Gaussian Processes with Derivative Observations

Xiaoke Yang (das.xiaoke@hotmail.com)
School of Automation Science and Electrical Engineering,
Beihang University

Abstract: This article summarises necessary derivations for Gaussian processes (GP) with derivative observations. Please refer to 'Solak et al, Derivative observations in Gaussian Process models of dynamic systems.' for more details on derivative observations. Uncertainties over derivative observations are not considered in this article.

1 Preliminary

Assume the underlying system is modelled by a multiple-input single-output function with noisy measurements of the output, i.e.

$$y = f(\mathbf{x}) + v, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^D$, $y \in \mathbb{R}$, and v is an additive Gaussian noise, i.e. $v \sim \mathcal{N}(0, v)$. The underlying function $f(\mathbf{x})$ is modelled by a GP, i.e.

$$f \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}^m, \mathbf{x}^n)). \tag{2}$$

A zero mean function $m(\mathbf{x}) = 0$ is used for the GP and the following covariance function is used

$$\mathrm{cov}[y^m, y^n] = k(\mathbf{x}^m, \mathbf{x}^n) = \alpha \exp\left(-\frac{1}{2}\|\mathbf{x}^m - \mathbf{x}^n\|^2_{\mathbf{\Gamma}}\right) + v\delta_{m,n} \tag{3}$$

where $\mathbf{x}^m, \mathbf{x}^n \in \mathbb{R}^D$ are two input points, $\mathbf{\Gamma} = \mathrm{diag}\left(\begin{bmatrix}\gamma_1 & \gamma_2 & \cdots & \gamma_D\end{bmatrix}\right)$, notation $\|\cdot\|_{\mathbf{\Gamma}}$ is defined as $\|\mathbf{x}\|^2_{\mathbf{\Gamma}} \triangleq \mathbf{x}^\top \mathbf{\Gamma} \mathbf{x}$, and the Kronecker delta is defined as

$$\delta_{m,n} = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases}. \tag{4}$$

We also define another covariance function for convenience,

$$k_f(\mathbf{x}^m, \mathbf{x}^n) = \alpha \exp\left(-\frac{1}{2}\|\mathbf{x}^m - \mathbf{x}^n\|^2_{\mathbf{\Gamma}}\right). \tag{5}$$

Basically, $k_f(\mathbf{x}^m, \mathbf{x}^n)$ calculates the covariance between the function values, instead of the measurements of the function outputs. The vector of hyper-parameters is defined as

$$\boldsymbol{\theta} = \begin{bmatrix}\gamma_1 & \gamma_2 & \cdots & \gamma_D & \alpha & v\end{bmatrix}. \tag{6}$$

2 Derivative Observations

With derivative observations, data for the GP become

Table 1: Input-output data for GP

Type	Derivative observation	Function observation
input	$\mathbf{x}_d \in \mathbb{R}^D$	$\mathbf{x} \in \mathbb{R}^D$
output	$\mathbf{y}_d \in \mathbb{R}^D$	$y \in \mathbb{R}$

If we define the input output data block as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_d^1 & \mathbf{x}_d^2 & \cdots & \mathbf{x}_d^M & \mathbf{x}^1 & \mathbf{x}^2 & \cdots & \mathbf{x}^N \end{bmatrix} \tag{7}$$

$$\mathbf{Y} = \begin{bmatrix} (\mathbf{y}_d^1)^\top & (\mathbf{y}_d^2)^\top & \cdots & (\mathbf{y}_d^M)^\top & y^1 & y^2 & \cdots & y^N \end{bmatrix} \tag{8}$$

Then the output covariance matrix is

$$\mathbf{K} = \begin{bmatrix} \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{\substack{m\in\{1,\ldots,M\} \\ n\in\{1,\ldots,M\}}} & \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{\substack{m\in\{1,\ldots,M\} \\ n\in\{1,\ldots,N\}}} \\ \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & k_{xd}(\mathbf{x}^m, \mathbf{x}_d^n) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{\substack{m\in\{1,\ldots,N\} \\ n\in\{1,\ldots,M\}}} & \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & k_{xx}(\mathbf{x}^m, \mathbf{x}^n) & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{\substack{m\in\{1,\ldots,N\} \\ n\in\{1,\ldots,N\}}} \end{bmatrix} \tag{9}$$

Things we need to compute include $k_{xx}(\mathbf{x}^m, \mathbf{x}^n) \in \mathbb{R}$, $k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n) \in \mathbb{R}^{D\times 1}$, and $k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n) \in \mathbb{R}^{D\times D}$. $k_{xx}(\mathbf{x}^m, \mathbf{x}^n) = k(\mathbf{x}^m, \mathbf{x}^n)$ as in (**??**). For the two derivatives The following results are from reference .

$$k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n)_i = -\alpha\gamma_i(x_{d,i}^m - x_i^n)\exp\left(-\frac{1}{2}\|\mathbf{x}_d^m - \mathbf{x}^n\|_{\mathbf{\Gamma}}^2\right), \tag{10}$$

$$= -\gamma_i(x_{d,i}^m - x_i^n)k_f(\mathbf{x}_d^m, \mathbf{x}^n), \tag{11}$$

$$k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n)_{i,j} = \alpha\gamma_i\left(\delta_{i,j} - \gamma_j(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right)\exp\left(-\frac{1}{2}\|\mathbf{x}_d^m - \mathbf{x}_d^n\|_{\mathbf{\Gamma}}^2\right), \tag{12}$$

$$= \gamma_i\left(\delta_{i,j} - \gamma_j(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right)k_f(\mathbf{x}_d^m, \mathbf{x}_d^n), \tag{13}$$

$$\tag{14}$$

where the subscripts i and j corresponds to the indices within the array or matrix, respectively. and

$$k_f(\mathbf{x}_d^m, \mathbf{x}^n) = \exp\left(-\frac{1}{2}\|\mathbf{x}_d^m - \mathbf{x}^n\|_{\mathbf{\Gamma}}^2\right) = k_f(\mathbf{x}^n, \mathbf{x}_d^m) \tag{15}$$

Then, in vector and matrix form, the above equations can be written as

$$k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n) = -k_f(\mathbf{x}_d^m, \mathbf{x}^n)\mathbf{\Gamma}(\mathbf{x}_d^m - \mathbf{x}^n), \tag{16}$$

$$k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n) = k_f(\mathbf{x}_d^m, \mathbf{x}_d^n)\left(\mathbf{\Gamma} - (\mathbf{\Gamma}(\mathbf{x}_d^m - \mathbf{x}_d^n))(\mathbf{\Gamma}(\mathbf{x}_d^m - \mathbf{x}_d^n))^\top\right), \tag{17}$$

$$= k_f(\mathbf{x}_d^m, \mathbf{x}_d^n)\left(\mathbf{\Gamma} - \mathbf{\Gamma}(\mathbf{x}_d^m - \mathbf{x}_d^n)(\mathbf{x}_d^m - \mathbf{x}_d^n)^\top\mathbf{\Gamma}\right), \tag{18}$$

The GP predictions are

$$p(y^*|\mathbf{X}, \mathbf{Y}, \mathbf{x}^*) \sim \mathcal{N}(\mu, \Sigma) \tag{19}$$

$$\tag{20}$$

where

$$\mu = k(\mathbf{x}^*, \mathbf{X})\mathbf{K}^{-1}\mathbf{Y} \tag{21}$$

$$\Sigma = k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\mathbf{K}^{-1}k(\mathbf{X}, \mathbf{x}^*) \tag{22}$$

where

$$k(\mathbf{x}^*, \mathbf{X}) = \begin{bmatrix} [k_{xd}(\mathbf{x}^*, \mathbf{x}_d^m)]_{m\in\{1,\ldots,M\}} & [k_{xx}(\mathbf{x}^*, \mathbf{x}^n)]_{n\in\{1,\ldots,N\}} \end{bmatrix} \tag{23}$$

$$k(\mathbf{X}, \mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})^\top \tag{24}$$

and the marginal likelihood is

$$L = \log p(\mathbf{Y}|\mathbf{X}) = -\frac{1}{2}\mathbf{Y}^\top \mathbf{K}^{-1}\mathbf{Y} - \frac{1}{2}\log|\mathbf{K}| - \frac{MD+N}{2}\log 2\pi \tag{25}$$

In order to train the GP, the derivative $\frac{\partial L}{\partial \boldsymbol{\theta}}$ needs to be computed, and are listed as follows.

$$\frac{\partial L}{\partial \theta_l} = \operatorname{vec}\left(\frac{\partial L}{\partial \mathbf{K}}\right)^\top \operatorname{vec}\left(\frac{\partial \mathbf{K}}{\partial \theta_l}\right), \quad l \in \{1, 2, \ldots, D+2\}, \tag{26}$$

where $\circ$ denotes the matrix Hadamard product or element-wise product. The first half of this product is a $(MD+N)\times(MD+N)$ matrix, whose i,j^{th} element is defined as $\frac{\partial L}{\partial K_{i,j}}$

$$\frac{\partial L}{\partial \mathbf{K}} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \frac{\partial L}{\partial K_{i,j}} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}_{\substack{i\in\{1,\ldots,MD+N\} \\ j\in\{1,\ldots,MD+N\}}} \tag{27}$$

from matrix cookbook

$$\frac{\partial L}{\partial \mathbf{K}} = \frac{1}{2}\mathbf{K}^{-\top}\mathbf{Y}\mathbf{Y}^T\mathbf{K}^{-\top} - \frac{1}{2}\mathbf{K}^{-\top} \tag{28}$$

Then we need to compute the right half of the product, i.e. $\frac{\partial \mathbf{K}}{\partial \theta_l}$, we still divide this into 4 blocks, in accordance with the definition of $\mathbf{K}$.

$$\frac{\partial k_{xx}(\mathbf{x}^m, \mathbf{x}^n)}{\partial \theta_l} = \frac{\partial k_f(\mathbf{x}^m, \mathbf{x}^n)}{\partial \theta_l} = \alpha \exp\left(-\frac{1}{2}\|\mathbf{x}^m - \mathbf{x}^n\|^2_{\boldsymbol{\Gamma}}\right) \frac{\partial\left(-\frac{1}{2}(\mathbf{x}^m - \mathbf{x}^n)^\top \boldsymbol{\Gamma}(\mathbf{x}^m - \mathbf{x}^n)\right)}{\partial \theta_l}, \tag{29}$$

$$= k_f(\mathbf{x}^m, \mathbf{x}^n)\frac{\partial\left(-\frac{1}{2}\sum_{i=1}^D \gamma_i (x_i^m - x_i^n)^2\right)}{\partial \theta_l}, \tag{30}$$

$$= -\frac{1}{2}k_f(\mathbf{x}^m, \mathbf{x}^n)(x_l^m - x_l^n)^2, \quad l \in \{1, \ldots, D\} \tag{31}$$

$$\frac{\partial k_{xx}(\mathbf{x}^m, \mathbf{x}^n)}{\partial \theta_l} = \frac{\partial k_f(\mathbf{x}^m, \mathbf{x}^n)}{\partial \theta_l} = \exp\left(-\frac{1}{2}\|\mathbf{x}^m - \mathbf{x}^n\|^2_{\boldsymbol{\Gamma}}\right), \quad l = D+1 \tag{32}$$

$$\frac{\partial k_{xx}(\mathbf{x}^m, \mathbf{x}^n)}{\partial \theta_l} = \delta_{m,n}, \quad l = D+2 \tag{33}$$

In the following, $i \in \{1, \ldots, D\}$.

$$\frac{\partial k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n)_i}{\partial \theta_l} = \frac{\partial\left(-\gamma_i(x_{d,i}^m - x_i^n)k_f(\mathbf{x}_d^m, \mathbf{x}^n)\right)}{\partial \theta_l}, \tag{34}$$

$$= -\gamma_i(x_{d,i}^m - x_i^n)\frac{\partial k_f(\mathbf{x}_d^m, \mathbf{x}^n)}{\partial \theta_l} - \delta_{i,l}(x_{d,i}^m - x_i^n)k_f(\mathbf{x}_d^m, \mathbf{x}^n), \quad l \in \{1, \ldots, D\} \tag{35}$$

$$\frac{\partial k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n)_i}{\partial \theta_l} = -\gamma_i(x_{d,i}^m - x_i^n)\frac{\partial k_f(\mathbf{x}_d^m, \mathbf{x}^n)}{\partial \theta_l}, \tag{36}$$

$$= -\gamma_i(x_{d,i}^m - x_i^n)\exp\left(-\frac{1}{2}\|\mathbf{x}_d^m - \mathbf{x}^n\|^2_{\boldsymbol{\Gamma}}\right), \quad l = D+1, \tag{37}$$

$$\frac{\partial k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n)_i}{\partial \theta_l} = 0, \quad l = D+2 \tag{38}$$

or by using matrix differentiation, we can get the following directly from xx.

$$\frac{\partial k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n)}{\partial \theta_l} = -\mathbf{\Gamma}(\mathbf{x}_d^m - \mathbf{x}^n)\frac{\partial k_f(\mathbf{x}_d^m, \mathbf{x}^n)}{\partial \theta_l} - \mathbf{\Delta}_l(\mathbf{x}_d^m - \mathbf{x}_d^n)k_f(\mathbf{x}_d^m, \mathbf{x}_d^n), \tag{39}$$

$$= \frac{1}{2}k_f(\mathbf{x}_d^m, \mathbf{x}^n)(x_{d,l}^m - x_l^n)^2\mathbf{\Gamma}(\mathbf{x}_d^m - \mathbf{x}^n) - \mathbf{\Delta}_l(\mathbf{x}_d^m - \mathbf{x}^n)k_f(\mathbf{x}_d^m, \mathbf{x}^n), \quad l \in \{1, \ldots, D\}, \tag{40}$$

$$\frac{\partial k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n)}{\partial \theta_l} = -\mathbf{\Gamma}(\mathbf{x}_d^m - \mathbf{x}^n)\exp\left(-\frac{1}{2}\|\mathbf{x}_d^m - \mathbf{x}^n\|_{\mathbf{\Gamma}}^2\right), \quad l = D+1, \tag{41}$$

$$\frac{\partial k_{dx}(\mathbf{x}_d^m, \mathbf{x}^n)}{\partial \theta_l} = \mathbf{0}_{D\times 1} \quad l = D+1, \tag{42}$$

$$\tag{43}$$

where $\mathbf{\Delta}_l \in \mathbb{R}^{D\times D}$ is a square matrix with the l^{th} diagonal element $\Delta_{l,l} = 1$ and all other elements as 0.

In the following $i \in \{1, \ldots, D\}, j \in \{1, \ldots, D\}$.

$$\frac{\partial k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n)_{i,j}}{\partial \theta_l} = \frac{\partial\left[\gamma_i\left(\delta_{i,j} - \gamma_j(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right)k_f(\mathbf{x}_d^m, \mathbf{x}_d^n)\right]}{\partial \theta_l}, \tag{44}$$

$$= \frac{\partial\left[\gamma_i\left(\delta_{i,j} - \gamma_j(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right)\right]}{\partial \theta_l}k_f(\mathbf{x}_d^m, \mathbf{x}_d^n) + \gamma_i\left(\delta_{i,j} - \gamma_j(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right)\frac{\partial k_f(\mathbf{x}_d^m, \mathbf{x}_d^n)}{\partial \theta_l}, \tag{45}$$

$$= \left[\delta_{i,l}\delta_{j,l} - (\delta_{i,l}\gamma_j + \delta_{j,l}\gamma_i)(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right]k_f(\mathbf{x}_d^m, \mathbf{x}_d^n) + \gamma_i\left(\delta_{i,j} - \gamma_j(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right)\frac{\partial k_f(\mathbf{x}_d^m, \mathbf{x}_d^n)}{\partial \theta_l}, \quad l \in \{1, \ldots, D\} \tag{46}$$

$$\frac{\partial k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n)_{i,j}}{\partial \theta_l} = \gamma_i\left(\delta_{i,j} - \gamma_j(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right)\frac{\partial k_f(\mathbf{x}_d^m, \mathbf{x}_d^n)}{\partial \theta_l} \tag{47}$$

$$= \gamma_i\left(\delta_{i,j} - \gamma_j(x_{d,i}^m - x_{d,i}^n)(x_{d,j}^m - x_{d,j}^n)\right)\exp\left(-\frac{1}{2}\|\mathbf{x}_d^m - \mathbf{x}_d^n\|_{\mathbf{\Gamma}}^2\right), \quad l = D+1 \tag{48}$$

$$\frac{\partial k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n)_{i,j}}{\partial \theta_l} = 0, \quad l = D+2 \tag{49}$$

or by using matrix differentiation, we can get the following directly from xx.

$$\frac{\partial k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n)}{\partial \theta_l} = \left(\mathbf{\Delta}_l - \mathbf{\Gamma}(\mathbf{x}_d^m - \mathbf{x}_d^n)(\mathbf{x}_d^m - \mathbf{x}_d^n)^\top\mathbf{\Delta}_l - \mathbf{\Delta}_l(\mathbf{x}_d^m - \mathbf{x}_d^n)(\mathbf{x}_d^m - \mathbf{x}_d^n)^\top\mathbf{\Gamma}\right)k_f(\mathbf{x}_d^m, \mathbf{x}_d^n) + \mathbf{\Gamma}\left(\mathbf{I} - (\mathbf{x}_d^m - \mathbf{x}_d^n)(\mathbf{x}_d^m - \mathbf{x}_d^n)^\top\mathbf{\Gamma}\right)(-\frac{1}{2}k_f(\mathbf{x}_d^m, \mathbf{x}_d^n)(x_{d,l}^m - x_{d,l}^n)^2), \quad l \in \{1, \ldots, D\} \tag{50}$$

$$\frac{\partial k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n)}{\partial \theta_l} = \mathbf{\Gamma}\left(\mathbf{I} - (\mathbf{x}_d^m - \mathbf{x}_d^n)(\mathbf{x}_d^m - \mathbf{x}_d^n)^\top\mathbf{\Gamma}\right)\exp\left(-\frac{1}{2}\|\mathbf{x}_d^m - \mathbf{x}_d^n\|_{\mathbf{\Gamma}}^2\right), \quad l = D+1 \tag{51}$$

$$\frac{\partial k_{dd}(\mathbf{x}_d^m, \mathbf{x}_d^n)}{\partial \theta_l} = \mathbf{0}_{D\times D}, \quad l = D+2 \tag{52}$$

Note, in some cases, the parameter γ_l take an inverse exponentiated square form of

$$\gamma_l = \frac{1}{\exp(\lambda_l)^2}, \quad l \in \{1, \ldots, D\} \tag{53}$$

In this case, all the above differentiation should have one more step, i.e.

$$\frac{\partial k(\cdot,\cdot)}{\partial \lambda_l} = \frac{\partial k(\cdot,\cdot)}{\partial \gamma_l}\frac{\mathrm{d}\gamma_l}{\mathrm{d}\lambda_l} = -2\exp(\lambda_l)^{-2}\frac{\partial k(\cdot,\cdot)}{\partial \gamma_l} = -2\gamma_l\frac{\partial k(\cdot,\cdot)}{\partial \gamma_l}, \quad l \in \{1,\ldots,D\}. \tag{54}$$