\mathcal{N}\left(\tilde{\boldsymbol{\mu}}|0, \boldsymbol{\Psi}_1^\top\mathbf{K_{uu}}^{-1}\boldsymbol{\Psi}_1 + \boldsymbol{\Lambda} + \tilde{\boldsymbol{\Sigma}}\right) - \tilde{\psi}_0 \\ &\quad + \mathrm{tr}\left(\mathbf{K_{uu}}^{-1}\tilde{\boldsymbol{\Psi}}_2\right) + \mathrm{KL}\left(q(\mathbf{X})\,\|\,p(\mathbf{X})\right) + \tilde{\mathbf{Z}} \\ &\lesssim \log p(\mathbf{Y}),\end{aligned} \tag{S.35}$$

where $\tilde{\psi}_0 = \mathrm{tr}\left(\tilde{\boldsymbol{\Sigma}}^{-1}\langle\mathbf{K_{ff}}\rangle_{q(\mathbf{X})}\right)$, $\boldsymbol{\Psi}_1 = \langle\mathbf{K_{uf}}\rangle_{q(\mathbf{X})}$, $\tilde{\boldsymbol{\Psi}}_2 = \langle\mathbf{K_{uf}}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{K_{fu}}\rangle_{q(\mathbf{X})}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix such that $\boldsymbol{\Lambda}_{ii} = \mathrm{tr}\left(\tilde{\boldsymbol{\Psi}}_{2(i)}\mathbf{K_{uu}}^{-1}\right) - \boldsymbol{\Psi}_{1(i)}^\top\mathbf{K_{uu}}^{-1}\boldsymbol{\Psi}_{1(i)}$. The sub-index (i) means that we are only taking the i-th column of the corresponding matrix.