1 Gaussian Process

[1] introduces Gaussian Process regression as an alternative view to Bayesian regression.

1.1 Bayesian Regression

In linear regression setup, we assume output $y \in \mathbb{R}$ is a linear function of inputs $x \in \mathbb{R}^d$, corrupted with additive iid normal noise,

$$y = f(x) + \epsilon \quad \text{where} \quad f(x) = w^T x \quad \text{and} \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2) \tag{1}$$

The Bayesian setup considers $w \in \mathbb{R}^d$ as a random variable, endowed with prior $w \sim \mathcal{N}(0, \Sigma_p)$. Using Bayes rule, we can find the posterior of weights given data, which is again a normal random variable $p(w \mid X, y) = \mathcal{N}\left(w \,;\, A^{-1}b, A^{-1}\right)$ where $A = \frac{1}{\sigma_n^2} X^T X + \Sigma_p^{-1}$ and $b = \frac{1}{\sigma_n^2} X^T y$ and $X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^{n \times 1}$ are design matrices. For test point x_*, the predictive distribution of $f_* = f(x_*)$ is the average likelihood of f_* under model $f(x; w)$ with respect to posterior of w.

$$p(f_* \mid x_*, X, y) = \int p(f_* \mid x_*, w) p(w \mid X, y) \, dw \tag{2}$$

We can think of the predictive distribution as a linear function $f_* = x_*^T w$ of weights, a normal random variable, and therefore is normal. Therefore, $f_* \mid x_*, X, y \sim \mathcal{N}(x_*^T A^{-1} b, x_*^T A^{-1} x_*)$. The natural extension to Bayesian linear regression is to kernelize it, for example assume a linear model in some feature space $f(x) = \phi(x)^T w$ for some feature map $\phi : \mathbb{R}^d \to \mathbb{R}^D$. We can write predictive distribution as

$$f_* \mid x_*, X, y \sim \mathcal{N}(k(x_*, X)(k(X, X) + \sigma_n^2 I)^{-1} y, \tag{3}$$

$$k(x_*, x_*) - k(x_*, X)(k(X, X) + \sigma_n^2 I)^{-1} k(X, x_*)) \tag{4}$$

where $k(X, X') = \Phi \Sigma_p \Phi'^T \in \mathbb{R}^{n \times n'}$ and $\Phi \in \mathbb{R}^{n \times D}$ are the feature vectors.

1.2 Gaussian Process

Definition 1. *(Gaussian Process) A Gaussian process is a stochastic process $\{X_t\}_{t \in T}$ such that, for every finite subset of indices $t_1, \cdots, t_k \in T$, $(X_{t_1}, \cdots, X_{t_k})$ is multivariate normal.*

A Gaussian process $f \sim \mathcal{GP}(m, k)$ over $\mathbb{R}^{\mathcal{X}}$ is fully specified by the mean function $m : \mathcal{X} \to \mathbb{R}$ and covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where

$$m(x) = \mathbb{E}\left[f(x)\right] \qquad k(x, x') = \mathbb{E}\left[(f(x) - m(x))(f(x') - m(x'))\right] \tag{5}$$

The covariance function determines function's behavior, for example its stationarity, smoothness, and periodicity etc. For example, the squared exponential covariance function $k_{\text{SE}}(x, x') = \exp(-\frac{1}{2\ell^2} \left\| x - x' \right\|_2^2)$ enforces the prior knowledge that functions are smooth, i.e. inputs are close in the Euclidean sense will have similar outputs. See Figure (1) for some examples of samples from Gaussian process.

Example 1. *(Intuition about covariance matrix for $\mathcal{GP}$) First consider $(y_1, y_2) \sim \mathcal{N}(\mathbf{0}, \Sigma)$ where $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = \rho_{12}$. We know $y_1 \mid y_2 = a \sim \mathcal{N}(\rho_{12} a, 1 - \rho_{12}^2)$. When $Cov(y_1, y_2) = \rho_{12} \uparrow 1$, y_1's samples conditioned on $y_2 = a$ fall close to a with high probability. When $\rho_{12} \downarrow 0$, $y_1 \mid y_2 = a$ will distribute like a unit normal, regardless of values that y_2 take. Extend this intuition to gaussian process: whenever $k(x_i, x_j)$ is large, y_i, y_j are correlated and so observing $y_i = a$ provides a strong prior on how y_j will behave, or in other words, reduce the uncertainty of values that y_j can take dramatically.*

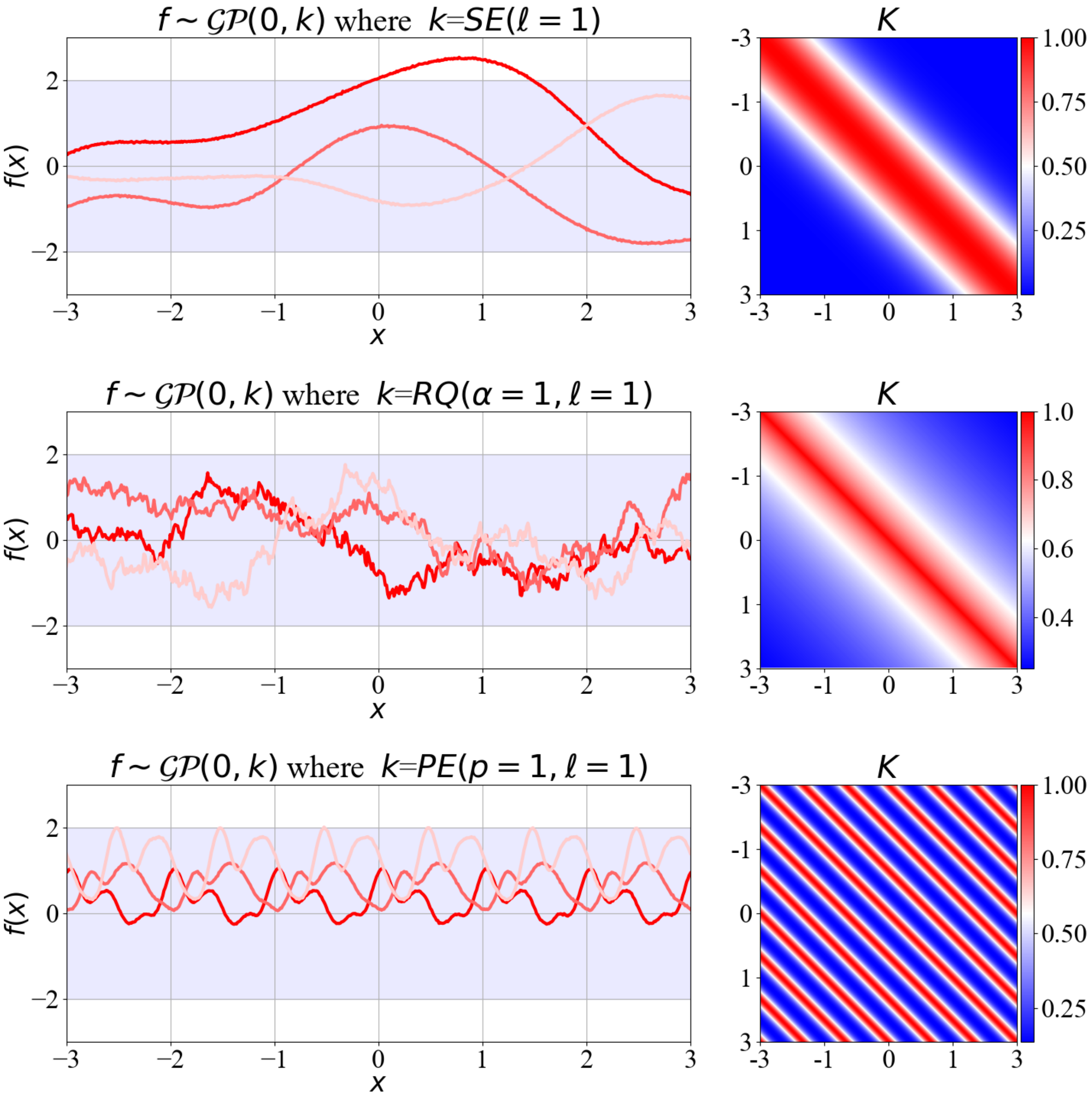

Figure 1: (Left) Samples from Gaussian process prior and (Right) covariance matrix at test locations.

1.3 Gaussian Process Regression

Instead of placing a prior over weights $p(w)$ to quantify randomness in function $f(x) = \phi(x)^T w$, we model function directly as a Gaussian process, $f \sim \mathcal{GP}(0, k)$. There is a one-to-one correspondence between the two views. For example, $f(x) = \phi(x)^T w$ with prior $w \sim \mathcal{N}(0, \Sigma_p)$ used in kernel Bayesian regression has

$$\mathbb{E}\left[f(x)\right] = \phi(x)^T \mathbb{E}\left[w\right] = 0 \qquad \mathbb{E}\left[f(x)f(x')\right] = \phi(x)^T \Sigma_p \phi(x') \tag{6}$$

Therefore, $f \sim \mathcal{GP}(0,k)$ where $k(x,x') = \phi(x)^T\Sigma_p\phi(x')$. Note k is in fact a valid kernel. (Since Σ_p is psd, $\Sigma_p = UDU^T$ by SVD. We can write $k(x,x') = \langle\psi(x),\psi(x')\rangle$ where $\psi(x) = \Sigma_p^{1/2}\phi(x)$ and $\Sigma_p^{1/2} = UD^{1/2}U^T$). Conversely, any valid kernel used in kernel Bayesian regression can be used to parameterize the covariance function of the Gaussian process model over f.

Since $y = f(x) + \epsilon$, we have $\mathbf{y} = \mathbf{f} + \sigma_n^2 I \sim \mathcal{N}(0, k(X,X) + \sigma_n^2 I)$ where $\mathbf{y}, \mathbf{f} \in \mathbb{R}^{n\times 1}$. We can write the joint distribution of observed values and function values at some test locations X_* as

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}_*\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}k(X,X)+\sigma_n^2 I & k(X,X_*)\\ k(X_*,X) & k(X_*,X_*)\end{bmatrix}\right) \tag{7}$$

We can derive the predictive distribution for $\mathbf{f}_* \mid \mathbf{y}$ by simply apply conditional distribution formula

$$\mathbf{f}_* \mid X, \mathbf{y}, X_* \sim \mathcal{N}(k(X_*,X)(k(X,X)+\sigma_n^2 I)^{-1}\mathbf{y} \tag{8}$$

$$k(X_*,X_*) - k(X_*,X)(k(X,X)+\sigma_n^2 I)^{-1}k(X,X_*)) \tag{9}$$

which has exact form compared to (4). More compactly, for a single test point x_*, $\mu_{\mathbf{f}_*} = k_*^T(K+\sigma_n^2 I)^{-1}$ and $\text{Var}(\mathbf{f}_*) = k(x_*,x_*) - k_*^T(K+\sigma_n^2 I)^{-1}k_*$ where $k_* = k(X,x_*) \in \mathbb{R}^{n\times 1}$. See Figure (2) for examples of Gaussian process regression with varying data size and hyperparameters.

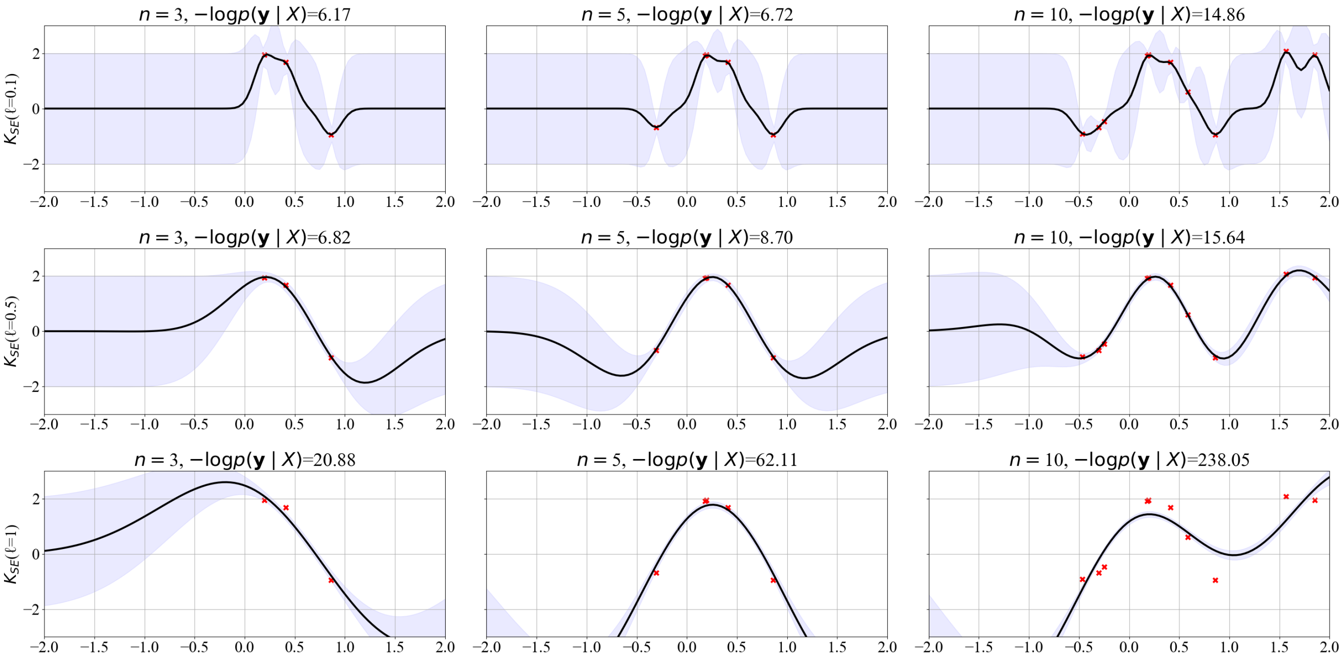

Figure 2: This plot shows predictive mean (black) and 95% confidence interval fit using Gaussian process regression assuming a SE prior over $f \sim \mathcal{GP}(0, k_{SE})$ of varying lengthscale ℓ and observed sample sizes n. Fitting using Cholesky factorization is more stable than applying matix inverse directly.

Given data $(X, \mathbf{y})$, marginal likelihood $p(\mathbf{y} \mid X)$ quantifies how likely data is observed under our additive noise model $\mathbf{y} \mid \mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 I)$ on average with respect to a Gaussian process prior over the function $\mathbf{f} \mid X \sim \mathcal{N}(0, K)$,

$$p(\mathbf{y} \mid X) = \int p(\mathbf{y} \mid \mathbf{f}, X)p(\mathbf{f} \mid X)\, d\mathbf{f} \tag{10}$$

Analogously in Bayesian linear regression, the marginal likelihood marginalize over weights $p(\mathbf{y} \mid X) = \int p(\mathbf{y} \mid X, w) p(w) \, dw$. We can obtain a closed form expression by reading off (7), i.e. $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, K + \sigma_n^2 I)$. In empirical Bayes setup, kernel hyperparameters can be found by maximizing log marginal likelihood,

$$\log p(\mathbf{y} \mid X) = -\frac{1}{2}\mathbf{y}^T (K + \sigma_n^2 I)^{-1} \mathbf{y} - \frac{1}{2} \log |K + \sigma_n^2 I| - \frac{n}{2} \log 2\pi \tag{11}$$

References

[1] Carl Edward Rasmussen and Williams Christopher. *Gaussian Process for Machine Learning.* MIT Press, 2006.