
Disease Trajectory Maps: Supplementary Material

Peter Schulam

Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218
pschulam@cs.jhu.edu

Raman Arora

Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21218
arora@cs.jhu.edu

1 Derivation of Evidence Lower Bound

When marginalizing over the rows of F , we induced a Gaussian process over the trajectories, but by doing so we implicitly induced a Gaussian process over the individual-specific basis coefficients. Let $\mathbf{w}_i \triangleq F\mathbf{x}_i \in \mathbb{R}^d$ denote the basis weights implied by the mapping F and representation \mathbf{x}_i in the reduced-rank LMM, and let $\mathbf{w}_{:,k}$ for $k \in [d]$ denote the k^{th} coefficient of all individuals in the dataset. After marginalizing the k^{th} row of F and applying the kernel trick, we see that the vector of coefficients $\mathbf{w}_{:,k}$ has a Gaussian process distribution with mean 0 and covariance

$$\text{Cov}(w_{ik}, w_{jk}) = \alpha k(\mathbf{x}_i, \mathbf{x}_j). \quad (1)$$

Moreover, the Gaussian processes across coefficients are statistically independent of one another. To construct our approximate objective, we first approximate each of the d coefficient Gaussian processes by introducing p inducing points (see e.g. Snelson and Ghahramani [2005], Titsias [2009]) with values $\mathbf{u}_k \in \mathbb{R}^p$ for each $k \in [d]$ observed at common inputs $\mathbf{z}_i \in \mathbb{R}^q$ for $i \in [p]$. We assume that each $\mathbf{w}_{:,k}$ and \mathbf{u}_k are sampled from a common Gaussian process, which implies the joint distribution:

$$\mathbf{u}_k \mid \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, K_{pp}) \quad (2)$$

$$\mathbf{w}_k \mid \mathbf{u}_k, \boldsymbol{\theta} \sim \mathcal{N}(K_{mp}K_{pp}^{-1}\mathbf{u}_k, \tilde{K}_{mm}). \quad (3)$$

where K_{pp} is the Gram matrix between inducing points, K_{mm} is the Gram matrix between individuals (based on their representations \mathbf{x}_i), K_{mp} is the cross Gram matrix between individuals and inducing points, and $\tilde{K}_{mm} \triangleq K_{mm} - K_{mp}K_{pp}^{-1}K_{pm}$.

Now, we stack the inducing point values $\mathbf{u}_{1:d}$ into the columns of a matrix $U \triangleq [\mathbf{u}_1, \dots, \mathbf{u}_d]$. We will use \mathbf{u} to denote the “vectorization” of U obtained by stacking the columns. Each row i of U can be thought of as the vector of coefficients belonging to a single *inducing individual* which has an associated representation $\mathbf{z}_i \in \mathbb{R}^q$. Let $\mathbf{y} \triangleq [\mathbf{y}_1^\top, \dots, \mathbf{y}_m^\top]^\top$ be the vector of concatenated trajectories and W be the matrix containing individual i 's coefficients \mathbf{w}_i in each row, then following the derivation of Hensman et al. [2013], we can lower bound the conditional log-probability of \mathbf{y} given \mathbf{u} and $\mathbf{x}_{1:m}$:

$$\log p(\mathbf{y} \mid \mathbf{u}, \mathbf{x}_{1:m}) = \log \int p(\mathbf{y} \mid W)p(W \mid \mathbf{u}, \mathbf{x}_{1:m})dW \quad (4)$$

$$= \log \int \prod_{i=1}^m p(\mathbf{y}_i \mid \mathbf{w}_i)p(W \mid \mathbf{u}, \mathbf{x}_{1:m})dW \quad (5)$$

$$\geq \int p(W \mid \mathbf{u}, \mathbf{x}_{1:m}) \sum_{i=1}^m \log p(\mathbf{y}_i \mid \mathbf{w}_i)dW \quad (6)$$

$$= \sum_{i=1}^m \mathbb{E}_{p(\mathbf{w}_i \mid \mathbf{u}, \mathbf{x}_i)} [\log p(\mathbf{y}_i \mid \mathbf{w}_i)]. \quad (7)$$

The expectation in each summand is easy to calculate because the mean of \mathbf{y}_i is linearly dependent on \mathbf{w}_i and because the conditional distribution \mathbf{w}_i given \mathbf{u} is multivariate normal. Specifically, we have that

$$\mathbf{w}_i \mid \mathbf{u}, \mathbf{x}_i \sim \mathcal{N}(\mathbf{U}^\top \mathbf{K}_{pp}^{-1} \mathbf{k}_i, \tilde{k}_{ii} \mathbf{I}_d), \quad (8)$$

where \mathbf{k}_i is a column vector filled with the i^{th} row of \mathbf{K}_{mp} and \tilde{k}_{ii} is the i^{th} diagonal element of $\tilde{\mathbf{K}}_{mm}$. Together with the conditional distribution of \mathbf{y}_i given \mathbf{w}_i , we have that each summand can be written as

$$\mathbb{E}_{p(\mathbf{w}_i \mid \mathbf{u}, \mathbf{x}_i)}[\log p(\mathbf{y}_i \mid \mathbf{w}_i)] \quad (9)$$

$$= -\frac{n_i}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \mathbb{E}_{p(\mathbf{w}_i \mid \mathbf{u}, \mathbf{x}_i)}[(\mathbf{y}_i - \mu - \mathbf{B}_i \mathbf{w}_i)^\top (\mathbf{y}_i - \mu - \mathbf{B}_i \mathbf{w}_i)] \quad (10)$$

$$= \log \mathcal{N}(\mathbf{y}_i \mid \mu + \mathbf{B}_i \mathbf{U}^\top \mathbf{K}_{pp}^{-1} \mathbf{k}_i, \sigma^2 \mathbf{I}_{n_i}) - \frac{\tilde{k}_{ii}}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i] \quad (11)$$

$$\triangleq \log \tilde{p}(\mathbf{y}_i \mid \mathbf{u}, \mathbf{x}_i). \quad (12)$$

We can now write the lower bound on the conditional log-probability as

$$\log p(\mathbf{y} \mid \mathbf{u}, \mathbf{x}_{1:m}) \geq \sum_{i=1}^m \log \tilde{p}(\mathbf{y}_i \mid \mathbf{u}, \mathbf{x}_i) \triangleq \log \tilde{p}(\mathbf{y} \mid \mathbf{u}, \mathbf{x}_{1:m}). \quad (13)$$

To complete the derivation of the approximate objective, we use the lower bound on $\log p(\mathbf{y} \mid \mathbf{u}, \mathbf{x}_{1:m})$ to create a variational lower bound on the marginal log-probability of the trajectories

$$\log p(\mathbf{y}) = \log \int p(\mathbf{y} \mid \mathbf{u}, \mathbf{x}_{1:m}) p(\mathbf{u}, \mathbf{x}_{1:m}) d\mathbf{u} \quad (14)$$

$$\geq \int q(\mathbf{u}, \mathbf{x}_{1:m}) (\log p(\mathbf{y} \mid \mathbf{u}, \mathbf{x}_{1:m}) - \log q(\mathbf{u}, \mathbf{x}_{1:m}) + \log p(\mathbf{u}, \mathbf{x}_{1:m})) d\mathbf{u} d\mathbf{x}_{1:m} \quad (15)$$

$$\geq \int q(\mathbf{u}, \mathbf{x}_{1:m}) (\log \tilde{p}(\mathbf{y} \mid \mathbf{u}, \mathbf{x}_{1:m}) - \log q(\mathbf{u}, \mathbf{x}_{1:m}) + \log p(\mathbf{u}, \mathbf{x}_{1:m})) d\mathbf{u} d\mathbf{x}_{1:m} \quad (16)$$

$$\triangleq \log \tilde{p}(\mathbf{y}). \quad (17)$$

We assume that $\mathbf{u}, \mathbf{x}_1, \dots, \mathbf{x}_m$ are all mutually independent in the variational posterior. We use a multivariate normal variational approximation for each \mathbf{x}_i with variational parameters \mathbf{m}_i and \mathbf{S}_i .

Fixing \mathbf{x}_i , to find the optimal form for $q(\mathbf{u})$, note that each $\log \tilde{p}(\mathbf{y}_i \mid \mathbf{u}, \mathbf{x}_i)$ is composed of a log-likelihood plus an additive term that is independent of \mathbf{u} . Therefore, the terms that depend on \mathbf{u} can be written as:

$$\mathbb{E}_{q(\mathbf{u})} \left[\sum_{i=1}^m \log \mathcal{N}(\mathbf{y}_i \mid \mu + \mathbf{B}_i \mathbf{U}^\top \mathbf{K}_{pp}^{-1} \mathbf{k}_i, \sigma^2 \mathbf{I}_{n_i}) \right] - \text{KL}(q \parallel p). \quad (18)$$

Now, note that the mean in any of the log-likelihood terms can be rewritten as

$$\mu + \mathbf{B}_i \mathbf{U}^\top \mathbf{K}_{pp}^{-1} \mathbf{k}_i = \mu + (\mathbf{B}_i \otimes \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1}) \mathbf{u}, \quad (19)$$

Let $\mathbf{C}_i \triangleq (\mathbf{B}_i \otimes \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1})$ denote the *extended design matrix* obtained through this rewriting, and recall that each column \mathbf{u}_k is normally distributed with mean zero and covariance \mathbf{K}_{pp} . The prior over the vectorized matrix \mathbf{u} is therefore also multivariate normal. The expression above is maximized when $q(\mathbf{u})$ is equal to the posterior over \mathbf{u} given the observed trajectories. Because the prior is multivariate normal and the mean of the likelihood depends linearly on \mathbf{u} , the posterior must also be multivariate normal. Moreover, we know its exact form:

$$\mathbf{m}_* = \mathbf{S}_* \left(\sigma^{-2} \sum_{i=1}^m \mathbf{C}_i^\top (\mathbf{y}_i - \mu) \right), \quad \mathbf{S}_* = \left(\sigma^{-2} \sum_{i=1}^m \mathbf{C}_i^\top \mathbf{C}_i + (\mathbf{I}_d \otimes \mathbf{K}_{pp}^{-1}) \right)^{-1}. \quad (20)$$

We therefore parameterize $q(\mathbf{u})$ as a multivariate normal distribution with variational parameters \mathbf{m} and \mathbf{S} .

We now derive a closed-form expression for the expectation of $\log \tilde{p}(\mathbf{y}_i | \mathbf{u}, \mathbf{x}_i)$ under variational posterior distribution. Because \mathbf{u} and \mathbf{x}_i are assumed to be independent in the variational posteriors, we can analyze the expectation in either order. Fix \mathbf{x}_i , then we see that $\log \tilde{p}(\mathbf{y}_i | \mathbf{u}, \mathbf{x}_i)$ depends on \mathbf{u} only through the mean of the Gaussian density, which is a quadratic term in log likelihood. Because $q(\mathbf{u})$ is multivariate normal, we can compute the expectation in closed form.

$$\begin{aligned} \mathbb{E}_{q(\mathbf{u})}[\log \tilde{p}(\mathbf{y}_i | \mathbf{u}, \mathbf{x}_i)] &= \mathbb{E}_{q(\mathbf{u})}[\log \mathcal{N}(\mathbf{y}_i | \mu + (\mathbf{B}_i \otimes \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1})\mathbf{u}, \sigma^2 \mathbf{I}_{n_i})] - \frac{\tilde{k}_{ii}}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i] \\ &= \log \mathcal{N}(\mathbf{y}_i | \mu + \mathbf{C}_i \mathbf{m}, \sigma^2 \mathbf{I}_{n_i}) - \frac{1}{2\sigma^2} \text{Tr}[\mathbf{S} \mathbf{C}_i^\top \mathbf{C}_i] - \frac{\tilde{k}_{ii}}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i], \end{aligned}$$

We can compute the expectation of $\mathbb{E}_{q(\mathbf{u})}[\log \tilde{p}(\mathbf{y}_i | \mathbf{u}, \mathbf{x}_i)]$ in closed form by noting that we need only compute expectations of \mathbf{k}_i and $\mathbf{k}_i \mathbf{k}_i^\top$. Specifically, we have that

$$\mathbb{E}_{q(\mathbf{x}_i)}[k(\mathbf{x}_i, \mathbf{z}_j)] = \frac{\alpha}{|\mathbf{S}_i|^{1/2} |\mathbf{A}|^{1/2}} \exp \left\{ \frac{1}{2} (\mathbf{B}^\top \mathbf{A}^{-1} \mathbf{b} - c) \right\}, \quad (21)$$

where $\mathbf{A} = \mathbf{S}_i^{-1} + \ell^{-2} \mathbf{I}_q$, $\mathbf{b} = \mathbf{S}_i^{-1} \mathbf{m}_i + \ell^{-2} \mathbf{z}_j$, and $c = \mathbf{m}_i^\top \mathbf{S}_i^{-1} \mathbf{m}_i + \ell^{-2} \mathbf{z}_j^\top \mathbf{z}_j$. Similarly, for the expected outer product, we have

$$\mathbb{E}_{q(\mathbf{x}_i)}[k(\mathbf{x}_i, \mathbf{z}_j) k(\mathbf{x}_i, \mathbf{z}_k)] = \frac{\alpha}{|\mathbf{S}_i|^{1/2} |\mathbf{A}|^{1/2}} \exp \left\{ \frac{1}{2} (\mathbf{B}^\top \mathbf{A}^{-1} \mathbf{b} - c) \right\}, \quad (22)$$

where $\mathbf{A} = \mathbf{S}_i^{-1} + 2\ell^{-2} \mathbf{I}_q$, $\mathbf{b} = \mathbf{S}_i^{-1} \mathbf{m}_i + \ell^{-2} \mathbf{z}_j + \ell^{-2} \mathbf{z}_k$, and $c = \mathbf{m}_i^\top \mathbf{S}_i^{-1} \mathbf{m}_i + \ell^{-2} \mathbf{z}_j^\top \mathbf{z}_j + \ell^{-2} \mathbf{z}_k^\top \mathbf{z}_k$. Importantly, we can simply substitute these expectations into $\mathbb{E}_{q(\mathbf{u})}[\log \tilde{p}(\mathbf{y}_i | \mathbf{u}, \mathbf{x}_i)]$ and the form of the lower bound does not change (it is still a Gaussian log-likelihood plus the additional trace terms).

2 Optimizing the Evidence Lower Bound

To formulate the complete objective, we use the lower bound derived above and place priors on the observation noise σ^2 , and the hyperparameters of the kernel $k(\cdot, \cdot)$. In this section and in our experiments we assume that the kernel is a radial basis function (RBF) with scale α and length-scale (or bandwidth) ℓ . We assume normal distributions over the log of σ^2 , α , and ℓ with mean parameters m_s , m_a , m_ℓ respectively and precision parameters ρ_s , ρ_a , and ρ_ℓ respectively. Our objective is therefore

$$\mathcal{J}_{\text{SA-DTM}}(\mathbf{m}, \mathbf{S}, \mathbf{m}_{1:m}, \mathbf{S}_{1:m}, \mu, \sigma^2, \alpha, \ell) = \quad (23)$$

$$\sum_{i=1}^m -\frac{n_i}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \mathbb{E}_{q(\mathbf{x}_i)}[\|\mathbf{y}_i - \mu - (\mathbf{B}_i \otimes \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1})\mathbf{m}\|_2^2] \quad (24)$$

$$+ \sum_{i=1}^m -\frac{1}{2\sigma^2} \text{Tr}[\mathbf{S}(\mathbf{B}_i^\top \mathbf{B}_i \otimes \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{k}_i^\top] \mathbf{K}_{pp}^{-1})] \quad (25)$$

$$+ \sum_{i=1}^m -\frac{1}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i] (\alpha - \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} \mathbf{k}_i]) \quad (26)$$

$$- \sum_{i=1}^m \frac{1}{2} (\text{Tr}[\mathbf{S}_i + \mathbf{m}_i \mathbf{m}_i^\top] - q - \log |\mathbf{S}_i|) \quad (27)$$

$$- \frac{1}{2} \left(\text{Tr}[(\mathbf{S} + \mathbf{m} \mathbf{m}^\top)(\mathbf{I}_d \otimes \mathbf{K}_{pp}^{-1})] - pd + \log \frac{|\mathbf{K}_{pp}|^d}{|\mathbf{S}|} \right) \quad (28)$$

$$- \frac{\rho_s}{2} \|\log \sigma^2 - m_s\|_2^2 - \frac{\rho_a}{2} \|\log \alpha - m_a\|_2^2 - \frac{\rho_\ell}{2} \|\log \ell - m_\ell\|_2^2. \quad (29)$$

Note that the last three lines above can be seen as regularizers (log priors for the hyperparameters and a KL divergence between the variational distribution q and the prior p). The first four lines can be decomposed across individuals, suggesting that we can use stochastic approximation of the objective and its gradients to derive a scalable algorithm for optimizing the objective.

We define an iterative first-order optimization algorithm. In broad strokes, within each iteration we will sample a single individual i (or a batch of patients), maximize the objective with respect to \mathbf{m}_i

and S_i while holding the global variables fixed, compute the approximate gradients of the objective, and take a small step in the direction of each gradient for each parameter (the step size is determined by a learning schedule, which may be specific to each global variable). We discuss each step in detail below. We do so assuming a single sampled individual i , although in principle we can sample a batch of individuals to reduce variance in the gradient estimate.

Maximizing wrt local variables (\mathbf{m}_i, S_i). Before computing gradients of the approximate objective with respect to the global parameters, we first do a block coordinate optimization over the local variational parameters of individual i . We optimize:

$$J_i(\mathbf{m}_i, S_i) = \quad (30)$$

$$- \frac{n_i}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \mathbb{E}_{q(\mathbf{x}_i)} [\|\mathbf{y}_i - \mu - (\mathbf{B}_i \otimes \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1}) \mathbf{m}\|_2^2] \quad (31)$$

$$- \frac{1}{2\sigma^2} \text{Tr}[S(\mathbf{B}_i^\top \mathbf{B}_i \otimes \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{k}_i^\top] \mathbf{K}_{pp}^{-1})] \quad (32)$$

$$- \frac{1}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i](\alpha - \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} \mathbf{k}_i]). \quad (33)$$

We can optimize this expression using a gradient-based optimizer. We use the scaled conjugate gradients algorithm.

Estimating gradients of global variables. Having sampled individual i and having refit her local variational parameters, we now want to approximate the gradient of the full objective with respect to the global variables \mathbf{m} , S , μ , σ^2 , α , and ℓ . We first look at the approximate gradient with respect to \mathbf{m} .

$$\hat{\nabla}_{\mathcal{J}_{SA\text{-DTM}}}(\mathbf{m}) = \mathbb{E}_{q(\mathbf{x}_i)} \left[\frac{m}{\sigma^2} (\mathbf{B}_i^\top \otimes \mathbf{K}_{pp}^{-1} \mathbf{k}_i) (\mathbf{y}_i - \mu - (\mathbf{B} \otimes \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1}) \mathbf{m}) \right] - (\mathbf{I}_d \otimes \mathbf{K}_{pp}^{-1}) \mathbf{m}. \quad (34)$$

The approximate gradient with respect to S is

$$\hat{\nabla}_{\mathcal{J}_{SA\text{-DTM}}}(S) = - \frac{m}{2\sigma^2} \text{Tr}[(\mathbf{B}_i^\top \mathbf{B}_i \otimes \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{k}_i^\top] \mathbf{K}_{pp}^{-1})] \quad (35)$$

$$- \frac{1}{2} \text{Tr}[(\mathbf{I}_d \otimes \mathbf{K}_{pp}^{-1})] + \frac{1}{2} \text{Tr}[S^{-1}]. \quad (36)$$

Note that if we set these approximate gradients to 0, we obtain the following estimates of \mathbf{m} and S :

$$\hat{\mathbf{m}} = \hat{S} \left(\frac{m}{\sigma^2} (\mathbf{B}_i^\top \otimes \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i]) (\mathbf{y} - \mu) \right) \quad (37)$$

$$\hat{S} = \left(\frac{m}{\sigma^2} (\mathbf{B}_i^\top \mathbf{B}_i \otimes \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{k}_i^\top] \mathbf{K}_{pp}^{-1}) + (\mathbf{I}_d \otimes \mathbf{K}_{pp}^{-1}) \right)^{-1} \quad (38)$$

We can improve the rate of convergence of our algorithm by taking the geometry of the space of distributions parameterized by \mathbf{m} and S into account. We do so by using the *natural gradients* for these two parameters instead of the approximations above. Let θ_1 and θ_2 denote the canonical parameterization of the variational multivariate normal, then the gradient updates at time t are Hoffman et al. [2013]:

$$\theta_1^t = \theta_1^{t-1} + \lambda_t (\eta_1^{t-1} - \theta_1^{t-1}) \quad (39)$$

$$\theta_2^t = \theta_2^{t-1} + \lambda_t (\eta_2^{t-1} - \theta_2^{t-1}), \quad (40)$$

where

$$\eta_1^{t-1} = \frac{m}{\sigma^2} (\mathbf{B}_i^\top \otimes \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i]) (\mathbf{y} - \mu) \quad (41)$$

$$\eta_2^{t-1} = - \frac{m}{2\sigma^2} (\mathbf{B}_i^\top \mathbf{B}_i \otimes \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{k}_i^\top] \mathbf{K}_{pp}^{-1}) \quad (42)$$

To update the hyperparameters, we need to compute the gradients with respect to μ , σ^2 , α , and ℓ . We parameterize σ^2 , α , and ℓ using their logarithms, and so present gradients with respect to that representation. To make the expressions more clear, we present the gradients as differentials with

respect to the kernel, which can be completed using the chain rule. The estimate of the gradient with respect to μ is

$$\hat{\nabla}_{\mathcal{J}_{\text{SA-DTM}}}(\mu) = \frac{m}{\sigma^2}(\mathbf{y}_i - \mu - (\mathbf{B}_i \otimes \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i^\top] \mathbf{K}_{pp}^{-1}) \mathbf{m})^\top \mathbf{1}_{n_i}. \quad (43)$$

The estimate of the gradient with respect to $\log \sigma^2$ is

$$\hat{\nabla}_{\mathcal{J}_{\text{SA-DTM}}}(\log \sigma^2) = -\frac{mn_i}{2} + \frac{m}{2\sigma^2} \mathbb{E}_{q(\mathbf{x}_i)}[\|\mathbf{y}_i - \mu - (\mathbf{B}_i \otimes \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1}) \mathbf{m}\|_2^2] \quad (44)$$

$$+ \frac{m}{2\sigma^2} \text{Tr}[\mathbf{S}(\mathbf{B}_i^\top \mathbf{B}_i \otimes \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{x}_i^\top] \mathbf{K}_{pp}^{-1})] \quad (45)$$

$$+ \frac{m}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i](\alpha - \text{Tr}[\mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{x}_i^\top]]) \quad (46)$$

$$- \rho_s(\log \sigma^2 - m_s). \quad (47)$$

The estimate of the gradient with respect to $\log \alpha$ is

$$\hat{\nabla}_{\text{SA-DTM}}(\log \alpha) = \quad (48)$$

$$\frac{m}{\sigma^2} \mathbb{E}_{q(\mathbf{x}_i)}[(\mathbf{y}_i - \mu - \mathbf{Cm})^\top (\mathbf{B}_i \otimes \partial \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} - \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp} \mathbf{K}_{pp}^{-1}) \mathbf{m}] \quad (49)$$

$$- \frac{m}{\sigma^2} \mathbb{E}_{q(\mathbf{x}_i)}[\text{Tr}[\mathbf{S} \mathbf{C}_i^\top (\mathbf{B}_i \otimes \partial \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} - \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp} \mathbf{K}_{pp}^{-1})]] \quad (50)$$

$$- \frac{m}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i] \alpha \quad (51)$$

$$+ \frac{m}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i](2\mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i^\top] \mathbf{K}_{pp}^{-1} \partial \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i] - \text{Tr}[\mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp} \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{x}_i^\top]]) \quad (52)$$

$$+ \frac{1}{2} (\text{Tr}[(\mathbf{S} + \mathbf{m} \mathbf{m}^\top)(\mathbf{I}_d \otimes \mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp} \mathbf{K}_{pp}^{-1})] - d \text{Tr}[\mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp}]). \quad (53)$$

The estimate of the gradient with respect to $\log \ell$ is

$$\hat{\nabla}_{\text{SA-DTM}}(\log \ell) = \quad (54)$$

$$\frac{m}{\sigma^2} \mathbb{E}_{q(\mathbf{x}_i)}[(\mathbf{y}_i - \mu - \mathbf{Cm})^\top (\mathbf{B}_i \otimes \partial \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} - \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp} \mathbf{K}_{pp}^{-1}) \mathbf{m}] \quad (55)$$

$$- \frac{m}{\sigma^2} \mathbb{E}_{q(\mathbf{x}_i)}[\text{Tr}[\mathbf{S} \mathbf{C}_i^\top (\mathbf{B}_i \otimes \partial \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} - \mathbf{k}_i^\top \mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp} \mathbf{K}_{pp}^{-1})]] \quad (56)$$

$$+ \frac{m}{2\sigma^2} \text{Tr}[\mathbf{B}_i^\top \mathbf{B}_i](2\mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i^\top] \mathbf{K}_{pp}^{-1} \partial \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i] - \text{Tr}[\mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp} \mathbf{K}_{pp}^{-1} \mathbb{E}_{q(\mathbf{x}_i)}[\mathbf{k}_i \mathbf{x}_i^\top]]) \quad (57)$$

$$+ \frac{1}{2} (\text{Tr}[(\mathbf{S} + \mathbf{m} \mathbf{m}^\top)(\mathbf{I}_d \otimes \mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp} \mathbf{K}_{pp}^{-1})] - d \text{Tr}[\mathbf{K}_{pp}^{-1} \partial \mathbf{K}_{pp}]). \quad (58)$$

References

- J. Hensman, N. Fusi, and N.D. Lawrence. Gaussian processes for big data. *arXiv:1309.6835*, 2013.
- M.D. Hoffman, D.M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 14(1):1303–1347, 2013.
- E. Snelson and Z. Ghahramani. Sparse gaussian processes using pseudo-inputs. In *NIPS*, 2005.
- M.K. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *AISTATS*, 2009.