Supplementary Material
Efficient Variational Inference for Gaussian Process Regression Networks

Trung V. Nguyen and Edwin V. Bonilla

This supplementary material concerns the derivation of two variational inference methods for GPRN. Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}_{i=1}^N$ be the set of training inputs and $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}_{i=1}^N$ be the set of corresponding training targets. We use P to denote the number of outputs and Q as the number of latent functions in a GPRN model.

As in the original paper, let us denote $\mathbf{u} = [\mathbf{f}; \mathbf{w}]$ as the concatenation of the latent functions and weights evaluated at the training points. Here $\mathbf{f} = vec(\mathbf{F}) = F(:) = [\mathbf{f}_1; \ldots; \mathbf{f}_Q]$ where $\mathbf{f}_j = [f_j(x_1), f_j(x_2), \ldots f_j(x_N)]^T$ is the j^{th} latent function. Hence $\mathbf{F}$ is a $N \times Q$ matrix where each column is a latent function and each row is the values of all latent functions at a particular input. Similarly $\mathbf{w} = vec(\mathbf{W}) = \mathbf{W}(:)$ where

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \ldots & w_{1Q} \\ w_{21} & w_{22} & \ldots & w_{2Q} \\ \vdots & \vdots & \vdots & \vdots \\ w_{P1} & w_{P2} & \ldots & w_{PQ} \end{bmatrix}$$

where each w_{ij} is a weight function, i.e., $w_{ij} = [w_{ij}(x_1), w_{ij}(x_2), \ldots, w_{ij}(x_N)]^T$. For mathematical and notational convenience, also define

$$\mathbf{W}(x_n) = \begin{bmatrix} w_{11}(x_n) & w_{12}(x_n) & \ldots & w_{1Q}(x_n) \\ w_{21}(x_n) & w_{22}(x_n) & \ldots & w_{2Q}(x_n) \\ \vdots & \vdots & \vdots & \vdots \\ w_{P1}(x_n) & w_{P2}(x_n) & \ldots & w_{PQ}(x_n) \end{bmatrix}.$$

1 Variational Inference for GPRNs

In variational inference for GPRN models, our goal is to find the distribution $q(\mathbf{u})$ closest to the posterior $p(\mathbf{u}|\mathcal{D},)$ with respect to the Kullback-Leibler divergence,

$$\mathrm{KL}[q(\mathbf{u})||p(\mathbf{u}|\mathcal{D})] = \mathbb{E}_q \left[\log \frac{q(\mathbf{u})}{p(\mathbf{u}|\mathcal{D})} \right].$$

As $p(\mathbf{u}|D)$ is unknown, the KL is intractable so finding the *best* distribution $q(\mathbf{u})$ that minimizes the divergence is not feasible. However, observe the fact that

$$\mathrm{KL}[q||p] = \log p(\mathcal{D}) - \underbrace{(\mathbb{E}_q[\log p(\mathcal{D}, \mathbf{u})] + \mathcal{H}_q[q(\mathbf{u})])}_{\textbf{evidence lower bound}} \tag{1}$$

where $\mathcal{H}_q[q(\mathbf{u})]$ is the entropy of $q(\mathbf{u})$ and the **evidence lower bound** term is also known as *negative free energy*

$$\mathcal{L}[q] = \mathbb{E}_q[\log p(\mathcal{D}, \mathbf{u})] + \mathcal{H}_q[q(\mathbf{u})]. \tag{2}$$

As the log evidence $\log p(\mathcal{D})$ in Equation(1) is a constant, minimizing the KL divergence with respect to $q(\mathbf{u})$ is equivalent to maximizing the evidence lower bound. The problem is now turned into an optimization problem whose objective function is the evidence lower bound which we will need to derive for different variational inference methods.

2 Mean-field Variational Inference

In mean-field approximation, we use a family of factorised distribution to approximate the posterior distribution $p(\mathbf{u}|\mathcal{D})$

$$q(\mathbf{u}) = q(\mathbf{f}, \mathbf{w}) = \prod_{j=1}^{Q} q(\mathbf{f}_j) \prod_{i=1}^{P} q(\mathbf{w}_{ij}) \tag{3}$$

Furthermore we limit the class of distribution of all factors to be Gaussians, i.e., $q(\mathbf{f}_j) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{f}_j}, \boldsymbol{\Sigma}_{\mathbf{f}_j})$ and $q(\mathbf{w}_{ij}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}_{ij}}, \boldsymbol{\Sigma}_{\mathbf{w}_{ij}})$. It can be shown that, due to $p(\mathbf{f}_j)$ and $p(\mathbf{w}_{ij})$ being generated from Gaussian processes, these Gaussians must have a full covariance matrix form. This approximation is often refered to as variational Gaussian processes in the literature.

2.1 Closed-From Evidence Lower Bound

In this section we derive the analytical form of the evidence lower bound $\mathcal{L}[q]$ for the assumed family of factorized distribution $q(\mathbf{f}, \mathbf{w})$ in Equation (3). Decomposing the *log-joint* term in Equation (2) into the sum of the *log-likelihood* and *log-prior* terms we have

$$\mathcal{L}[q] = \mathbb{E}_q[\log p(\mathcal{D}|\mathbf{f}, \mathbf{w})] + \mathbb{E}_q[\log p(\mathbf{f}, \mathbf{w})] + \mathcal{H}[q(\mathbf{f}, \mathbf{w})] \tag{4}$$

As we shall see shortly, each of the additive terms in the above equation can be computed analytically, leading to the exact solution of the evidence lower bound.

2.1.1 Expected Log Likelihood

We first compute the expected log likelihood,

$$\mathbb{E}_q[\log p(\mathcal{D}|\mathbf{f},\mathbf{w})] = \int q(\mathbf{f})q(\mathbf{w})\log p(\mathcal{D}|\mathbf{f},\mathbf{w})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{w} = \sum_{n=1}^{N}\int q(\mathbf{f})q(\mathbf{w})\log p(\mathbf{y}_n|\mathbf{f}_n,\mathbf{W}_n)\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{w}. \tag{5}$$

Here the subscript n corresponds to the nth observation; $\mathbf{y}_n$, $\mathbf{f}_n$, and $\mathbf{W}_n$, respectively, are the outputs, latent values, and weights corresponding to observation n. Note that we slight abuse the notations but it should be clear from the context and the subscripts being used. More specifically, subscript n always indicates the nth observation, subscript i indicates the ith output, and subscript j indicates the jth latent function.
Using the identity in Equation (42) (see Appendix) for $\mathbf{f}_n$ and $\mathbf{W}_n$ we get

$$\begin{aligned}
&\int q(\mathbf{f})q(\mathbf{w})\log p(\mathbf{y}_n|\mathbf{f}_n,\mathbf{W}_n)\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{w}\\
=&\int q(\mathbf{f}_n,\{-\mathbf{f}_n\})q(\mathbf{W}_n,\{-\mathbf{W}_n\})\log p(\mathbf{y}_n|\mathbf{f}_n,\mathbf{W}_n)\mathrm{d}\mathbf{f}_n\mathrm{d}\{-\mathbf{f}_n\}\mathrm{d}\mathbf{W}_n\mathrm{d}\{-\mathbf{W}_n\}\\
=&\sum_{n=1}^{N}\int q(\mathbf{f}_n)q(\mathbf{W}_n)\log p(\mathbf{y}_n|\mathbf{f}_n,\mathbf{W}_n)\mathrm{d}\mathbf{f}_n\mathrm{d}\mathbf{W}_n
\end{aligned}$$

where $q(\mathbf{f}_n)$ and $q(\mathbf{W}_n)$ are the posterior marginals of $\mathbf{f}_n$ and $\mathbf{W}_n$ and $\{-\mathbf{f}_n\}$, $\{-\mathbf{W}_n\}$ are the latent and weight function values excluding observation n. Then

$$\mathbb{E}_q[\log p(\mathcal{D}|\mathbf{f},\mathbf{w})] = \sum_{n=1}^{N}\int q(\mathbf{f}_n)q(\mathbf{W}_n)\log p(\mathbf{y}_n|\mathbf{f}_n,\mathbf{W}_n)\mathrm{d}\mathbf{f}_n\mathrm{d}\mathbf{W}_n, \tag{6}$$

hence the *log-likelihood* is the sum of the log-likelihood of individual observations.

Before proceeding to the derivation of the individual log-likelihood we take a look at an interesting property originating from our assumed factorization form for the posterior. This property is indeed what allows us to evaluate the individual log-likelihood term analytically. Since $q(\mathbf{f},\mathbf{w}) = \prod_j \mathbf{f}_j \prod_{i,j}\mathbf{w}_{ij}$, the marginals $q(\mathbf{f}_n)$ and $q(\mathbf{W}_n)$ are Gaussian with diagonal covariance matrices and fully factorize as

$$q(\mathbf{f}_n) = \prod_j q(\mathbf{f}(x_n)_j) = \prod_i \mathcal{N}(\mathbf{f}_j(x_n); (\boldsymbol{\mu}_{\mathbf{f}_j})_n, (\boldsymbol{\Sigma}_{\mathbf{f}_j})_{n,n})$$

$$q(\mathbf{W}_n) = \prod_{i,j} q(\mathbf{W}(x_n)_{i,j}) = \prod_i \mathcal{N}(\mathbf{w}_{ni}; \boldsymbol{\mu}_{\mathbf{w}_{ni}}, \boldsymbol{\Sigma}_{\mathbf{w}_{ni}})$$

where we denote the i-th row of the weight matrix at the nth data point $\mathbf{W}_n = \mathbf{W}(x_n)$ as $\mathbf{w}_{ni}$.

Each individual log-likelihood term in the summation above is the crux of the inference problem so we derive its form analytically in details.

$$\int q(\mathbf{f}_n)q(\mathbf{W}_n)\log p(\mathbf{y}_n|\mathbf{f}_n,\mathbf{W}_n)\mathrm{d}\mathbf{f}_n\mathrm{d}\mathbf{W}_n$$

$$
\begin{aligned}
&= \sum_{i=1}^{P} \int q(\mathbf{f}_n) q(\mathbf{W}_n) \log \mathcal{N}(\mathbf{y}_{ni}; \mathbf{w}_{ni}^T \mathbf{f}_n, \sigma_y^2) \mathrm{d}\mathbf{f}_n \mathrm{d}\mathbf{W}_n \\
&= \sum_{i=1}^{P} \int q(\mathbf{f}_n) q(\mathbf{w}_{ni}) \log \mathcal{N}(\mathbf{y}_{ni}; \mathbf{w}_{ni}^T \mathbf{f}_n, \sigma_y^2) \mathrm{d}\mathbf{f}_n \mathrm{d}\mathbf{w}_{ni} \\
&= -\frac{P}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} \sum_{i=1}^{P} \int q(\mathbf{f}_n) q(\mathbf{w}_{ni}) (\mathbf{y}_{ni} - \mathbf{w}_{ni}^T \mathbf{f}_n)^T (\mathbf{y}_{ni} - \mathbf{w}_{ni}^T \mathbf{f}_n) \mathrm{d}\mathbf{f}_n \mathrm{d}\mathbf{w}_{ni} \\
&= -\frac{P}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} \sum_{i=1}^{P} \int \mathcal{N}(\mathbf{w}_{ni}; \boldsymbol{\mu}_{\mathbf{w}_{ni}}, \boldsymbol{\Sigma}_{\mathbf{w}_{ni}}) \int (\mathbf{w}_{ni}^T \mathbf{f}_n - \mathbf{y}_{ni})^T (\mathbf{w}_{ni}^T \mathbf{f}_n - \mathbf{y}_{ni}) \mathcal{N}(\mathbf{f}_n; \boldsymbol{\mu}_{\mathbf{f}_n}, \boldsymbol{\Sigma}_{\mathbf{f}_n}) \mathrm{d}\mathbf{f}_n \mathrm{d}\mathbf{w}_{ni}
\end{aligned} \tag{7}
$$

where $\mathbf{y}_{ni}$ is the ith output corresponding to observation n.
Applying the identity in Equation (43) for the expectation of a quadratic form wrt a Gaussian twice we get

$$
\begin{aligned}
&\int q(\mathbf{f}_n) q(\mathbf{W}_n) \log p(\mathbf{y}_n | \mathbf{f}_n, \mathbf{W}_n) \mathrm{d}\mathbf{f}_n \mathrm{d}\mathbf{W}_n \\
&= -\frac{P}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} \sum_{i=1}^{P} \Big((\mathbf{y}_{ni} - \boldsymbol{\mu}_{\mathbf{f}_n}^T \boldsymbol{\mu}_{\mathbf{w}_{ni}})^T (\mathbf{y}_{ni} - \boldsymbol{\mu}_{\mathbf{f}_n}^T \boldsymbol{\mu}_{\mathbf{w}_{ni}}) \\
&\qquad + \boldsymbol{\mu}_{\mathbf{w}_{ni}}^T \boldsymbol{\Sigma}_{\mathbf{f}_n} \boldsymbol{\mu}_{\mathbf{w}_{ni}} + \boldsymbol{\mu}_{\mathbf{f}_n}^T \boldsymbol{\Sigma}_{\mathbf{w}_{ni}} \boldsymbol{\mu}_{\mathbf{f}_n} + \mathrm{Tr}[\boldsymbol{\Sigma}_{\mathbf{f}_n} \boldsymbol{\Sigma}_{\mathbf{w}_{ni}}] \Big)
\end{aligned} \tag{8}
$$

The final expression for the expected log likelihood can now be obtained

$$
\begin{aligned}
\mathbb{E}_q[\log p(\mathcal{D}|\mathbf{f}, \mathbf{w})] =& -\frac{NP}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} \sum_{n=1}^{N} \sum_{i=1}^{P} (\mathbf{y}_{ni} - \boldsymbol{\mu}_{\mathbf{f}_n}^T \boldsymbol{\mu}_{\mathbf{w}_{ni}})^T (\mathbf{y}_{ni} - \boldsymbol{\mu}_{\mathbf{f}_n}^T \boldsymbol{\mu}_{\mathbf{w}_{ni}}) \\
&- \frac{1}{2\sigma_y^2} \sum_{n=1}^{N} \sum_{i=1}^{P} \left(\boldsymbol{\mu}_{\mathbf{w}_{ni}}^T \boldsymbol{\Sigma}_{\mathbf{f}_n} \boldsymbol{\mu}_{\mathbf{w}_{ni}} + \boldsymbol{\mu}_{\mathbf{f}_n}^T \boldsymbol{\Sigma}_{\mathbf{w}_{ni}} \boldsymbol{\mu}_{\mathbf{f}_n} + \mathrm{Tr}[\boldsymbol{\Sigma}_{\mathbf{f}_n} \boldsymbol{\Sigma}_{\mathbf{w}_{ni}}] \right) \\
=& -\frac{NP}{2} \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} \sum_{n=1}^{N} (\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{W}_n} \boldsymbol{\mu}_{\mathbf{f}_n})^T (\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{W}_n} \boldsymbol{\mu}_{\mathbf{f}_n}) \\
&- \frac{1}{2\sigma_y^2} \sum_{i=1}^{P} \sum_{j=1}^{Q} \mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{f}_j})^T (\boldsymbol{\mu}_{\mathbf{w}_{ij}} \bullet \boldsymbol{\mu}_{\mathbf{w}_{ij}}) + \mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{w}_{ij}})^T (\boldsymbol{\mu}_{\mathbf{f}_j} \bullet \boldsymbol{\mu}_{\mathbf{f}_j}) \\
=& -\frac{1}{2} NP \log(2\pi\sigma_y^2) - \frac{1}{2\sigma_y^2} \sum_{n=1}^{N} (\mathbf{Y}_{\cdot n}^T - \boldsymbol{\Omega}_{\mathrm{w_n}} \boldsymbol{\nu}_{\mathrm{f_n}})^T (\mathbf{Y}_{\cdot n}^T - \boldsymbol{\Omega}_{\mathrm{w_n}} \boldsymbol{\nu}_{\mathrm{f_n}}) \\
&- \frac{1}{2\sigma_y^2} \sum_{i=1}^{P} \sum_{j=1}^{Q} \Big[\mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{f}_j})^T (\boldsymbol{\mu}_{\mathrm{w_{ij}}} \bullet \boldsymbol{\mu}_{\mathrm{w_{ij}}}) + \mathrm{diag}(\boldsymbol{\Sigma}_{\mathbf{w}_{ij}})^T (\boldsymbol{\mu}_{\mathrm{f_j}} \bullet \boldsymbol{\mu}_{\mathrm{f_j}}) \Big].
\end{aligned} \tag{9}
$$

Here $\mathbf{Y}_{:n}^T$ is the P-dimensional vector of training targets corresponding to observation n; $\mathbf{\Omega}_{\mathrm{w_n}}$ is the $(P \times Q)$-dimensional matrix containing the means for the weight parameters; $\boldsymbol{\nu}_{\mathrm{f_n}}$ is the Q-dimensional vector of means for the latent function parameters; $\mathrm{diag}(\cdot)$ turns the diagonal elements of a matrix into a vector (or viceversa); and $\bullet$ denotes the Hadamard product.

2.1.2 Expected Log Prior

We now compute the expected log prior term

$$\begin{aligned}\mathbb{E}_q[\log p(\mathbf{f}, \mathbf{w})] &= \sum_{j=1}^{Q} \int q(\mathbf{f}_j) \log p(\mathbf{f}_j) \mathrm{d}\mathbf{f}_j + \sum_{i,j} \int q(\mathbf{w}_{ij}) \log p(\mathbf{w}_{ij}) \mathrm{d}\mathbf{w}_{ij} \\ &= \sum_{j=1}^{Q} \mathbb{E}_{q(\mathbf{f}_j)}[\log \mathcal{N}(\mathbf{f}_j; 0, \mathbf{K}_f)] + \sum_{i,j} \mathbb{E}_{q(\mathbf{w}_{ij})}[\log \mathcal{N}(\mathbf{w}_{ij}; 0, \mathbf{K}_w)] \end{aligned} \tag{10}$$

For each latent function $\mathbf{f}_j$,

$$\begin{aligned}\mathbb{E}_{q(\mathbf{f}_j)}[\log \mathcal{N}(\mathbf{f}_j; 0, \mathbf{K}_f)] &= \int \mathcal{N}(\mathbf{f}_j; \boldsymbol{\mu}_{\mathbf{f}_j}, \mathbf{\Sigma}_{\mathbf{f}_j}) \log \mathcal{N}(\mathbf{f}_j; 0, \mathbf{K}_f) \mathrm{d}\mathbf{f}_j \\ &= -\frac{1}{2} \int \mathbf{f}_j^T \mathbf{K}_f^{-1} \mathbf{f}_j \mathcal{N}(\mathbf{f}_j; \boldsymbol{\mu}_{\mathbf{f}_j}, \mathbf{\Sigma}_{\mathbf{f}_j}) - \frac{1}{2} \log |\mathbf{K}_f| \\ &= -\frac{1}{2} \left(\log |\mathbf{K}_f| + \boldsymbol{\mu}_{\mathbf{f}_j}^T \mathbf{K}_f^{-1} \boldsymbol{\mu}_{\mathbf{f}_j} + \mathrm{Tr}(\mathbf{K}_f^{-1} \mathbf{\Sigma}_{\mathbf{f}_j}) \right) \end{aligned}$$

Similarly for each weight function $\mathbf{w}_{ij}$,

$$\mathbb{E}_{q(\mathbf{w}_{ij})}[\log \mathcal{N}(\mathbf{w}_{ij}; 0, \mathbf{K}_w)] = -\frac{1}{2} \left(\log |\mathbf{K}_{\mathbf{w}_{ij}}| + \boldsymbol{\mu}_{\mathbf{w}_{ij}}^T \mathbf{K}_w^{-1} \boldsymbol{\mu}_{\mathbf{w}_{ij}} + \mathrm{Tr}(\mathbf{K}_{\mathbf{w}_{ij}} \mathbf{\Sigma}_{\mathbf{w}_{ij}}^{-1}) \right)$$

The final expression for the expected log prior is

$$\begin{aligned}\mathbb{E}_q[\log p(\mathbf{f}, \mathbf{w})] = &-\frac{1}{2} \sum_{j=1}^{Q} \left(\log |\mathbf{K}_f| + \boldsymbol{\mu}_{\mathbf{f}_j}^T \mathbf{K}_f^{-1} \boldsymbol{\mu}_{\mathbf{f}_j} + \mathrm{Tr}(\mathbf{K}_f^{-1} \mathbf{\Sigma}_{\mathbf{f}_j}) \right) \\ &-\frac{1}{2} \sum_{i,j} \left(\log |\mathbf{K}_{\mathbf{w}_{ij}}| + \boldsymbol{\mu}_{\mathbf{w}_{ij}}^T \mathbf{K}_w^{-1} \boldsymbol{\mu}_{\mathbf{w}_{ij}} + \mathrm{Tr}(\mathbf{K}_{\mathbf{w}_{ij}} \mathbf{\Sigma}_{\mathbf{w}_{ij}}^{-1}) \right) \end{aligned} \tag{11}$$

2.1.3 Entropy

The entropy term is

$$\begin{aligned}\mathcal{H}[q(\mathbf{f}, \mathbf{w})] &= -\int q(\mathbf{f}, \mathbf{w}) \log q(\mathbf{f}, \mathbf{w}) \mathrm{d}\mathbf{f} \mathrm{d}\mathbf{w} \\ &= \sum_{j=1}^{Q} \mathcal{H}[q(\mathbf{f}_j)] + \sum_{i,j} \mathcal{H}[q(\mathbf{w}_{ij})] \end{aligned}$$

$$= \frac{1}{2}\sum_{j=1}^{Q} \log |\mathbf{\Sigma}_{\mathbf{f}_j}| + \frac{1}{2}\sum_{i,j} \log |\mathbf{\Sigma}_{\mathbf{w}_{ij}}| + \text{const} \tag{12}$$

2.2 Efficient Closed-Form Updates for Variational Parameters

We now derive the best approximate distribution $q(\mathbf{u})$ using standard results in the mean-field theory. The optimum distributions for the latent and weight functions are

$$\log q(\mathbf{f}_j) = \mathbb{E}_{-\mathbf{f}_j} \log p(\mathcal{D}, \mathbf{u}) + \text{const} = \mathbb{E}_{-\mathbf{f}_j} \log p(\mathbf{u}) + \mathbb{E}_{-\mathbf{f}_j} \log p(\mathcal{D}|\mathbf{u}) + \text{const} \tag{13}$$

$$\log q(\mathbf{w}_{ij}) = \mathbb{E}_{-\mathbf{w}_{ij}} \log p(\mathcal{D}, \mathbf{u}) + \text{const} = \mathbb{E}_{-\mathbf{w}_{ij}} \log p(\mathbf{u}) + \mathbb{E}_{-\mathbf{w}_{ij}} \log p(\mathcal{D}|\mathbf{u}) + \text{const} \tag{14}$$

where the expectations are taken with respect to $q(\mathbf{u})$ and $-\mathbf{f}_j$ and $-\mathbf{w}_{ij}$ denote the set of all variables excluding $\mathbf{f}_j$ and $\mathbf{w}_{ij}$. Identical results can be obtained by setting the gradients of the evidence lower bound to zero.

2.2.1 Updating Equations for the Latent Functions

We first derive the updating equations for the latent functions. Since each $q(\mathbf{f}_j)$ is a full Gaussian, its variational parameters are the mean and covariance matrix which we denote as $\boldsymbol{\mu}_{\mathbf{f}_j}$ and $\mathbf{\Sigma}_{\mathbf{f}_j}$.
Expectation from the prior term in Equation (13) is

$$\mathbb{E}_{-\mathbf{f}_j} \log p(\mathbf{u}) = \mathbb{E}_{-\mathbf{f}_j} \log p(\mathbf{f}_j) = -\frac{1}{2}\log |K_f| - \frac{1}{2}\mathbf{f}_j^T K_f^{-1} \mathbf{f}_j \tag{15}$$

Expectation from the likelihood term in Equation (13) is

$$\begin{aligned}\mathbb{E}_{-\mathbf{f}_j} \log p(\mathbf{y}|\mathbf{u}) &= \mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{W}_n, \mathbf{f}_n) \\ &= -\frac{1}{2\sigma_y^2}\left(\mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \mathbf{f}_n^T \mathbf{W}_n^T \mathbf{W}_n \mathbf{f}_n - 2\mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \mathbf{y}_n^T \mathbf{W}_n \mathbf{f}_n\right)\end{aligned} \tag{16}$$

Omitting a few steps of derivation (see Appendix), expectation of the quadratic term in Equation (16) is

$$\mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \mathbf{f}_n^T \mathbf{W}_n^T \mathbf{W}_n \mathbf{f}_n = \mathbf{f}_j^T \mathbb{E}\left[\sum_{i=1}^{P} \text{diag}(\mathbf{w}_{ij} \bullet \mathbf{w}_{ij})\right] \mathbf{f}_j + 2\mathbf{f}_j^T \sum_{k \neq j} \mathbb{E}\left[\sum_{i=1}^{P} \text{diag}(\mathbf{w}_{ik} \bullet \mathbf{w}_{ij})\right] \mathbb{E}[\mathbf{f}_k] \tag{17}$$

and expectation of the linear term (see Appendix) is

$$\mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \mathbf{y}_n^T \mathbf{W}_n \mathbf{f}_n = \mathbf{f}_j^T \sum_{i=1}^{P} \mathbf{Y}_{\cdot i} \bullet \mathbb{E}[\mathbf{w}_{ij}]. \tag{18}$$

where $\mathbf{Y}_{\cdot i}$ is the N-dimensional vector of observations corresponding to output i. All terms that do not contain $\mathbf{f}_j$ are absorbed into the const term in Equation (13).

Combining all equations we get

$$\begin{aligned}\log q(\mathbf{f}_j) &= -\frac{1}{2}\mathbf{f}_j^T K_f^{-1}\mathbf{f}_j - \frac{1}{2\sigma_y^2}\mathbf{f}_j^T E[\sum_{i=1}^{P}\operatorname{diag}(\mathbf{w}_{ij}\bullet\mathbf{w}_{ij})]\mathbf{f}_j \\ &\quad - \frac{1}{\sigma_y^2}\mathbf{f}_j^T\sum_{k\neq j}\mathbb{E}[\sum_{i=1}^{P}\operatorname{diag}(\mathbf{w}_{ik}\bullet\mathbf{w}_{ij})]E[\mathbf{f}_k] + \frac{1}{\sigma_y^2}\mathbf{f}_j^T\sum_{i=1}^{P}\mathbf{Y}_{\cdot i}\bullet\mathbb{E}[\mathbf{w}_{ij}] + \text{const} \\ &= -\frac{1}{2}\mathbf{f}_j^T\boldsymbol{\Sigma}_{\mathbf{f}_j}\mathbf{f}_j + \mathbf{f}_j^T\boldsymbol{\Sigma}_{\mathbf{f}_j}^{-1}\boldsymbol{\mu}_{\mathbf{f}_j} + \text{const.} \qquad (19)\end{aligned}$$

Completing the square gives the following update equations for the variational parameters of a latent function $\mathbf{f}_j$

$$\boldsymbol{\Sigma}_{\mathbf{f}_j} = \left(K_f^{-1} + \frac{1}{\sigma_y^2}\sum_{i=1}^{P}\operatorname{diag}(\boldsymbol{\mu}_{\mathbf{w}_{ij}}\bullet\boldsymbol{\mu}_{\mathbf{w}_{ij}} + \operatorname{Var}(\mathbf{w}_{ij}))\right)^{-1} \qquad (20)$$

$$\boldsymbol{\mu}_{\mathbf{f}_j} = \frac{1}{\sigma_y^2}\boldsymbol{\Sigma}_{\mathbf{f}_j}\sum_{i=1}^{P}\left(\mathbf{Y}_{\cdot i} - \sum_{k\neq j}\boldsymbol{\mu}_{\mathbf{w}_{ik}}\bullet\boldsymbol{\mu}_{\mathbf{f}_k}\right)\bullet\boldsymbol{\mu}_{\mathbf{w}_{ij}} \qquad (21)$$

where $\operatorname{Var}(\mathbf{w}_{ij})$ is the diagonal of $\boldsymbol{\Sigma}_{\mathbf{w}_{ij}}$.

2.2.2 Updating Equations for the Weight Functions

Derivations of the update equations for the weight functions are done in analogy to the previous section. Expectation from the prior term in 14 is

$$\mathbb{E}_{-\mathbf{w}_{ij}}\log p(\mathbf{u}) = \mathbb{E}_{-\mathbf{w}_{ij}}\log p(\mathbf{w}_{ij}) = -\frac{1}{2}\log|K_w| - \frac{1}{2}\mathbf{w}_{ij}^T K_w^{-1}\mathbf{w}_{ij} \qquad (22)$$

Expection from the likelihood term in 14 is

$$\begin{aligned}\mathbb{E}_{-\mathbf{w}_{ij}}\log p(\mathbf{y}|\mathbf{u}) &= \mathbb{E}_{-\mathbf{w}_{ij}}\sum_{n=1}^{N}\log p(\mathbf{y}_n|\mathbf{W}_n,\mathbf{f}_n) \\ &= \mathbb{E}_{-\mathbf{w}_{ij}}\sum_{n=1}^{N}\left(-\frac{1}{2\sigma_y^2}(\mathbf{y}_n - \mathbf{W}_n\mathbf{f}_n)^T(\mathbf{y}_n - \mathbf{W}_n\mathbf{f}_n)\right) \\ &= -\frac{1}{2\sigma_y^2}\left(\mathbb{E}_{-\mathbf{w}_{ij}}\sum_{n=1}^{N}\mathbf{f}_n^T\mathbf{W}_n^T\mathbf{W}_n\mathbf{f}_n - 2\mathbb{E}_{-\mathbf{w}_{ij}}\sum_{n=1}^{N}\mathbf{y}_n^T\mathbf{W}_n\mathbf{f}_n\right) \qquad (23)\end{aligned}$$

Omitting a few steps of derivation, expectation of the quadratic term in Equation (23) is

$$\mathbb{E}_{-\mathbf{w}_{ij}}\sum_{n=1}^{N}\mathbf{f}_n^T\mathbf{W}_n^T\mathbf{W}_n\mathbf{f}_n = \mathbf{w}_{ij}^T\mathbb{E}[\operatorname{diag}(\mathbf{f}_j\bullet\mathbf{f}_j)]\mathbf{w}_{ij} + 2\mathbf{w}_{ij}^T\sum_{k\neq j}\mathbb{E}[\operatorname{diag}(\mathbf{f}_k\bullet\mathbf{f}_j)]\mathbb{E}[\mathbf{w}_{ik}] \qquad (24)$$

and expectation of the linear term is

$$\mathbb{E}_{-\mathbf{w}_{ij}} \sum_{n=1}^{N} \mathbf{y}_n^T \mathbf{W}_n \mathbf{f}_n = \mathbf{w}_{ij}^T (\mathbf{Y}_{\cdot i} \bullet \boldsymbol{\mu}_{\mathbf{f}_j}). \tag{25}$$

The updating equations for variational parameters of $\mathbf{w}_{ij}$ are

$$\boldsymbol{\Sigma}_{\mathbf{w}_{ij}} = \left(K_w^{-1} + \frac{1}{\sigma_y^2} \operatorname{diag}(\boldsymbol{\mu}_{\mathbf{f}_j} \bullet \boldsymbol{\mu}_{\mathbf{f}_j} + \operatorname{Var}(\mathbf{f}_j)) \right)^{-1} \tag{26}$$

$$\boldsymbol{\mu}_{\mathbf{w}_{ij}} = \frac{1}{\sigma_y^2} \boldsymbol{\Sigma}_{\mathbf{w}_{ij}} (\mathbf{Y}_{\cdot i} - \sum_{k \neq j} \boldsymbol{\mu}_{\mathbf{f}_k} \bullet \boldsymbol{\mu}_{\mathbf{w}_{ik}}) \bullet \boldsymbol{\mu}_{\mathbf{f}_j} \tag{27}$$

where $\operatorname{Var}(\mathbf{f}_j)$ is the diagonal of $\boldsymbol{\Sigma}_{\mathbf{f}_j}$. From Equations 20 and 26, we see that only N parameters on the diagonals are needed to parametrize the full covariance matrices of all latent and weight functions. Effectively the number of parameters to be kept in memory and optimized is significantly reduced, which can lead to better learning of parameters (e.g. reduce overfitting) and allow us to handle larger datasets.

2.3 Learning of the Hyper-parameters

Variational inference for Gaussian processes often allows easy handling of the hyper-parameters. Specifically, the **evidence lower bound** in Equation (4) is implicitly conditioned on the hyper-parameters. Keeping the variational parameters fixed, the objective function is a function of the hyper-parameters and as a result, it can also be optimized wrt to the hyper-parameters using, for example, gradient-based methods.

2.3.1 Derivatives wrt the Noise σ_y

As a function of σ_y, the only contribution in the **evidence lower bound** (Equation (4)) is from the likelihood term. Taking the derivatives of $\mathbb{E}_q[\log p(\mathcal{D}|\mathbf{f}, \mathbf{w})]$ in Equation (9) σ_y w.r.t is straightforward algebra and hence omitted.

2.3.2 Derivatives wrt to $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_w$

As a function of $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_w$, the only contribution in the **evidence lower bound** is from the prior term,

$$\mathcal{L}[q] = \mathbb{E}_q[\log p(\mathcal{D}|\mathbf{f}, \mathbf{w})] + \mathbb{E}_q[\log p(\mathbf{f}, \mathbf{w})] + \mathcal{H}[q] = \mathbb{E}_q[\log p(\mathbf{f}, \mathbf{w})] + \text{const}$$

The derivative of a latent function $\mathbf{f}_j$ w.r.t to the $t-$th element of the $\boldsymbol{\theta}_f$ is

$$\begin{aligned} \frac{\mathrm{d}\mathbb{E}_{q(\mathbf{f}_j)}[\log \mathcal{N}(\mathbf{f}_j; 0, \mathbf{K}_f)]}{\mathrm{d}(\boldsymbol{\theta}_f)_t} &= -\frac{1}{2} \left[\operatorname{Tr}(\mathbf{K}_f^{-1} \frac{\mathrm{d}\mathbf{K}_f}{\mathrm{d}(\boldsymbol{\theta}_f)_t}) - \boldsymbol{\mu}_{\mathbf{f}_j}^T \mathbf{K}_f^{-1} \frac{\mathrm{d}\mathbf{K}_f}{\mathrm{d}(\boldsymbol{\theta}_f)_t} \mathbf{K}_f^{-1} \boldsymbol{\mu}_{\mathbf{f}_j} - \operatorname{Tr}(\mathbf{K}_f^{-1} \frac{\mathrm{d}\mathbf{K}_f}{\mathrm{d}(\boldsymbol{\theta}_f)_t} \mathbf{K}_f^{-1} \boldsymbol{\Sigma}_{\mathbf{f}_j}) \right] \\ &= \frac{1}{2} \operatorname{Tr} \left(-\mathbf{K}_f^{-1} \frac{\mathrm{d}\mathbf{K}_f}{\mathrm{d}(\boldsymbol{\theta}_f)_t} + \boldsymbol{\mu}_{\mathbf{f}_j}^T \mathbf{K}_f^{-1} \frac{\mathrm{d}\mathbf{K}_f}{\mathrm{d}(\boldsymbol{\theta}_f)_t} \mathbf{K}_f^{-1} \boldsymbol{\mu}_{\mathbf{f}_j} + \mathbf{K}_f^{-1} \frac{\mathrm{d}\mathbf{K}_f}{\mathrm{d}(\boldsymbol{\theta}_f)_t} \mathbf{K}_f^{-1} \boldsymbol{\Sigma}_{\mathbf{f}_j} \right) \end{aligned}$$

$$= \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_f^{-1}(\boldsymbol{\Sigma}_{\mathbf{f}_j} + \boldsymbol{\mu}_{\mathbf{f}_j}\boldsymbol{\mu}_{\mathbf{f}_j}^T - \mathbf{K}_f)\mathbf{K}_f^{-1}\frac{\mathrm{d}\mathbf{K}_f}{\mathrm{d}(\boldsymbol{\theta}_f)_t}\right). \tag{28}$$

Hence the derivatives of the lower bound $\mathcal{L}[q]$ w.r.t the hyperparameters of the covariance function of the latent functions is

$$\frac{\mathrm{d}\mathcal{L}[q]}{\mathrm{d}\theta_{f_t}} = \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}_f^{-1}(\sum_{j=1}^{Q}\boldsymbol{\Sigma}_{\mathbf{f}_j} + \boldsymbol{\mu}_{\mathbf{f}_j}\boldsymbol{\mu}_{\mathbf{f}_j}^T - \mathbf{K}_f)\mathbf{K}_f^{-1}\frac{\mathrm{d}\mathbf{K}_f}{\mathrm{d}(\boldsymbol{\theta}_f)_t}\right) \tag{29}$$

Similarly for the hyper-parameters of covariance function of weights:

$$\frac{\mathrm{d}\mathcal{L}[q]}{\mathrm{d}(\boldsymbol{\theta}_w)_t} = \frac{1}{2}\mathrm{Tr}\left(\sum_{i,j}\mathbf{K}_w^{-1}(\boldsymbol{\Sigma}_{\mathbf{w}_{ij}} + \boldsymbol{\mu}_{\mathbf{w}_{ij}}\boldsymbol{\mu}_{\mathbf{w}_{ij}}^T - \mathbf{K}_{\mathbf{w}_{ij}})\mathbf{K}_w^{-1}\frac{\mathrm{d}\mathbf{K}_{\mathbf{w}_{ij}}}{\mathrm{d}(\boldsymbol{\theta}_w)_t}\right) \tag{30}$$

The derivatives of the covariance matrices $\mathbf{K}_f$ and $\mathbf{K}_w$ w.r.t the hyper-parameters $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_w$ depend on the functional form of the covariance functions. For squared exponential covariance functions, the derivatives are straightforward.

3 Nonparametric Variational Inference for GPRN

In nonparametric variational inference we approximate the posterior $p(\mathbf{u}|\mathbf{y})$ with the mixture of isotropic Gaussians family

$$q(\mathbf{u}) = \frac{1}{K}\sum_{k=1}^{K} q^{(k)}(\mathbf{u}) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\mathbf{u}; \boldsymbol{\mu}^{(k)}, \sigma_k\mathbf{I}) \tag{31}$$

The evidence lower bound corresponding to this posterior approximation is

$$\mathcal{L}[q] = \mathbb{E}_q[\log p(\mathcal{D}, \mathbf{u})] + \mathcal{H}[q(\mathbf{u})]$$

The entropy term $\mathcal{H}[q(\mathbf{u})]$ does not have closed-form expression as $q(\mathbf{u})$ is a Gaussian mixture. However it can be bounded using Jensen's inequality,

$$\mathcal{H}[q(\mathbf{u})] \geq -\frac{1}{K}\sum_{k=1}^{K}\log\frac{1}{K}\sum_{j=1}^{K}\mathcal{N}(\boldsymbol{\mu}^{(k)}; \boldsymbol{\mu}^{(j)}, (\sigma_k^2 + \sigma_j^2)\mathbf{I}). \tag{32}$$

It is instructive to note that an upper bound for this entropy also exists (see Huber et al. (2008)) and so an average of the lower and upper bound may be used to approximate the entropy. Alternatively a second-order Taylor approximation may also suffice. Our use of the lower bound only is mainly for saving of computation.
The expected log joint $\mathbb{E}_q[\log p(\mathcal{D}, \mathbf{u})]$ is the expectation of a non-linear function under a mixture of Gaussian and there is no analytical form in general. However we show that a close-form expression can be obtained under the likelihood model of GPRN and the

factorized assumption of the mixture component of the approximate posterior. First we have,

$$\mathbb{E}_q[\log p(\mathcal{D}, \mathbf{u})] = \int q(\mathbf{u}) \log p(\mathcal{D}, \mathbf{u}) \mathrm{d}\mathbf{u} = \sum_{k=1}^{K} \int q^{(k)}(\mathbf{u}) \log p(\mathcal{D}, \mathbf{u}) \mathrm{d}\mathbf{u} = \sum_{k=1}^{K} \mathbb{E}_{q^{(k)}}[\log p(\mathcal{D}, \mathbf{u})]$$

Each term in the summation is the expectation of the log joint under a component of the mixture posterior. As the covariance matrices of the components are isotropic, the Gaussians fully factorize over the latent functions $\mathbf{f}_j$ and weight functions $\mathbf{w}_{ij}$. Therefore we can apply our results in mean-field to derive an exact expression for the expected log joint,

$$\begin{aligned}
\mathbb{E}_q[\log p(\mathcal{D}, \mathbf{f}, \mathbf{w})] = & -\frac{1}{2K} \sum_{k=1}^{K} \sum_{j=1}^{Q} \left[\log |\mathbf{K}_f| + (\boldsymbol{\mu}_{\mathbf{f}_j}^{(k)})^T \left(\mathbf{K}_f^{-1} + \frac{P\sigma_k^2}{\sigma_y^2} \mathbf{I} \right) \boldsymbol{\mu}_{\mathbf{f}_j}^{(k)} + \sigma_k^2 \mathrm{Tr}(\mathbf{K}_f^{-1}) \right] \\
& - \frac{1}{2K} \sum_{k=1}^{K} \sum_{i=1}^{P} \sum_{j=1}^{Q} \left[\log |\mathbf{K}_w| + (\boldsymbol{\mu}_{\mathbf{w}_{ij}}^{(k)})^T \left(\mathbf{K}_w^{-1} + \frac{P\sigma_k^2}{\sigma_y^2} \mathbf{I} \right) \boldsymbol{\mu}_{\mathbf{w}_{ij}}^{(k)} + \sigma_k^2 \mathrm{Tr}(\mathbf{K}_w^{-1}) \right] \\
& - \frac{1}{2K\sigma_y^2} \sum_{k=1}^{K} \sum_{n=1}^{N} \left(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{W}_n}^{(k)} \boldsymbol{\mu}_{\mathbf{f}_n}^{(k)} \right)^T \left(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{W}_n}^{(k)} \boldsymbol{\mu}_{\mathbf{f}_n}^{(k)} \right) \\
& - \frac{1}{2K} \sum_{k=1}^{K} \frac{\sigma_k^4}{\sigma_y^2} NPQ - \frac{1}{2} NP \log(2\pi\sigma_y^2)
\end{aligned} \tag{33}$$

From Equations (32) and (33) we get a proper bound for the **evidence lower bound** of nonparametric variational inference for the GPRN model. This is a significant improvement over the original paper where the expected *log-joint* is approximated using the 2-nd order Taylor expansion. Having an analytical lower bound guarantees that our optimization procedure will converge. This also enables the optimization of the hyper-parameters in a variational EM-like manner.

3.1 Derivatives of Evidence Lower Bound wrt the Variational Parameters and Hyper-parameters

Here again we draw the reader's attention to Equation 33 which contains the analytical form of the evidence lower bound. As has been done multiple times in this note, the derivatives of the variational parameters as well as the hyper-parameters can be taken almost exactly from the respective equations for the mean-field methods. The key difference between the lower bound GPRN-NPV and GPRN-MF is from the entropy term of which the derivatives can be compute as follows.

3.1.1 Derivatives of the Entropy Lower Bound wrt Variational Parameters

The derivatives of the (negative) lower bound of mixtures (not including the $1/K$ factor yet) with respect to a mean $\boldsymbol{\mu}_k$ is

$$\sum_{j=1}^{K} \frac{\mathrm{d}\log q_j}{\mathrm{d}\boldsymbol{\mu}_k} = \frac{\mathrm{d}\log q_k}{\mathrm{d}\boldsymbol{\mu}_k} + \sum_{j\neq k} \frac{\mathrm{d}\log q_j}{\mathrm{d}\boldsymbol{\mu}_k} = \frac{1}{K}\sum_{j=1}^{K} \left(\frac{\mathcal{N}_{kj}}{q_k} + \frac{\mathcal{N}_{kj}}{q_j}\right) \frac{\boldsymbol{\mu}_j - \boldsymbol{\mu}_k}{\sigma_k^2 + \sigma_j^2} \tag{34}$$

where $q_k = \frac{1}{K}\sum_{j=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k; \boldsymbol{\mu}_j, (\sigma_k^2+\sigma_j^2)\mathbf{I})$ and $\mathcal{N}_{kj}$ is a short-hand notation for $\mathcal{N}(\boldsymbol{\mu}_k; \boldsymbol{\mu}_j, (\sigma_k^2+\sigma_j^2)\mathbf{I})$.

The term $\frac{\mathcal{N}_{kj}}{q_k} + \frac{\mathcal{N}_{kj}}{q_j}$ in Equation (34) contains very small probabilities. We must express it log space and carry the computation in that space to ensure numerical stability.

$$\frac{1}{K}\left(\frac{\mathcal{N}_{kj}}{q_n} + \frac{\mathcal{N}_{kj}}{q_j}\right) = \frac{1}{K}\left(\frac{\mathcal{N}_{kj}}{\frac{1}{K}\sum_{j'=1}^{K}\mathcal{N}_{kj'}} + \frac{\mathcal{N}_{kj}}{\frac{1}{K}\sum_{j'=1}^{K}\mathcal{N}_{jj'}}\right) \tag{35}$$

$$= \left(\frac{1}{\sum_{k=1}^{K}\exp(\log\mathcal{N}_{nk} - \log\mathcal{N}_{nj})} + \frac{1}{\sum_{k=1}^{K}\exp(\log\mathcal{N}_{jk} - \log\mathcal{N}_{jn})}\right) \tag{36}$$

3.1.2 Derivatives of the Entropy Lower Bound wrt the Hyperparameters

The derivatives with respect to a covariance scale σ_n^2 (not including the $1/K$ factor yet) is

$$\sum_{j=1}^{K} \frac{\mathrm{d}\log q_j}{\mathrm{d}\sigma_k^2} = \frac{\mathrm{d}\log q_k}{\mathrm{d}\sigma_k^2} + \sum_{j\neq k} \frac{\mathrm{d}\log q_j}{\mathrm{d}\sigma_k^2}$$

$$= \frac{1}{2K}\sum_{j=1}^{K} \left(\frac{\mathcal{N}_{kj}}{q_k} + \frac{\mathcal{N}_{kj}}{q_j}\right)\left(\frac{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)}{(\sigma_k^2+\sigma_j^2)^2} - \frac{D}{\sigma_k^2+\sigma_j^2}\right) \tag{37}$$

4 Predictive Distributions

For a new input location $\mathbf{x}^*$ we can use the approximate posterior to predict its outputs $\mathbf{y}^* = \mathbf{y}(\mathbf{x}^*)$. The predictive distribution by mean-field is

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathcal{D}) = \int p(\mathbf{y}^*|\mathbf{W}^*, \mathbf{f}^*)p(\mathbf{W}^*, \mathbf{f}^*|\mathbf{u})q(\mathbf{u})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{W}^*\mathrm{d}\mathbf{f}^* \tag{38}$$

and that of nonparametric approximation is

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathcal{D}) = \frac{1}{K}\sum_{k=1}^{K} \int p(\mathbf{y}^*|\mathbf{W}^*, \mathbf{f}^*)p(\mathbf{W}^*, \mathbf{f}^*|\mathbf{u})q^{(k)}(\mathbf{u})\mathrm{d}\mathbf{u}\mathrm{d}\mathbf{W}^*\mathrm{d}\mathbf{f}^* \tag{39}$$

where $\mathbf{W}^* = \mathbf{W}(\mathbf{x}^*)$ and $\mathbf{f}^* = \mathbf{f}(\mathbf{x}^*)$ and subscript k denotes the k-th component of the mixture posterior. The predictive distributions for both approximations are analytically intractable due to the non-Gaussian likelihood of the GPRN models. However their predicted means can be computed analytically.

We first derive the predictive mean for the mean-field method:

$$\begin{aligned}
\mathbb{E}[\mathbf{y}^*|\mathcal{D},\mathbf{x}^*] &= \int \mathbf{y}^* p(\mathbf{y}^*|\mathcal{D},\mathbf{x}^*)\mathrm{d}\mathbf{y}^* \\
&= \int\int\int \mathbf{y}^* p(\mathbf{y}^*,\mathbf{f}^*,\mathbf{W}^*|\mathcal{D},\mathbf{x}^*)\mathrm{d}\mathbf{y}^*\mathrm{d}\mathbf{f}^*\mathrm{d}\mathbf{W}^* \\
&= \int\int\int \mathbf{y}^* p(\mathbf{y}^*|\mathbf{f}^*,\mathbf{W}^*)p(\mathbf{f}^*,\mathbf{W}^*|\mathcal{D},\mathbf{x}^*)\mathrm{d}\mathbf{f}^*\mathrm{d}\mathbf{W}^*\mathrm{d}\mathbf{y}^* \\
&= \int\int \mathbb{E}[\mathbf{y}^*|\mathbf{f}^*,\mathbf{W}^*]p(\mathbf{f}^*,\mathbf{W}^*|\mathcal{D},\mathbf{x}^*)\mathrm{d}\mathbf{f}^*\mathrm{d}\mathbf{W}^* \\
&= \int\int \mathbf{W}^*\mathbf{f}^* p(\mathbf{f}^*,\mathbf{W}^*|\mathcal{D},\mathbf{x}^*)\mathrm{d}\mathbf{f}^*\mathrm{d}\mathbf{W}^* \\
&= \int\int\int\int \mathbf{W}^*\mathbf{f}^* p(\mathbf{f}^*,\mathbf{W}^*|\mathbf{f},\mathbf{w},\mathbf{x}^*)p(\mathbf{f},\mathbf{w}|\mathcal{D})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{f}^*\mathrm{d}\mathbf{W}^* \\
&= \int\int \mathbb{E}[\mathbf{W}^*\mathbf{f}^*|\mathbf{f},\mathbf{w},\mathbf{x}^*]p(\mathbf{f},\mathbf{w})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{w} \\
&= \int\int \mathbb{E}[\mathbf{W}^*|\mathbf{w},\mathbf{x}^*]\mathbb{E}[\mathbf{f}^*|\mathbf{f},\mathbf{x}^*]p(\mathbf{f},\mathbf{w}|\mathcal{D})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{w} \\
&= \mathbb{E}[\mathbf{W}^*|\mathbf{w},\mathbf{x}^*]\mathbb{E}[\mathbf{f}^*|\mathbf{f},\mathbf{x}^*] \\
&= \mathbf{K}_w^*\mathbf{K}_w^{-1}\boldsymbol{\mu}_\mathbf{w}\mathbf{K}_f^*\mathbf{K}_f^{-1}\boldsymbol{\mu}_\mathbf{f} \qquad (40)
\end{aligned}$$

Here $\mathbf{K}_w^*$ and $\mathbf{K}_f^*$ are the covariance matrices corresponding to the covariance functions κ_w and κ_f evaluated on the test point $\mathbf{x}^*$ wrt all of the training data; $\boldsymbol{\mu}_\mathbf{w}$ and $\boldsymbol{\mu}_\mathbf{f}$ are the mean of the latent and weight functions, respectively.

For nonparametric variational inference with mixture posterior, the predictive mean is simply the average of the predictions made by all components:

$$\mathbb{E}[\mathbf{y}^*|\mathbf{x}^*,\mathcal{D}] = \frac{1}{K}\sum_{k=1}^{K}\mathbf{K}_w^*\mathbf{K}_w^{-1}\boldsymbol{\mu}_\mathbf{w}^{(k)}\mathbf{K}_f^*\mathbf{K}_f^{-1}\boldsymbol{\mu}_\mathbf{f}^{(k)} \qquad (41)$$

5 Appendix

Convenient facts and identities used in derivations. Below $\mathbf{x}$ and $\mathbf{y}$ are vectors and $\mathbf{D}$ is diagonal matrix.

$$\int p(a,b)h(a)\mathrm{d}a\mathrm{d}b = \int p(a)h(a)\int p(b|a)\mathrm{d}b\mathrm{d}a = \int p(a)h(a)\mathrm{d}a \qquad (42)$$

$$\int_x (\mathbf{W}\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{W}\mathbf{x}-\boldsymbol{\mu})^T\mathcal{N}(\mathbf{m},\mathbf{S}) = (\boldsymbol{\mu}-\mathbf{W}\mathbf{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{W}\mathbf{m}) + \mathrm{Tr}[\mathbf{W}^T\boldsymbol{\Sigma}^{-1}\mathbf{W}\mathbf{S}] \qquad (43)$$

where $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$.

$$\mathbf{x}^T \mathbf{D} \mathbf{x} = \sum \mathbf{x}_i \mathbf{D}_{ii} \mathbf{x}_i = \sum \mathbf{D}_{ii} \mathbf{x}_i^2 \tag{44}$$

$$\mathbf{x}^T \mathbf{D} \mathbf{y} = \sum \mathbf{x}_i \mathbf{D}_{ii} \mathbf{y}_i \tag{45}$$

$$\mathbf{f}_n^T \mathbf{W}_n^T \mathbf{W}_n \mathbf{f}_n = \sum_{k=1}^{Q} \sum_{k'=1}^{Q} \mathbf{f}_k(\mathbf{x}_n) \mathbf{f}_{k'}(\mathbf{x}_n) \sum_{i=1}^{P} \mathbf{w}_{ik}(\mathbf{x}_n) \mathbf{w}_{ik'}(\mathbf{x}_n) \tag{46}$$

Detailed derivation for Equation (17):

$$\begin{aligned}
&\mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \mathbf{f}_n^T \mathbf{W}_n^T \mathbf{W}_n \mathbf{f}_n \\
&= \mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \sum_{k=1}^{Q} \sum_{k'=1}^{Q} \mathbf{f}_k(\mathbf{x}_n) \mathbf{f}_{k'}(\mathbf{x}_n) \sum_{i=1}^{P} \mathbf{w}_{ik}(\mathbf{x}_n) \mathbf{w}_{ik'}(\mathbf{x}_n) \\
&= \mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \mathbf{f}_j(\mathbf{x}_n) \mathbf{f}_j(\mathbf{x}_n) \sum_{i=1}^{P} \mathbf{w}_{ij}(\mathbf{x}_n) \mathbf{w}_{ij}(\mathbf{x}_n) + 2\mathbb{E}_{-\mathbf{f}_j} \sum_{k \neq j}^{Q} \sum_{n=1}^{N} \mathbf{f}_j(\mathbf{x}_n) \mathbf{f}_k(\mathbf{x}_n) \sum_{i=1}^{P} \mathbf{w}_{ij}(\mathbf{x}_n) \mathbf{w}_{ik}(\mathbf{x}_n) \\
&= \mathbf{f}_j^T \mathbb{E}[\sum_{i=1}^{P} \operatorname{diag}(\mathbf{w}_{ij} \bullet \mathbf{w}_{ij})] \mathbf{f}_j + 2\mathbf{f}_j^T \sum_{k \neq j} \sum_{i=1}^{P} \operatorname{diag}(\mathbf{w}_{ik} \bullet \mathbf{w}_{ij}) \mathbb{E}[\mathbf{f}_k].
\end{aligned} \tag{47}$$

Here we have used Equation (44) with $\mathbf{x} = \mathbf{f}_j$ and $\mathbf{D} = \mathbb{E}[\sum_{i=1}^{P} \operatorname{diag}(\mathbf{w}_{ij} \bullet \mathbf{w}_{ij})]$ and Equation (45) with $\mathbf{x} = \mathbf{f}_j$, $\mathbf{y} = \mathbf{f}_k$ and $\mathbf{D} = \mathbb{E}[\sum_{i=1}^{P} \operatorname{diag}(\mathbf{w}_{ik} \bullet \mathbf{w}_{ij})]$.

Detailed derivation for Equation 18:

$$\begin{aligned}
\mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \mathbf{y}_n^T \mathbf{W}_n \mathbf{f}_n &= \mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \sum_{i=1}^{P} \mathbf{y}_{ni} \sum_{j=1}^{Q} \mathbf{w}_{ij}(\mathbf{x}_n) \mathbf{f}_j(\mathbf{x}_n) \\
&= \mathbb{E}_{-\mathbf{f}_j} \sum_{n=1}^{N} \mathbf{f}_j(\mathbf{x}_n) \sum_{i=1}^{P} \mathbf{w}_{ij}(\mathbf{x}_n) \mathbf{y}_{ni} \\
&= \mathbb{E}_{-\mathbf{f}_j} \mathbf{f}_j^T \sum_{i=1}^{P} \mathbf{w}_{ij} \bullet \mathbf{Y}_{\cdot i} \\
&= \mathbf{f}_j^T \sum_{i=1}^{P} \mathbf{Y}_{\cdot i} \bullet \boldsymbol{\mu}_{\mathbf{w}_{ij}}
\end{aligned} \tag{48}$$