Derivations for Generalized Gaussian Process Models

Antoni B. Chan Daxiang Dong
Department of Computer Science
City University of Hong Kong
abchan@cityu.edu.hk, daxidong@cityu.edu.hk

Abstract

This is the supplemental material for the CVPR 2011 paper, "Generalized Gaussian Process Models" [1]. It contains derivations for the Taylor and Laplace approximations, and a summary of EP.

I. Closed-form Taylor Approximation

In this section, we derive the closed-form Taylor approximation proposed in Section 5.1 of the paper [1].

A. Joint likelihood approximation

The joint likelihood of data and parameters is,

$$\log p(\mathbf{y}, \boldsymbol{\eta}|\mathbf{X}) = \log p(\mathbf{y}|\boldsymbol{\theta}(\boldsymbol{\eta})) + \log p(\boldsymbol{\eta}|\mathbf{X}) \tag{S.1}$$

$$= \sum_{i=1}^{n} \left(\frac{1}{a(\phi)} \left[y_i \theta(\eta_i) - b(\theta(\eta_i)) \right] + \log h(y_i, \phi) \right) - \frac{1}{2} \boldsymbol{\eta}^T \mathbf{K}^{-1} \boldsymbol{\eta} - \frac{1}{2} \log |\mathbf{K}| - \frac{n}{2} \log 2\pi \tag{S.2}$$

Note that the terms that prevents the posterior of $\boldsymbol{\eta}$ from being Gaussian are the functions $b(\theta(\eta_i))$ and $\theta(\eta_i)$. Let us consider a second-order Taylor expansion of the likelihood term at the point $\tilde{\eta}_i$,

$$\log p(y_i|\theta(\eta_i)) = \frac{1}{a(\phi)} \left[y_i \theta(\eta_i) - b(\theta(\eta_i)) \right] + \log h(y_i, \phi) \tag{S.3}$$

$$\approx \log p(y_i|\theta(\tilde{\eta}_i)) + \tilde{u}_i(\eta_i - \tilde{\eta}_i) - \frac{1}{2} \tilde{w}_i^{-1} (\eta_i - \tilde{\eta}_i)^2 \tag{S.4}$$

where $\tilde{u}_i = u(\tilde{\eta}_i, y_i)$ and $\tilde{w}_i = w(\tilde{\eta}_i, y_i)$ as defined in (19). Summing over (S.4), we have the approximation

$$\sum_{i=1}^{n} \log p(y_i|\theta(\eta_i)) \approx \sum_{i=1}^{n} \log p(y_i|\theta(\tilde{\eta}_i)) + \tilde{u}_i(\eta_i - \tilde{\eta}_i) - \frac{1}{2} \tilde{w}_i^{-1} (\eta_i - \tilde{\eta}_i)^2 \tag{S.5}$$

$$= -\frac{1}{2} (\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}})^T \tilde{\mathbf{W}}^{-1} (\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}) + \tilde{\mathbf{u}}^T (\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}) + \log p(\mathbf{y}|\theta(\tilde{\boldsymbol{\eta}})) \tag{S.6}$$

where $\tilde{\mathbf{u}}$ is the vector of $\tilde{u}_i$, and $\tilde{\mathbf{W}} = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$. Substituting the Taylor approximation into (S.2), we obtain an approximation to the joint posterior,

$$\begin{aligned}\log q(\mathbf{y}, \boldsymbol{\eta}|\mathbf{X}) &= \log p(\mathbf{y}|\theta(\tilde{\boldsymbol{\eta}})) - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \\ &\quad - \frac{1}{2}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}})^T\tilde{\mathbf{W}}^{-1}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}) + \tilde{\mathbf{u}}^T(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}) - \frac{1}{2}\boldsymbol{\eta}^T\mathbf{K}^{-1}\boldsymbol{\eta} \qquad \text{(S.7)} \\ &= \log p(\mathbf{y}|\theta(\tilde{\boldsymbol{\eta}})) - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \\ &\quad - \frac{1}{2}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}} - \tilde{\mathbf{W}}\tilde{\mathbf{u}})^T\tilde{\mathbf{W}}^{-1}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}} - \tilde{\mathbf{W}}\tilde{\mathbf{u}}) - \frac{1}{2}\boldsymbol{\eta}^T\mathbf{K}^{-1}\boldsymbol{\eta} + \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}} \qquad \text{(S.8)}\end{aligned}$$

Next, we note that the first two terms of the second line are of the form

$$\begin{aligned}&(\mathbf{x} - \mathbf{a})^T\mathbf{B}(\mathbf{x} - \mathbf{a}) + \mathbf{x}^T\mathbf{C}\mathbf{x} = \mathbf{x}^T\mathbf{D}\mathbf{x} - 2\mathbf{x}^T\mathbf{B}\mathbf{a} + \mathbf{a}^T\mathbf{B}\mathbf{a} && \text{(S.9)} \\ &= \mathbf{x}^T\mathbf{D}\mathbf{x} - 2\mathbf{x}^T\mathbf{D}\mathbf{D}^{-1}\mathbf{B}\mathbf{a} + \mathbf{a}^T\mathbf{B}^T\mathbf{D}^{-1}\mathbf{B}\mathbf{a} - \mathbf{a}^T\mathbf{B}^T\mathbf{D}^{-1}\mathbf{B}\mathbf{a} + \mathbf{a}^T\mathbf{B}\mathbf{a} && \text{(S.10)} \\ &= \left\|\mathbf{x} - \mathbf{D}^{-1}\mathbf{B}\mathbf{a}\right\|^2_{\mathbf{D}^{-1}} + \mathbf{a}^T(\mathbf{B} - \mathbf{B}^T\mathbf{D}^{-1}\mathbf{B})\mathbf{a} && \text{(S.11)} \\ &= \left\|\mathbf{x} - \mathbf{D}^{-1}\mathbf{B}\mathbf{a}\right\|^2_{\mathbf{D}^{-1}} + \|\mathbf{a}\|^2_{\mathbf{B}^{-1}+\mathbf{C}^{-1}} && \text{(S.12)}\end{aligned}$$

where $\mathbf{D} = \mathbf{B} + \mathbf{C}$, and the last line uses the matrix inversion lemma. Using this property ($\mathbf{x} = \boldsymbol{\eta}$, $\mathbf{a} = \tilde{\boldsymbol{\eta}} + \tilde{\mathbf{W}}\tilde{\mathbf{u}}$, $\mathbf{B} = \mathbf{W}^{-1}$, $\mathbf{C} = \mathbf{K}^{-1}$), the joint likelihood can be rewritten as

$$\begin{aligned}\log q(\mathbf{y}, \boldsymbol{\eta}|\mathbf{X}) &= \log p(\mathbf{y}|\theta(\tilde{\boldsymbol{\eta}})) - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \\ &\quad - \frac{1}{2}\left\|\boldsymbol{\eta} - \mathbf{A}^{-1}\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{t}}\right\|^2_{\mathbf{A}^{-1}} - \frac{1}{2}\left\|\tilde{\mathbf{t}}\right\|^2_{\tilde{\mathbf{W}}+\mathbf{K}} + \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}} \qquad \text{(S.13)}\end{aligned}$$

where $\mathbf{A} = \tilde{\mathbf{W}}^{-1} + \mathbf{K}^{-1}$, and $\tilde{\mathbf{t}} = \tilde{\boldsymbol{\eta}} + \tilde{\mathbf{W}}\tilde{\mathbf{u}}$ is the target vector.

B. Approximate Posterior

Removing terms in (S.13) that are not dependent on $\boldsymbol{\eta}$, the approximate posterior of $\boldsymbol{\eta}$ is

$$\begin{aligned}&\log q(\boldsymbol{\eta}|\mathbf{X}, \mathbf{y}) \propto \log q(\mathbf{y}, \boldsymbol{\eta}|\mathbf{X}) \propto -\frac{1}{2}\left\|\boldsymbol{\eta} - \mathbf{A}^{-1}\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{t}}\right\|^2_{\mathbf{A}^{-1}} && \text{(S.14)} \\ \Rightarrow \quad &q(\boldsymbol{\eta}|\mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\boldsymbol{\eta}\middle|\mathbf{A}^{-1}\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{t}}, \mathbf{A}^{-1}\right), && \text{(S.15)}\end{aligned}$$

and hence

$$\hat{\mathbf{m}} = \hat{\mathbf{V}}\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{t}}, \quad \hat{\mathbf{V}} = \left(\tilde{\mathbf{W}}^{-1} + \mathbf{K}^{-1}\right)^{-1}. \qquad \text{(S.16)}$$

C. Approximate Marginal

The approximate marginal is obtained by substituting the approximate joint in (S.13)

$$\log p(\mathbf{y}|\mathbf{X}) = \log \int \exp(\log p(\mathbf{y}, \boldsymbol{\eta}|\mathbf{X}))d\boldsymbol{\eta} \approx \log \int \exp(\log q(\mathbf{y}, \boldsymbol{\eta}|\mathbf{X}))d\boldsymbol{\eta} \tag{S.17}$$

$$\begin{aligned} &= \log p(\mathbf{y}|\theta(\tilde{\boldsymbol{\eta}})) - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \\ &\quad - \frac{1}{2}\left\|\tilde{\mathbf{t}}\right\|^2_{\tilde{\mathbf{W}}+\mathbf{K}} + \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}} + \log \int e^{-\frac{1}{2}\left\|\boldsymbol{\eta}-\mathbf{A}^{-1}\tilde{\mathbf{W}}^{-1}\tilde{\mathbf{t}}\right\|^2_{\mathbf{A}^{-1}}} d\boldsymbol{\eta} \end{aligned} \tag{S.18}$$

$$\begin{aligned} &= \log p(\mathbf{y}|\theta(\tilde{\boldsymbol{\eta}})) - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \\ &\quad - \frac{1}{2}\left\|\tilde{\mathbf{t}}\right\|^2_{\tilde{\mathbf{W}}+\mathbf{K}} + \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}} + \log(2\pi)^{\frac{n}{2}}\left|\mathbf{A}^{-1}\right|^{\frac{1}{2}} \end{aligned} \tag{S.19}$$

$$= \log p(\mathbf{y}|\theta(\tilde{\boldsymbol{\eta}})) - \frac{1}{2}\log|\mathbf{K}|\,|\mathbf{A}| - \frac{1}{2}\left\|\tilde{\mathbf{t}}\right\|^2_{\tilde{\mathbf{W}}+\mathbf{K}} + \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}} \tag{S.20}$$

Looking at the determinant term,

$$\log|\mathbf{A}|\,|\mathbf{K}| = \log\left|(\tilde{\mathbf{W}}^{-1} + \mathbf{K}^{-1})\mathbf{K}\right| = \log\left|\mathbf{I} + \tilde{\mathbf{W}}^{-1}\mathbf{K}\right| = \log\left|\tilde{\mathbf{W}} + \mathbf{K}\right|\left|\tilde{\mathbf{W}}^{-1}\right|. \tag{S.21}$$

Hence, the approximate marginal is

$$\log q(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\tilde{\mathbf{t}}^T(\tilde{\mathbf{W}} + \mathbf{K})^{-1}\tilde{\mathbf{t}} - \frac{1}{2}\log\left|\tilde{\mathbf{W}} + \mathbf{K}\right| + r(\phi) \tag{S.22}$$

where

$$r(\phi) = \log p(\mathbf{y}|\theta(\tilde{\boldsymbol{\eta}})) + \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}} + \frac{1}{2}\log\left|\tilde{\mathbf{W}}\right|. \tag{S.23}$$

For an individual data point, we have

$$r_i(\phi) = \log p(y_i|\theta(\tilde{\eta}_i)) + \frac{1}{2}\tilde{w}_i\tilde{u}_i^2 + \frac{1}{2}\log|\tilde{w}_i|, \tag{S.24}$$

and hence

$$r(\phi) = \sum_{i=1}^{n} r_i(\phi). \tag{S.25}$$

1) Derivatives wrt hyperparameters: The derivative of (S.22) with respect to the kernel hyperparameter α_j is

$$\frac{\partial}{\partial\alpha_j}\log q(\mathbf{y}|\mathbf{X}) = \frac{1}{2}\tilde{\mathbf{t}}^T\left(\mathbf{K} + \tilde{\mathbf{W}}\right)^{-1}\frac{\partial\mathbf{K}}{\partial\alpha_j}\left(\mathbf{K} + \tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{t}} - \frac{1}{2}\mathrm{tr}\left[(\mathbf{K} + \tilde{\mathbf{W}})^{-1}\frac{\partial\mathbf{K}}{\partial\alpha_j}\right] \tag{S.26}$$

$$= \frac{1}{2}\mathrm{tr}\left[\left(\mathbf{z}\mathbf{z}^T - (\mathbf{K} + \tilde{\mathbf{W}})^{-1}\right)\frac{\partial\mathbf{K}}{\partial\alpha_j}\right], \quad \mathbf{z} = (\mathbf{K} + \tilde{\mathbf{W}})^{-1}\tilde{\mathbf{t}}. \tag{S.27}$$

where $\frac{\partial\mathbf{K}}{\partial\alpha_j}$ is the element-wise derivative of the kernel matrix with respect to the kernel hyperparameter α_j, and we use the derivative properties,

$$\frac{\partial}{\partial\alpha}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\alpha}\mathbf{A}^{-1}, \qquad \frac{\partial}{\partial\alpha}\log|\mathbf{A}| = \mathrm{tr}(\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\alpha}). \tag{S.28}$$

2) Derivative wrt dispersion: For the derivative with respect to the dispersion parameter, we first note that

$$\frac{\partial}{\partial\phi}\tilde{w}_i = a'(\phi)\left\{b''(\theta(\eta))\theta'(\eta)^2 - [y - b'(\theta(\eta))]\,\theta''(\eta)\right\}^{-1} = \frac{a'(\phi)}{a(\phi)}\tilde{w}_i, \tag{S.29}$$

$$\frac{\partial}{\partial\phi}\tilde{\mathbf{W}} = \frac{a'(\phi)}{a(\phi)}\tilde{\mathbf{W}}, \tag{S.30}$$

$$\frac{\partial}{\partial\phi}\tilde{u}_i = -\frac{a'(\phi)}{a(\phi)^2}\theta'(\eta)\,[y - b'(\theta(\eta))] = -\frac{a'(\phi)}{a(\phi)}\tilde{u}_i, \tag{S.31}$$

$$\frac{\partial}{\partial\phi}\tilde{w}_i\tilde{u}_i^2 = \frac{a'(\phi)}{a(\phi)}\tilde{w}_i\tilde{u}_i^2 + 2\tilde{w}_i\tilde{u}_i\left(-\frac{a'(\phi)}{a(\phi)}\tilde{u}_i\right) = -\frac{a'(\phi)}{a(\phi)}\tilde{w}_i\tilde{u}_i^2 \tag{S.32}$$

$$\frac{\partial}{\partial\phi}\log p(y_i|\theta(\tilde{\eta}_i)) = \frac{\partial}{\partial\phi}\left\{\frac{1}{a(\phi)}\left[y_i\theta(\tilde{\eta}_i) - b(\theta(\tilde{\eta}_i))\right] + \log h(y_i,\phi)\right\} \tag{S.33}$$

$$= -\frac{a'(\phi)}{a(\phi)^2}\left[y_i\theta(\tilde{\eta}_i) - b(\theta(\tilde{\eta}_i))\right] + \frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.34}$$

$$= -\frac{a'(\phi)}{a(\phi)}\tilde{v}_i + \frac{\partial}{\partial\phi}\log h(y_i,\phi), \tag{S.35}$$

where $\tilde{v}_i = \frac{1}{a(\phi)}\left[y_i\theta(\tilde{\eta}_i) - b(\theta(\tilde{\eta}_i))\right]$. Thus,

$$\frac{\partial}{\partial\phi}r_i(\phi) = -\frac{a'(\phi)}{a(\phi)}\tilde{v}_i + \frac{\partial}{\partial\phi}\log h(y_i,\phi) - \frac{1}{2}\frac{a'(\phi)}{a(\phi)}\tilde{w}_i\tilde{u}_i^2 + \frac{1}{2}\frac{1}{\tilde{w}_i}\frac{a'(\phi)}{a(\phi)}\tilde{w}_i \tag{S.36}$$

$$= \frac{a'(\phi)}{a(\phi)}\left(\frac{1}{2} - \tilde{v}_i - \frac{1}{2}\tilde{w}_i\tilde{u}_i^2\right) + \frac{\partial}{\partial\phi}\log h(y_i,\phi). \tag{S.37}$$

Summing over i,

$$\frac{\partial}{\partial\phi}r(\phi) = \sum_i \frac{a'(\phi)}{a(\phi)}\left(\frac{1}{2} - \tilde{v}_i - \frac{1}{2}\tilde{w}_i\tilde{u}_i^2\right) + \frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.38}$$

$$= \frac{a'(\phi)}{a(\phi)}\left(\frac{n}{2} - \mathbf{1}^T\tilde{\mathbf{v}} - \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}}\right) + \sum_{i=1}^{n}\frac{\partial}{\partial\phi}\log h(y_i,\phi). \tag{S.39}$$

Also note that $\tilde{\mathbf{t}}$ is not a function of ϕ, as the term cancels out in $\tilde{\mathbf{W}}\tilde{\mathbf{u}}$. Finally,

$$\frac{\partial}{\partial\phi}\log q(\mathbf{y}|\mathbf{X}) \tag{S.40}$$

$$= \frac{1}{2}\tilde{\mathbf{t}}^T\left(\mathbf{K} + \tilde{\mathbf{W}}\right)^{-1}\frac{\partial\tilde{\mathbf{W}}}{\partial\phi}\left(\mathbf{K} + \tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{t}} - \frac{1}{2}\mathrm{tr}((\mathbf{K} + \tilde{\mathbf{W}})^{-1}\frac{\partial\tilde{\mathbf{W}}}{\partial\phi}) + \frac{\partial}{\partial\phi}r(\phi)$$

$$= \frac{1}{2}\mathrm{tr}\left[\left(\mathbf{z}\mathbf{z}^T - (\mathbf{K} + \tilde{\mathbf{W}})^{-1}\right)\frac{\partial\tilde{\mathbf{W}}}{\partial\phi}\right] + \frac{a'(\phi)}{a(\phi)}\left(\frac{n}{2} - \mathbf{1}^T\tilde{\mathbf{v}} - \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}}\right) + \sum_{i=1}^{n}\frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.41}$$

$$= \frac{a'(\phi)}{a(\phi)}\left\{\frac{1}{2}\mathrm{tr}\left[\left(\mathbf{z}\mathbf{z}^T - (\mathbf{K} + \tilde{\mathbf{W}})^{-1}\right)\tilde{\mathbf{W}}\right] + \frac{n}{2} - \mathbf{1}^T\tilde{\mathbf{v}} - \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}}\right\} + \sum_{i=1}^{n}\frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.42}$$

$$= \frac{a'(\phi)}{a(\phi)}\left\{\frac{1}{2}\mathbf{z}^T\tilde{\mathbf{W}}\mathbf{z} - \frac{1}{2}\mathrm{tr}\left[(\mathbf{K} + \tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}\right] + \frac{n}{2} - \mathbf{1}^T\tilde{\mathbf{v}} - \frac{1}{2}\tilde{\mathbf{u}}^T\tilde{\mathbf{W}}\tilde{\mathbf{u}}\right\} + \sum_{i=1}^{n}\frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.43}$$

3) Derivative wrt dispersion when $b_\phi(\theta)$: We next look at the special case where the term $b_\phi(\theta)$ is also a function of ϕ. We first note that

$$\begin{aligned}\frac{\partial}{\partial\phi}\tilde{w}_i &= a'(\phi)\left\{b''_\phi(\theta(\eta))\theta'(\eta)^2 - \left[y - b'_\phi(\theta(\eta))\right]\theta''(\eta)\right\}^{-1} \\ &\quad - a(\phi)\frac{\theta'(\eta)^2\frac{\partial}{\partial\phi}b''_\phi(\theta(\eta)) + \theta''(\eta)\frac{\partial}{\partial\phi}b'_\phi(\theta(\eta))}{\left\{b''_\phi(\theta(\eta))\theta'(\eta)^2 - \left[y - b'_\phi(\theta(\eta))\right]\theta''(\eta)\right\}^2} \qquad \text{(S.44)}\\ &= \frac{a'(\phi)}{a(\phi)}\tilde{w}_i - \frac{1}{a(\phi)}\tilde{w}_i^2\left[\theta'(\eta)^2\frac{\partial}{\partial\phi}b''_\phi(\theta(\eta)) + \theta''(\eta)\frac{\partial}{\partial\phi}b'_\phi(\theta(\eta))\right], \qquad \text{(S.45)}\\ \frac{\partial}{\partial\phi}\tilde{u}_i &= -\frac{a'(\phi)}{a(\phi)^2}\theta'(\eta)\left[y - b'_\phi(\theta(\eta))\right] - \frac{\theta'(\eta)}{a(\phi)}\frac{\partial}{\partial\phi}b'_\phi(\theta(\eta)) \qquad \text{(S.46)}\\ &= -\frac{a'(\phi)}{a(\phi)}\tilde{u}_i - \frac{\theta'(\eta)}{a(\phi)}\frac{\partial}{\partial\phi}b'_\phi(\theta(\eta)), \qquad \text{(S.47)}\\ \frac{\partial}{\partial\phi}(\tilde{w}_i\tilde{u}_i^2) &= \frac{\partial\tilde{w}_i}{\partial\phi}\tilde{u}_i^2 + 2\tilde{w}_i\tilde{u}_i\frac{\partial\tilde{u}_i}{\partial\phi}, \qquad \text{(S.48)}\\ \frac{\partial}{\partial\phi}\tilde{t}_i &= \frac{\partial}{\partial\phi}(\tilde{\eta}_i + \tilde{w}_i\tilde{u}_i) = \tilde{w}_i\frac{\partial\tilde{u}_i}{\partial\phi} + \tilde{u}_i\frac{\partial\tilde{w}_i}{\partial\phi}, \qquad \text{(S.49)}\\ \frac{\partial}{\partial\phi}\log p(y_i|\theta(\tilde{\eta}_i)) &= \frac{\partial}{\partial\phi}\left\{\frac{1}{a(\phi)}\left[y_i\theta(\tilde{\eta}_i) - b_\phi(\theta(\tilde{\eta}_i))\right] + \log h(y_i,\phi)\right\} \qquad \text{(S.50)}\\ &= -\frac{a'(\phi)}{a(\phi)^2}\left[y_i\theta(\tilde{\eta}_i) - b_\phi(\theta(\tilde{\eta}_i))\right] - \frac{1}{a(\phi)}\frac{\partial}{\partial\phi}b_\phi(\theta(\tilde{\eta}_i)) + \frac{\partial}{\partial\phi}\log h(y_i,\phi) \qquad \text{(S.51)}\\ &= -\frac{a'(\phi)}{a(\phi)}\tilde{v}_i + \frac{\partial}{\partial\phi}\log h(y_i,\phi) - \frac{1}{a(\phi)}\frac{\partial}{\partial\phi}b_\phi(\theta(\tilde{\eta}_i)). \qquad \text{(S.52)}\end{aligned}$$

Thus,

$$\frac{\partial}{\partial\phi}r_i(\phi) = \frac{\partial}{\partial\phi}\log p(y_i|\theta(\tilde{\eta}_i)) + \frac{1}{2}\left[\frac{\partial\tilde{w}_i}{\partial\phi}\tilde{u}_i^2 + 2\tilde{w}_i\tilde{u}_i\frac{\partial\tilde{u}_i}{\partial\phi}\right] + \frac{1}{2}\frac{1}{\tilde{w}_i}\frac{\partial\tilde{w}_i}{\partial\phi} \qquad \text{(S.53)}$$

For the first term in (S.22),

$$\begin{aligned}\frac{\partial}{\partial\phi}\left[\tilde{\mathbf{t}}^T(\tilde{\mathbf{W}} + \mathbf{K})^{-1}\tilde{\mathbf{t}}\right] &= \tilde{\mathbf{t}}^T\frac{\partial(\tilde{\mathbf{W}} + \mathbf{K})^{-1}}{\partial\phi}\tilde{\mathbf{t}} + 2\tilde{\mathbf{t}}^T(\tilde{\mathbf{W}} + \mathbf{K})^{-1}\frac{\partial\tilde{\mathbf{t}}}{\partial\phi} \qquad \text{(S.54)}\\ &= -\tilde{\mathbf{t}}^T\left(\mathbf{K} + \tilde{\mathbf{W}}\right)^{-1}\frac{\partial\tilde{\mathbf{W}}}{\partial\phi}\left(\mathbf{K} + \tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{t}} + 2\tilde{\mathbf{t}}^T(\tilde{\mathbf{W}} + \mathbf{K})^{-1}\frac{\partial\tilde{\mathbf{t}}}{\partial\phi}. \qquad \text{(S.55)}\end{aligned}$$

For the second term in (S.22),

$$\frac{\partial}{\partial\phi}\log\left|\mathbf{K} + \tilde{\mathbf{W}}\right| = \mathrm{tr}((\mathbf{K} + \tilde{\mathbf{W}})^{-1}\frac{\partial\tilde{\mathbf{W}}}{\partial\phi}). \qquad \text{(S.56)}$$

Finally, we have

$$\begin{aligned}\frac{\partial}{\partial\phi}\log q(\mathbf{y}|\mathbf{X}) &= \frac{1}{2}\tilde{\mathbf{t}}^T\left(\mathbf{K} + \tilde{\mathbf{W}}\right)^{-1}\frac{\partial\tilde{\mathbf{W}}}{\partial\phi}\left(\mathbf{K} + \tilde{\mathbf{W}}\right)^{-1}\tilde{\mathbf{t}} - \tilde{\mathbf{t}}^T(\tilde{\mathbf{W}} + \mathbf{K})^{-1}\frac{\partial\tilde{\mathbf{t}}}{\partial\phi} \\ &\quad - \frac{1}{2}\mathrm{tr}((\mathbf{K} + \tilde{\mathbf{W}})^{-1}\frac{\partial\tilde{\mathbf{W}}}{\partial\phi}) + \sum_i\frac{\partial}{\partial\phi}r_i(\phi).\end{aligned} \qquad \text{(S.57)}$$

II. Laplace Approximation

In this section, we present more details of the Laplace approximation for GGPMs, which is discussed in Section 5.2 of the paper. The Laplace's approximation is a Gaussian approximation of the posterior $p(\boldsymbol{\eta}|\mathbf{X},\mathbf{y})$ using a 2nd-order Taylor expansion at its maximum (mode). Hence, the Laplace approximation is a special case of the closed-form Taylor approximation in the previous section, where the expansion points $\tilde{\eta}_i$ are selected so that $\hat{\mathbf{m}}$ is the maximum of the true posterior $p(\boldsymbol{\eta}|\mathbf{X},\mathbf{y})$.

A. Approximate Posterior

The maximum of the posterior is given by

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \log p(\boldsymbol{\eta}|\mathbf{X},\mathbf{y}) = \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \log p(\mathbf{y}|\boldsymbol{\theta}(\boldsymbol{\eta})) + \log p(\boldsymbol{\eta}|\mathbf{X}). \tag{S.58}$$

In vector notation, the derivatives of the posterior are

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\boldsymbol{\eta}|\mathbf{X},\mathbf{y}) = \mathbf{u} - \mathbf{K}^{-1}\boldsymbol{\eta} \tag{S.59}$$

$$\frac{\partial^2}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T} \log p(\boldsymbol{\eta}|\mathbf{X},\mathbf{y}) = -\mathbf{W}^{-1} - \mathbf{K}^{-1} \tag{S.60}$$

At the maximum of the posterior, we have the condition

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\boldsymbol{\eta}|\mathbf{X},\mathbf{y}) = \hat{\mathbf{u}} - \mathbf{K}^{-1}\boldsymbol{\eta} = \mathbf{0} \tag{S.61}$$

$$\Rightarrow \quad \hat{\mathbf{u}} = \mathbf{K}^{-1}\hat{\boldsymbol{\eta}}, \tag{S.62}$$

where $\hat{\mathbf{u}}$ is $\mathbf{u}$ evaluated at $\hat{\boldsymbol{\eta}}$. The maximum can be obtained iteratively using the Newton-Raphson method. In particular, each iteration is

$$\hat{\boldsymbol{\eta}}^{(new)} = \hat{\boldsymbol{\eta}} - \left(\frac{\partial}{\partial \boldsymbol{\eta}\boldsymbol{\eta}^T} \log p(\hat{\boldsymbol{\eta}}|\mathbf{X},\mathbf{y})\right)^{-1} \left(\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\hat{\boldsymbol{\eta}}|\mathbf{X},\mathbf{y})\right) \tag{S.63}$$

$$= \hat{\boldsymbol{\eta}} + \left(\hat{\mathbf{W}}^{-1} + \mathbf{K}^{-1}\right)^{-1} (\hat{\mathbf{u}} - \mathbf{K}^{-1}\hat{\boldsymbol{\eta}}) \tag{S.64}$$

$$= \left(\hat{\mathbf{W}}^{-1} + \mathbf{K}^{-1}\right)^{-1} (\hat{\mathbf{u}} + \hat{\mathbf{W}}^{-1}\hat{\boldsymbol{\eta}}) \tag{S.65}$$

$$= \left(\hat{\mathbf{W}}^{-1} + \mathbf{K}^{-1}\right)^{-1} \hat{\mathbf{W}}^{-1}(\hat{\mathbf{W}}\hat{\mathbf{u}} + \hat{\boldsymbol{\eta}}), \tag{S.66}$$

where $\hat{\mathbf{W}}$ is $\mathbf{W}$ evaluated at $\hat{\boldsymbol{\eta}}$. Note that this is exactly the same form as the closed-form Taylor expansion in the previous section. In particular, at each iteration the expansion point $\hat{\boldsymbol{\eta}}$ is moved closer to the maximum. As a result, the target vector $\hat{\mathbf{t}}^{(new)} = \hat{\mathbf{W}}\hat{\mathbf{u}} + \hat{\boldsymbol{\eta}}$ is also updated.

B. Approximate Marginal

The marginal likelihood is approximated by applying a Taylor approximation to the marginal likelihood at the mode of the posterior, $\hat{\boldsymbol{\eta}}$. The marginal distribution is

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\theta(\boldsymbol{\eta}))p(\boldsymbol{\eta}|\mathbf{X})d\boldsymbol{\eta} = \int e^{\kappa(\boldsymbol{\eta})} d\boldsymbol{\eta} \tag{S.67}$$

where

$$\kappa(\boldsymbol{\eta}) = \log p(\mathbf{y}|\theta(\boldsymbol{\eta})) + \log p(\boldsymbol{\eta}|\mathbf{X}). \tag{S.68}$$

The second-order Taylor approximation of $\kappa(\boldsymbol{\eta})$ around $\hat{\boldsymbol{\eta}}$ is

$$\kappa(\boldsymbol{\eta}) \approx \kappa(\hat{\boldsymbol{\eta}}) - \frac{1}{2}\left\|\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}\right\|^2_{\hat{\mathbf{V}}} \tag{S.69}$$

$$= \log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}})) + \log p(\hat{\boldsymbol{\eta}}|\mathbf{X}) - \frac{1}{2}\left\|\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}\right\|^2_{\hat{\mathbf{V}}} \tag{S.70}$$

Substituting into (S.67), we obtain the approximate marginal distribution

$$\log q(\mathbf{y}|\mathbf{X}) = \log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}})) + \log p(\hat{\boldsymbol{\eta}}|\mathbf{X}) + \int e^{-\frac{1}{2}\|\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}\|^2_{\hat{\mathbf{V}}}} d\boldsymbol{\eta} \tag{S.71}$$

$$= \log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}})) - \frac{1}{2}\hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\hat{\boldsymbol{\eta}} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi + \frac{n}{2}\log 2\pi + \frac{1}{2}\log\left|\hat{\mathbf{V}}\right| \tag{S.72}$$

$$= \log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}})) - \frac{1}{2}\hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\hat{\boldsymbol{\eta}} - \frac{1}{2}\log\left|\left(\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1}\right)\mathbf{K}\right| \tag{S.73}$$

$$= \log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}})) - \frac{1}{2}\hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\hat{\boldsymbol{\eta}} - \frac{1}{2}\log\left|\hat{\mathbf{W}}^{-1}\mathbf{K}+\mathbf{I}\right| \tag{S.74}$$

Note that $\hat{\boldsymbol{\eta}}$ is dependent on the kernel matrix $\mathbf{K}$ and $a(\phi)$. Hence, at each iteration during optimization of $q(\mathbf{y}|\mathbf{X})$, we have to recompute its value.

1) Derivative wrt hyperparameters: We next look at the derivatives of each term in (S.74). Note that $\hat{\boldsymbol{\eta}}$ is a function of the kernel hyperparameters. Using the chain rule, the derivative of the first term is

$$\frac{\partial}{\partial\alpha_j}\log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}})) = \frac{\partial}{\partial\hat{\boldsymbol{\eta}}^T}\log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}}))\frac{\partial\hat{\boldsymbol{\eta}}}{\partial\alpha_j} = \hat{\mathbf{u}}^T\frac{\partial\hat{\boldsymbol{\eta}}}{\partial\alpha_j}. \tag{S.75}$$

For the second term, we use the product rule, $\frac{\partial\mathbf{AB}}{\partial\alpha} = \mathbf{A}\frac{\partial\mathbf{B}}{\partial\alpha} + \frac{\partial\mathbf{A}}{\partial\alpha}\mathbf{B}$,

$$\frac{\partial}{\partial\alpha_j}\hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\hat{\boldsymbol{\eta}} = \frac{\partial}{\partial\alpha_j}\hat{\boldsymbol{\eta}}^T\left(\mathbf{K}^{-1}\hat{\boldsymbol{\eta}}\right) \tag{S.76}$$

$$= \hat{\boldsymbol{\eta}}^T\frac{\partial}{\partial\alpha_j}\left(\mathbf{K}^{-1}\hat{\boldsymbol{\eta}}\right) + \frac{\partial\hat{\boldsymbol{\eta}}^T}{\partial\alpha_j}\left(\mathbf{K}^{-1}\hat{\boldsymbol{\eta}}\right) \tag{S.77}$$

$$= \hat{\boldsymbol{\eta}}^T\left(\mathbf{K}^{-1}\frac{\partial\hat{\boldsymbol{\eta}}}{\partial\alpha_j} + \frac{\partial\mathbf{K}^{-1}}{\partial\alpha_j}\hat{\boldsymbol{\eta}}\right) + \frac{\partial\hat{\boldsymbol{\eta}}^T}{\partial\alpha_j}\left(\mathbf{K}^{-1}\hat{\boldsymbol{\eta}}\right) \tag{S.78}$$

$$= -\hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\frac{\partial\mathbf{K}}{\partial\alpha_j}\mathbf{K}^{-1}\hat{\boldsymbol{\eta}} + 2\hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\frac{\partial\hat{\boldsymbol{\eta}}}{\partial\alpha_j}. \tag{S.79}$$

For the third term,

$$\frac{\partial}{\partial\alpha_j}\log\left|\hat{\mathbf{W}}^{-1}\mathbf{K}+\mathbf{I}\right| = \mathrm{tr}\left[(\hat{\mathbf{W}}^{-1}\mathbf{K}+\mathbf{I})^{-1}\frac{\partial}{\partial\alpha_j}(\hat{\mathbf{W}}^{-1}\mathbf{K})\right] \tag{S.80}$$

$$= \mathrm{tr}\left[(\hat{\mathbf{W}}^{-1}\mathbf{K}+\mathbf{I})^{-1}(\hat{\mathbf{W}}^{-1}\frac{\partial\mathbf{K}}{\partial\alpha_j} + \frac{\partial\hat{\mathbf{W}}^{-1}}{\partial\alpha_j}\mathbf{K})\right] \tag{S.81}$$

$$= \mathrm{tr}\left[(\mathbf{K}+\hat{\mathbf{W}})^{-1}\frac{\partial\mathbf{K}}{\partial\alpha_j}\right] + \mathrm{tr}\left[(\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1})^{-1}\frac{\partial\hat{\mathbf{W}}^{-1}}{\partial\alpha_j}\right] \tag{S.82}$$

Putting it all together, we have

$$\frac{\partial}{\partial \alpha_j}\log q(\mathbf{y}|\mathbf{X}) = \hat{\mathbf{u}}^T\frac{\partial \hat{\boldsymbol{\eta}}}{\partial \alpha_j} + \frac{1}{2}\hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \alpha_j}\mathbf{K}^{-1}\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\frac{\partial \hat{\boldsymbol{\eta}}}{\partial \alpha_j} - \frac{1}{2}\mathrm{tr}\left[(\mathbf{K}+\hat{\mathbf{W}})^{-1}\frac{\partial \mathbf{K}}{\partial \alpha_j}\right] - \frac{1}{2}\mathrm{tr}\left[(\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1})^{-1}\frac{\partial \hat{\mathbf{W}}^{-1}}{\partial \alpha_j}\right] \tag{S.83}$$

$$= \frac{1}{2}\mathrm{tr}\left[\left(\hat{\mathbf{u}}\hat{\mathbf{u}}^T - (\mathbf{K}+\hat{\mathbf{W}})^{-1}\right)\frac{\partial \mathbf{K}}{\partial \alpha_j}\right] - \frac{1}{2}\mathrm{tr}\left[(\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1})^{-1}\frac{\partial \hat{\mathbf{W}}^{-1}}{\partial \alpha_j}\right], \tag{S.84}$$

which follows from $\hat{\mathbf{u}} = \mathbf{K}^{-1}\hat{\boldsymbol{\eta}}$.

Next, we note that $\hat{\mathbf{W}}$ is a function of $\hat{\boldsymbol{\eta}}$, which in turn is a function of the kernel hyperparameter α_j, and thus

$$\frac{\partial \hat{\mathbf{W}}^{-1}}{\partial \alpha_j} = \sum_{i=1}^{n}\frac{\partial \hat{\mathbf{W}}^{-1}}{\partial \hat{\eta}_i}\frac{\partial \hat{\eta}_i}{\partial \alpha_j} \tag{S.85}$$

The first term is

$$\frac{\partial \hat{\mathbf{W}}^{-1}}{\partial \hat{\eta}_i} = \left(-\frac{\partial^3}{\partial \hat{\eta}_i^3}\log p(y_i|\theta(\hat{\eta}_i))\right)\mathbf{e}_i\mathbf{e}_i^T. \tag{S.86}$$

Elements of the second term are given by,

$$\frac{\partial \hat{\boldsymbol{\eta}}}{\partial \alpha_j} = \frac{\partial}{\partial \alpha_j}(\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1})^{-1}\hat{\mathbf{W}}^{-1}\hat{\mathbf{t}} \tag{S.87}$$

$$= -(\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1})^{-1}\frac{\partial \mathbf{K}^{-1}}{\partial \alpha_j}(\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1})^{-1}\hat{\mathbf{W}}^{-1}\hat{\mathbf{t}} \tag{S.88}$$

$$= (\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1})^{-1}\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \alpha_j}\mathbf{K}^{-1}(\hat{\mathbf{W}}^{-1}+\mathbf{K}^{-1})^{-1}\hat{\mathbf{W}}^{-1}\hat{\mathbf{t}} \tag{S.89}$$

$$= \hat{\mathbf{W}}(\hat{\mathbf{W}}+\mathbf{K})^{-1}\frac{\partial \mathbf{K}}{\partial \alpha_j}\hat{\mathbf{u}}. \tag{S.90}$$

Hence, (S.85) can be expressed as

$$\frac{\partial \mathbf{W}^{-1}}{\partial \alpha_j} = \mathrm{diag}\left(\mathbf{A}\frac{\partial \hat{\boldsymbol{\eta}}}{\partial \alpha_j}\right), \tag{S.91}$$

where $\mathbf{A}$ is a diagonal matrix of with entries

$$\mathbf{A} = \mathrm{diag}\left(-\frac{\partial^3}{\partial \hat{\eta}_i^3}\log p(y_i|\theta(\hat{\eta}_i))\right) \tag{S.92}$$

and

$$-\frac{\partial^3}{\partial \hat{\eta}_i^3}\log p(y_i|\theta(\hat{\eta}_i)) = \frac{\partial}{\partial \hat{\eta}_i}\left[\frac{1}{a(\phi)}\left\{\theta'(\hat{\eta}_i)[g^{-1}]'(\hat{\eta}_i) - \theta''(\hat{\eta}_i)\left(y - g^{-1}(\hat{\eta}_i)\right)\right\}\right] \tag{S.93}$$

$$= \frac{1}{a(\phi)}\left\{\theta'(\hat{\eta}_i)[g^{-1}]''(\hat{\eta}_i) + \theta''(\hat{\eta}_i)[g^{-1}]'(\hat{\eta}_i) + \theta''(\hat{\eta}_i)[g^{-1}]'(\hat{\eta}_i) - \theta'''(\hat{\eta}_i)\left(y - g^{-1}(\hat{\eta}_i)\right)\right\} \tag{S.94}$$

$$= \frac{1}{a(\phi)}\left\{\theta'(\hat{\eta}_i)[g^{-1}]''(\hat{\eta}_i) + 2\theta''(\hat{\eta}_i)[g^{-1}]'(\hat{\eta}_i) - \theta'''(\hat{\eta}_i)\left(y - g^{-1}(\hat{\eta}_i)\right)\right\}. \tag{S.95}$$

2) Derivative wrt dispersion: Looking at the derivatives of each term in (S.74), for the first term,

$$\frac{\partial}{\partial\phi}\log p(y_i|\theta(\hat{\eta}_i)) = \frac{\partial}{\partial\phi}\left\{\frac{1}{a(\phi)}\left[y_i\theta(\hat{\eta}_i) - b(\theta(\hat{\eta}_i))\right] + \log h(y_i,\phi)\right\} \tag{S.96}$$

$$= -\frac{a'(\phi)}{a(\phi)^2}\left[y_i\theta(\hat{\eta}_i) - b(\theta(\hat{\eta}_i))\right] + \frac{1}{a(\phi)}\frac{\partial}{\partial\phi}\left[y_i\theta(\hat{\eta}_i) - b(\theta(\hat{\eta}_i))\right] + \frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.97}$$

$$= -\frac{a'(\phi)}{a(\phi)}\hat{v}_i + \frac{1}{a(\phi)}\frac{\partial}{\partial\hat{\eta}_i}\left[y_i\theta(\hat{\eta}_i) - b(\theta(\hat{\eta}_i))\right]\frac{\partial\hat{\eta}_i}{\partial\phi} + \frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.98}$$

$$= -\frac{a'(\phi)}{a(\phi)}\hat{v}_i + \hat{u}_i\frac{\partial\hat{\eta}_i}{\partial\phi} + \frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.99}$$

where $\hat{v}_i = v(\hat{\eta}_i, y_i)$ and $\hat{u}_i = u(\hat{\eta}_i, y_i)$. Hence,

$$\frac{\partial}{\partial\phi}\log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}})) = \sum_{i=1}^{n} -\frac{a'(\phi)}{a(\phi)}\hat{v}_i + \hat{u}_i\frac{\partial\hat{\eta}_i}{\partial\phi} + \frac{\partial}{\partial\phi}\log h(y_i,\phi) \tag{S.100}$$

$$= -\frac{a'(\phi)}{a(\phi)}\mathbf{1}^T\hat{\mathbf{v}} + \hat{\mathbf{u}}^T\frac{\partial\hat{\boldsymbol{\eta}}}{\partial\phi} + \sum_{i=1}^{n}\frac{\partial}{\partial\phi}\log h(y_i,\phi). \tag{S.101}$$

The derivative of $\hat{\boldsymbol{\eta}}$ is

$$\frac{\partial\hat{\boldsymbol{\eta}}}{\partial\phi} = \frac{\partial}{\partial\phi}(\hat{\mathbf{W}}^{-1} + \mathbf{K}^{-1})^{-1}\hat{\mathbf{W}}^{-1}\hat{\mathbf{t}} = \frac{\partial}{\partial\phi}(\mathbf{I} + \hat{\mathbf{W}}\mathbf{K}^{-1})^{-1}\hat{\mathbf{t}} \tag{S.102}$$

$$= -(\mathbf{I} + \hat{\mathbf{W}}\mathbf{K}^{-1})^{-1}\frac{\partial}{\partial\phi}(\hat{\mathbf{W}}\mathbf{K}^{-1})(\mathbf{I} + \hat{\mathbf{W}}\mathbf{K}^{-1})^{-1}\hat{\mathbf{t}} \tag{S.103}$$

$$= -\frac{a'(\phi)}{a(\phi)}(\mathbf{I} + \hat{\mathbf{W}}\mathbf{K}^{-1})^{-1}\hat{\mathbf{W}}\mathbf{K}^{-1}(\mathbf{I} + \hat{\mathbf{W}}\mathbf{K}^{-1})^{-1}\hat{\mathbf{t}} \tag{S.104}$$

$$= -\frac{a'(\phi)}{a(\phi)}(\hat{\mathbf{W}}^{-1} + \mathbf{K}^{-1})^{-1}(\mathbf{K} + \hat{\mathbf{W}})^{-1}\hat{\mathbf{t}}, \tag{S.105}$$

$$= -\frac{a'(\phi)}{a(\phi)}(\hat{\mathbf{W}}^{-1} + \mathbf{K}^{-1})^{-1}\hat{\mathbf{u}}. \tag{S.106}$$

Hence,

$$\frac{\partial}{\partial\phi}\log p(\mathbf{y}|\theta(\hat{\boldsymbol{\eta}})) = -\frac{a'(\phi)}{a(\phi)}\left[\mathbf{1}^T\hat{\mathbf{v}} + \hat{\mathbf{u}}^T(\mathbf{W}^{-1} + \mathbf{K}^{-1})^{-1}\hat{\mathbf{u}}\right] + \sum_{i=1}^{n}\frac{\partial}{\partial\phi}\log h(y_i,\phi). \tag{S.107}$$

For the second term,

$$\frac{\partial}{\partial\phi}\frac{1}{2}\hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}^T\mathbf{K}^{-1}\frac{\partial\hat{\boldsymbol{\eta}}}{\partial\phi} = \hat{\mathbf{u}}^T\frac{\partial\hat{\boldsymbol{\eta}}}{\partial\phi}, \tag{S.108}$$

$$= -\frac{a'(\phi)}{a(\phi)}\hat{\mathbf{u}}^T(\mathbf{W}^{-1} + \mathbf{K}^{-1})^{-1}\hat{\mathbf{u}}. \tag{S.109}$$

For the third term,

$$\frac{\partial}{\partial\phi}\log\left|\hat{\mathbf{W}}^{-1}\mathbf{K}+\mathbf{I}\right| = \mathrm{tr}\left[(\hat{\mathbf{W}}^{-1}\mathbf{K}+\mathbf{I})^{-1}\frac{\partial}{\partial\phi}(\hat{\mathbf{W}}^{-1}\mathbf{K})\right] \tag{S.110}$$

$$= -\mathrm{tr}\left[(\hat{\mathbf{W}}^{-1}\mathbf{K}+\mathbf{I})^{-1}\hat{\mathbf{W}}^{-1}\frac{\partial\hat{\mathbf{W}}}{\partial\phi}\hat{\mathbf{W}}^{-1}\mathbf{K})\right] \tag{S.111}$$

$$= -\frac{a'(\phi)}{a(\phi)}\mathrm{tr}\left[(\hat{\mathbf{W}}^{-1}\mathbf{K}+\mathbf{I})^{-1}\hat{\mathbf{W}}^{-1}\mathbf{K})\right] \tag{S.112}$$

$$= -\frac{a'(\phi)}{a(\phi)}\mathrm{tr}\left[(\mathbf{K}+\hat{\mathbf{W}})^{-1}\mathbf{K})\right]. \tag{S.113}$$

Finally, putting it all together,

$$\frac{\partial}{\partial\phi}\log q(\mathbf{y}|\mathbf{X}) = \frac{a'(\phi)}{a(\phi)}\left\{-\mathbf{1}^T\hat{\mathbf{v}}+\frac{1}{2}\mathrm{tr}\left[(\mathbf{K}+\hat{\mathbf{W}})^{-1}\mathbf{K})\right]\right\}+\sum_{i=1}^{n}\frac{\partial}{\partial\phi}\log h(y_i,\phi). \tag{S.114}$$

III. Expectation Propagation

In this section, we present more details on expectation propagation (EP) for for GGPMs, which is discussed in Section 5.3 of the paper.

A. Computing the site parameters

Instead of computing the optimal site parameters all at once, EP works by iteratively updating each individual site using the other site approximations [2], [3]. In particular, assume that we are updating site t_i. We first compute the *cavity distribution*, which is the marginalization over all sites except t_i,

$$q_{\neg i}(\eta_i) \propto \int p(\boldsymbol{\eta}|\mathbf{X})\prod_{j\neq i}t_j(\eta_j|\tilde{Z}_j,\tilde{\mu}_j,\tilde{\sigma}_j^2)d\eta_j, \tag{S.115}$$

where the notation $\neg i$ indicates the sites without t_i. $q_{\neg i}(\eta_i)$ is an approximation to the posterior distribution of η_i, given all observations except y_i. Since both terms are Gaussian, this integral can be computed in closed-form, leading to the cavity distribution

$$q_{\neg i}(\eta_i) = \mathcal{N}\left(\eta_i\middle|\mu_{\neg i},\sigma_{\neg i}^2\right), \tag{S.116}$$

$$\mu_{\neg i} = \sigma_{\neg i}^2(\sigma_i^{-2}\mu_i-\tilde{\sigma}_i^{-2}\tilde{\mu}_i), \quad \sigma_{\neg i}^2 = (\sigma_i^{-2}-\tilde{\sigma}_i^{-2})^{-1} \tag{S.117}$$

where $\sigma_i^2 = [\hat{\mathbf{V}}]_{ii}$ and $\mu_i = [\hat{\mathbf{m}}]_i$.

Next, multiplying the cavity distribution by the *true* data likelihood of y_i gives an approximation to the unnormalized posterior of η_i, given the observations,

$$q(\eta_i|\mathbf{X},\mathbf{y}) = q_{\neg i}(\eta_i)p(y_i|\theta(\eta_i)). \tag{S.118}$$

Note that this incorporates the true likelihood of y_i and the approximate likelihoods of the remaining observations. On the other hand, the approximation to the posterior of η_i using the site function is

$$\hat{q}(\eta_i) = q_{\neg i}(\eta_i)t_i(\eta_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_i^2) \tag{S.119}$$

Hence, the new parameters for site t_i can be computed by minimizing the KL divergence between (S.118) and (S.119), i.e., between the approximate posterior using the true likeli-

hood and that using the site t_i,

$$\{\tilde{Z}_i^*, \tilde{\mu}_i^*, \tilde{\sigma}_i^{*2}\} = \underset{\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2}{\operatorname{argmin}} \mathrm{KL}\left(q(\eta_i|\mathbf{X}, \mathbf{y}) \,\|\hat{q}(\eta_i)\right) \tag{S.120}$$

$$= \underset{\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2}{\operatorname{argmin}} \mathrm{KL}\left(q_{\neg i}(\eta_i) p(y_i|\theta(\eta_i)) \,\middle\|\, q_{\neg i}(\eta_i) t_i(\eta_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)\right) \tag{S.121}$$

Note that $\hat{q}(\eta_i)$ in (S.119) is an unnormalized Gaussian, i.e. $\hat{q}(\eta_i) = \hat{Z}_i \mathcal{N}\left(\eta_i|\hat{\mu}_i, \hat{\sigma}_i^2\right)$, with parameters (given by the product of Gaussians)

$$\hat{\mu}_i = \hat{\sigma}_i^2(\sigma_{\neg i}^{-2}\mu_{\neg i} + \tilde{\sigma}_i^{-2}\tilde{\mu}_i), \tag{S.122}$$

$$\hat{\sigma}_i^2 = (\sigma_{\neg i}^{-2} + \tilde{\sigma}_i^{-2})^{-1}, \tag{S.123}$$

$$\hat{Z}_i = \tilde{Z}_i(2\pi)^{-\frac{1}{2}}(\tilde{\sigma}_i^2 + \sigma_{\neg i}^2)^{-\frac{1}{2}} \exp\left(-\frac{(\mu_{\neg i} - \tilde{\mu}_i)^2}{2(\sigma_{\neg i}^2 + \tilde{\sigma}_i^2)}\right). \tag{S.124}$$

Hence, to compute (S.121), it suffices to first find the parameters of the unnormalized Gaussian that minimizes the KL divergence to $q(\eta_i|\mathbf{X}, \mathbf{y})$, and then "subtract" $q_{\neg i}(\eta_i)$ from this Gaussian. To find $\hat{q}(\eta_i)$ we minimize the KL divergence,

$$\{\hat{Z}_i^*, \hat{\mu}_i^*, \hat{\sigma}_i^{*2}\} = \underset{\hat{Z}_i, \hat{\mu}_i, \hat{\sigma}_i^2}{\operatorname{argmin}} \mathrm{KL}\left(q_{\neg i}(\eta_i) p(y_i|\theta(\eta_i)) \,\middle\|\, \hat{Z}_i \mathcal{N}\left(\eta_i\middle|\hat{\mu}_i, \hat{\sigma}_i^2\right)\right). \tag{S.125}$$

In particular, it is well known that the KL divergence in (S.125) is minimized when the moments of the Gaussian match those of the first argument. In addition to the mean and variance (1st and 2nd moments), we also must match the normalization constant (0th moment), since $\hat{q}(\eta_i)$ is unnormalized. The optimal parameters are

$$\hat{\mu}_i = \mathbb{E}_q[\eta_i] = \frac{1}{\hat{Z}_i}\int \eta_i q_{\neg i}(\eta_i) p(y_i|\theta(\eta_i)) d\eta_i \tag{S.126}$$

$$\hat{\sigma}_i^2 = \mathrm{var}_q(\eta_i) = \frac{1}{\hat{Z}_i}\int (\eta_i - \hat{\mu}_i)^2 q_{\neg i}(\eta_i) p(y_i|\theta(\eta_i)) d\eta_i \tag{S.127}$$

$$\hat{Z}_i = Z_q = \int q_{\neg i}(\eta_i) p(y_i|\theta(\eta_i)) d\eta_i, \tag{S.128}$$

where $\mathbb{E}_q$ and var_q are the expectation and variance with respect to $q(\eta_i|\mathbf{X}, \mathbf{y})$. These moments are discussed later in this section. Finally, the new parameters for site t_i are obtained by "subtracting" the two Gaussians, $q_{\neg i}(\theta_i)$ from $\hat{q}(\theta_i)$, leading to the site updates

$$\tilde{\mu}_i = \tilde{\sigma}_i^2(\hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma_{\neg i}^{-2}\mu_{\neg i}) \tag{S.129}$$

$$\tilde{\sigma}_i^2 = (\hat{\sigma}_i^{-2} - \sigma_{\neg i}^{-2})^{-1} \tag{S.130}$$

$$\tilde{Z}_i = \hat{Z}_i\sqrt{2\pi(\sigma_{\neg i}^2 + \tilde{\sigma}_i^2)} \exp\left(\frac{(\mu_{\neg i} - \tilde{\mu}_i)^2}{2(\sigma_{\neg i}^2 + \tilde{\sigma}_i^2)}\right). \tag{S.131}$$

EP iterates over each of the site t_i, i.e. each observation y_i, iteratively until convergence.

1) Computing the moments: Each EP iteration requires the moments of $q(\eta_i)$ in (S.126-S.128), where

$$q(\eta_i) = \frac{1}{\hat{Z}_i} p(y_i|\theta(\eta_i)) \mathcal{N}\left(\eta_i\middle|\mu_{\neg i}, \sigma_{\neg i}^2\right), \tag{S.132}$$

$$\hat{Z}_i = \int p(y_i|\theta(\eta_i)) \mathcal{N}\left(\eta_i\middle|\mu_{\neg i}, \sigma_{\neg i}^2\right) d\eta_i. \tag{S.133}$$

Depending on the form of the likelihood $p(y_i|\theta(\eta_i))$ the integrals may not be analytically tractable. Hence, approximate integration is required. In this paper, we use numerical integration with 10,000 equally spaced points between $\mu_{\neg i} \pm 5\sigma_{\neg i}$. We also tried Gaussian-Hermite quadrature to approximate the integral, but this caused EP to fail to converge in many cases, due to inaccuracy in the approximation.

B. Marginal Likelihood

The EP approximation to the marginal log-likelihood is

$$\log p(\mathbf{y}|\mathbf{X}) \approx \log q(\mathbf{y}|\mathbf{X}) = \log Z_{EP} \tag{S.134}$$

$$= \log \int q(y|\theta(\boldsymbol{\eta}))p(\boldsymbol{\eta}|\mathbf{X})d\boldsymbol{\eta} \tag{S.135}$$

$$= \log \int \mathcal{N}\left(\boldsymbol{\eta}\middle|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right) \prod_{i=1}^{n} \tilde{Z}_i \mathcal{N}\left(\boldsymbol{\eta}|\mathbf{0}, \mathbf{K}\right) d\boldsymbol{\eta} \tag{S.136}$$

$$= \sum_{i=1}^{n} \log \tilde{Z}_i + \log \int \mathcal{N}\left(\tilde{\boldsymbol{\mu}}\middle|\boldsymbol{\eta}, \tilde{\boldsymbol{\Sigma}}\right) \mathcal{N}\left(\boldsymbol{\eta}|\mathbf{0}, \mathbf{K}\right) d\boldsymbol{\eta} \tag{S.137}$$

Hence, we have

$$\log q(\mathbf{y}|\mathbf{X}) = \sum_{i=1}^{n} \log \tilde{Z}_i + \log \mathcal{N}\left(\tilde{\boldsymbol{\mu}}\middle|\mathbf{0}, \tilde{\boldsymbol{\Sigma}} + \mathbf{K}\right) \tag{S.138}$$

$$\begin{aligned} = &-\frac{1}{2}\tilde{\boldsymbol{\mu}}^T(\mathbf{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}} - \frac{1}{2}\log\left|\mathbf{K} + \tilde{\boldsymbol{\Sigma}}\right| - \frac{n}{2}\log 2\pi \\ &+ \sum_i \left\{ \log \hat{Z}_i + \frac{1}{2}\log 2\pi + \frac{1}{2}\log(\sigma_{\neg i}^2 + \tilde{\sigma}_i^2) + \frac{(\mu_{\neg i} - \tilde{\mu}_i)^2}{2(\sigma_{\neg i}^2 + \tilde{\sigma}_i^2)} \right\}. \end{aligned} \tag{S.139}$$

Finally,

$$\begin{aligned} &\log q(\mathbf{y}|\mathbf{X}) \\ &= -\frac{1}{2}\tilde{\boldsymbol{\mu}}^T(\mathbf{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}} - \frac{1}{2}\log\left|\mathbf{K} + \tilde{\boldsymbol{\Sigma}}\right| + \sum_i \left\{ \log \hat{Z}_i + \frac{1}{2}\log(\sigma_{\neg i}^2 + \tilde{\sigma}_i^2) + \frac{(\mu_{\neg i} - \tilde{\mu}_i)^2}{2(\sigma_{\neg i}^2 + \tilde{\sigma}_i^2)} \right\}. \end{aligned} \tag{S.140}$$

1) Derivatives wrt hyperparameters: Note that $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\Sigma}}$, $\boldsymbol{\mu}_{\neg}$, and $\boldsymbol{\Sigma}_{\neg}$ are all implicitly functions of the kernel hyperparameter, due to the EP update steps. Hence, the derivative of the marginal is composed of the explicit derivative wrt the hyperparameter ($\frac{\partial}{\partial \alpha_j}$) as well as several implicit derivative (e.g., $\frac{\partial \tilde{\boldsymbol{\mu}}}{\partial \alpha_j}\frac{\partial}{\partial \tilde{\boldsymbol{\mu}}}$) It can be shown that the implicit derivatives are all zero [4], and hence

$$\frac{\partial \log q(\mathbf{y}|\mathbf{X})}{\partial \alpha_j} = \left.\frac{\partial \log q(\mathbf{y}|\mathbf{X})}{\partial \alpha_j}\right|_{\text{explicit}} \tag{S.141}$$

$$= \frac{1}{2}\tilde{\boldsymbol{\mu}}^T(\mathbf{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\frac{\partial \mathbf{K}}{\partial \alpha_j}(\mathbf{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}} - \frac{1}{2}\mathrm{tr}\left((\mathbf{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\frac{\partial \mathbf{K}}{\partial \alpha_j}\right) \tag{S.142}$$

$$= \frac{1}{2}\mathrm{tr}\left(\left[\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T - (\mathbf{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\right]\frac{\partial \mathbf{K}}{\partial \alpha_j}\right), \quad \tilde{\mathbf{z}} = (\mathbf{K} + \tilde{\boldsymbol{\Sigma}})^{-1}\tilde{\boldsymbol{\mu}}. \tag{S.143}$$

2) Derivative wrt dispersion: The only term in the marginal that depends on ϕ directly is $\log \hat{Z}_i$. Its derivative is

$$\frac{\partial}{\partial \phi} \log \hat{Z}_i = \frac{\partial}{\partial \phi} \log \int p(y_i|\theta(\eta_i)) \mathcal{N}\left(\eta_i \middle| \mu_{\neg i}, \sigma^2_{\neg i}\right) d\eta_i \tag{S.144}$$

$$= \frac{1}{\hat{Z}_i} \int \frac{\partial}{\partial \phi} p(y_i|\theta(\eta_i)) \mathcal{N}\left(\eta_i \middle| \mu_{\neg i}, \sigma^2_{\neg i}\right) d\eta_i \tag{S.145}$$

$$= \frac{1}{\hat{Z}_i} \int \left[\frac{h'(y_i,\phi)}{h(y_i,\phi)} + \frac{a'(\phi)}{a(\phi)^2}(\theta(\eta_i) y_i - b(\theta(\eta_i)))\right] p(y_i|\theta(\eta_i)) \mathcal{N}\left(\eta_i \middle| \mu_{\neg i}, \sigma^2_{\neg i}\right) d\eta_i \tag{S.146}$$

$$= \frac{\partial}{\partial \phi} \log h(y_i,\phi) + \frac{a'(\phi)}{a(\phi)^2} \left(y_i \mathbb{E}_q[\theta(\eta_i)] - \mathbb{E}_q[b(\theta(\eta_i))]\right), \tag{S.147}$$

where $\mathbb{E}_q$ is the expectation under $q(\eta_i|\mathbf{X}, \mathbf{y})$. Hence, the derivative of the marginal wrt the dispersion parameter is

$$\frac{\partial \log q(\mathbf{y}|\mathbf{X})}{\partial \phi} = \left.\frac{\partial \log q(\mathbf{y}|\mathbf{X})}{\partial \phi}\right|_{\text{explicit}} = \sum_{i=1}^{n} \frac{\partial}{\partial \phi} \log \hat{Z}_i \tag{S.148}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \phi} \log h(y_i,\phi) + \frac{a'(\phi)}{a(\phi)^2} \sum_{i=1}^{n} \left(y_i \mathbb{E}_q[\theta(\eta_i)] - \mathbb{E}_q[b(\theta(\eta_i))]\right). \tag{S.149}$$

Hence, additional expectations $\mathbb{E}_q[\theta(\eta_i)]$ and $\mathbb{E}_q[b(\theta(\eta_i))]$ are required.

REFERENCES

[1] A. B. Chan and D. Dong, "Generalized gaussian process models," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[2] T. Minka, *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[3] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[4] M. Seeger, "Expectation propagation for exponential families," tech. rep., 2005.