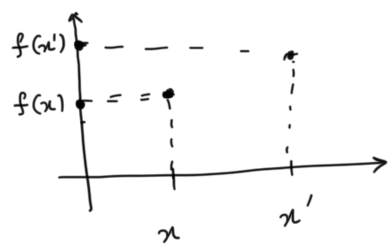
Gaussian Process Regression

Setup: Given $D := \{(x_1, y_1), \dots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

1. Model definition:

$$y = f(x) + \varepsilon \quad \begin{cases} f(x) \sim GP(\cancel{\mu_p(x)}^{\,0}, \underline{k(x, x'; \theta)}) & (1) \\ \varepsilon \sim \mathcal{N}(0, \sigma_n^2 \underset{n \times n}{I}) \end{cases}$$



$$\begin{bmatrix} f(x) \\ f(x') \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_p(x) \\ \mu_p(x') \end{bmatrix}, \begin{bmatrix} k(x, x; \theta) & k(x, x \\ k(x', x; \theta) & k(x', x \end{bmatrix} \right.$$

RBF/SE: $\quad k(x, x'; \theta) = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{i=1}^{d} \frac{(x_i - x_i')^2}{l_i^2}\right)$

Trainable parameters: $\Theta := \{\sigma_f^2, l_1, \dots, l_d, \sigma_n^2\} \geq 0$

* How to sample a GP prior:

X = np.linspace(0, 1, 200)

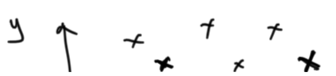
$\underset{200 \times 200}{K} = k(X, X; \theta) \leftarrow$

$L \rightarrow$ cholesky(K)

$\underset{200 \times 1}{f} = \cancel{\mu(X)}^{\,0} + z L^{\frac{1}{2}}, \quad z \sim \mathcal{N}(0, I)$

2. Model training:

Data: $\mathcal{D} = \{X, y\}$, $\underset{n \times d}{X}$, $\underset{n \times 1}{y}$



$p(y|X, f)\, p(f|X)$

$$p(f \mid X, y) = \frac{\quad}{p(y \mid X)}$$

$$p(y \mid x) = \int \underbrace{\underbrace{p(y \mid x, f)}}_{\mathcal{N}(y \mid f(x),\, \sigma_n^2 I)} \overbrace{p(f \mid x)}^{\mathcal{N}(0, K)}\, df$$

Log-marginal likelihood: $\log p(y \mid x) = \underbrace{-\frac{1}{2}\log|K + \sigma_n^2 I|}_{\text{model complexity}} \underbrace{- \frac{1}{2} y^T (K + \sigma_n^2 I)^{-1} y}_{\text{data fit}} - \underbrace{\frac{n}{2}\log}_{\text{co}}$

$\uparrow$

Log-likelihood $\log p(y \mid x, f) = \underbrace{-\frac{n}{2}\log 2\pi\sigma_n^2}\ \underbrace{- \frac{1}{2\sigma_n^2}(y - f(x))^T (y - f(x))}_{\text{data fit}}$

MLE

$\uparrow$

MAP $\quad \log p(y \mid x, f) + \log p(f \mid x) =$

$$-\frac{n}{2}\log 2\pi\sigma_n^2 - \frac{1}{2\sigma_n^2}(y - f(x))^T (y - f(x)) - \|f\|_{\mathcal{H}}$$

Training objective:

Type-2 MLE estimation

$$\theta = \arg\min \mathcal{L}(\theta) := -\log p(y \mid X)$$

$\hookrightarrow \{\sigma_f^2, \ell_1, \ldots, \ell_d, \sigma_n^2\}$

$$-\log p(y \mid x) = \frac{1}{2}\log|K| + \frac{1}{2} y^T \underset{n\times n}{K^{-1}} y + \frac{n}{2}\log 2\pi :=$$

$$\underset{n\times n}{K} = k(X, X; \theta) + \sigma_n^2 I$$

Cholesky decomposition $\mathcal{O}(n^3)$

$\longrightarrow$ Gradients : $\nabla_\theta \log p(y \mid x)$ (analytically or via AD)

3.) Predictive posterior:

$$\begin{bmatrix} f(x^*) \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k(x^*, x^*) & k(x^*, X) \\ & \underbrace{k(X, X)}_{K} \end{bmatrix} \right)$$

$$\boxed{P(f(x^*) \mid X, y) = \mathcal{N}\left(\mu(x^*), \Sigma(x^*) \right)}$$

$$\left\{ \begin{array}{l} \mu(x^*) = k(x^*, X) K^{-1} y \\ \\ \Sigma(x^*) = k(x^*, x^*) - k(x^*, X) K^{-1} k(X, x^*) \end{array} \right\}$$