1. Key quantities in variational inference

Variational E-step:

\[
\frac{dF_1}{d\mu^u} = \sum_{u=1}^{U} \sum_{i=1}^{N^u} \left\{ -\Delta_i^u \left[ 2b_i^u K_{MM}^{-1} K_{NM,i}^u \right] \right\} + \mathbb{I}(t_i^u \in T^u) \left[ \tilde{G}' \left( -\frac{(b_i^u)^2}{2B_i^u} \right) \right],
\]

\[
\frac{dF_1}{d\Sigma^u} = \sum_{u=1}^{U} \sum_{i=1}^{N^u} \left\{ -\Delta_i^u \left[ K_{MM}^{-1} K_{NM,i}^u \right] \right\} + \mathbb{I}(t_i^u \in T^u) \left[ \left( -\frac{(b_i^u)^2}{2B_i^u} \right) \right].
\]

Variational M-step:

Updates for \(\mu_M^0\) can be derived in closed form as follows,

\[
\mu_M^0 = \frac{1}{\xi + U} \left( \xi g + \sum_{u=1}^{U} \mu^u \right).
\]

Gradient methods are needed for updating other parameters (denoted as \(\theta_k\)), including pseudo input positions and GP hyper-parameters, with key quantities summarized below:

\[
\frac{dF_1}{d\theta_k} = \sum_{u=1}^{U} \sum_{i=1}^{N^u} \left\{ \mathbb{I}(t_i^u \in T^u) \left[ -G' \left( -\frac{(b_i^u)^2}{2B_i^u} \right) \right] \right\} \times \left\{ 2b_i^u (\mu^u - g) \right\} \frac{dK_{MM}^{-1} K_{NM,i}^u}{d\theta_k},
\]

\[
\frac{dF_2}{d\mu^u} = -K_{MM}^{-1} (\mu^u - \mu_M^0),
\]

\[
\frac{dF_2}{d\Sigma^u} = -\frac{1}{2} K_{MM}^{-1} + \frac{1}{2} \Sigma^{-1},
\]

\[
\frac{dF_3}{d\mu^u} = 0,
\]

\[
\frac{dF_3}{d\Sigma^u} = 0.
\]

2. Hierarchical Gaussian process construction

Using the construction in (2) and (3), we can integrate over \(\mu_N^0\) to obtain the marginal prior distribution for \(f_N^a\) as

\[
f_N^a \sim \mathcal{N} \left( g, \left( 1 + \frac{1}{\xi} \right) K_{NN} \right).
\]
Using the construction in (9)-(11), we can integrate out $\mathbf{f}_M^u$ using (10) and (11), obtaining,

$$f_N^u | \mu_M^u \sim \mathcal{N} \left( g + \mathbf{K}_{NM}^{-1} \mathbf{K}_{M}^{-1} (\mu_M^0 - g), \right.$$

$$\left. \mathbf{K}_{NN}^{-1} + \frac{1}{\xi} \left( \mathbf{K}_{NN}^{-1} - \mathbf{K}_{NM}^{-1} \mathbf{K}_{MM}^{-1} \mathbf{K}_{NM} \right) \right).$$

Then, integrating out $\mu_M^u$ using (9), we can get the same marginal prior as in (S1),

$$f_N^u \sim \mathcal{N} \left(g, \left(1 + \frac{1}{\xi}\right) \mathbf{K}_{NN}^{-1}\right). \tag{S2}$$

The only difference between (S1) and (S2) is that we need a realization of the function at possible features from all tasks in (S1), while only the features from task $u$ need to be specified in (S2).

### 3. Confluent hypergeometric functions

When $|x|$ is small, for example $|x| \leq 30$, we use the power series to compute the confluent hypergeometric function:

$$\mathbf{1}_F(a, b, x) = \sum_{k=0}^{\infty} \frac{(a)_k x^k}{(b)_k k!},$$

where $(a)_0 = 1$, $(a)_k = \prod_{j=0}^{k-1} (a + j)$. In practice, the summation can be terminated at a sufficiently large number to achieve a given error tolerance level.

When $|x|$ is large, for example $|x| > 30$, we use the following computation (Thompson, 1997):

$$\mathbf{1}_F(a, b, x) = \frac{\Gamma(b)(-x)^{-a}}{\Gamma(b-1)} \sum_{k=0}^{\infty} \frac{(a)_k (a + a - b)_k}{k! (-x)^k}.$$

Having $\mathbf{1}_F(a, b, x)$, the gradient $\mathbf{G}(x) = G(0, \frac{1}{2}, z)$ can be numerically computed, where $G(a, b, x) = \frac{\partial \mathbf{1}_F(a, b, x)}{\partial a}$ (Ancarani & Gasaneo, 2008).

Finally, the expectation needed to compute $\mathbf{F}_1$ during variational E-step can be estimated as (Lloyd et al., 2014):

$$\mathbb{E}[\log(f_{N,i}^u)^2] = -\mathbf{G}(\frac{b_i^u}{2B_{ii}}) + \log(\frac{B_{ii}^u}{2}) - \text{const}.$$

### References

