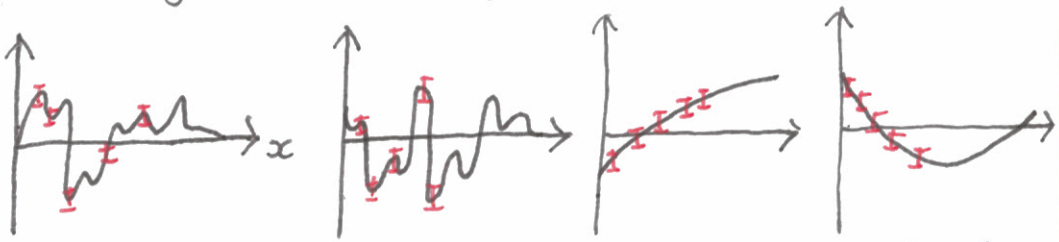
Two Bayesian Linear Regression models

Both $\underline{w} \sim N(\underline{0}, \sigma_w^2 \mathbb{I})$, $\underline{y} \sim N(\underline{\Phi}\,\underline{w}, \sigma_y^2)$

↑ from RBFs bandwidth h

Samples synthetic $\underline{w}$ and $\underline{y}$



$h = 0.05$ $h = 0.05$ $h = 1$ $h = 1$

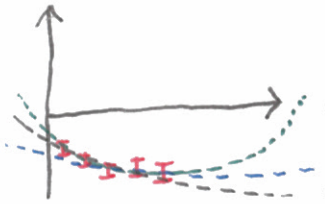
Observe real data

Model posterior:

$$p(h \mid \underline{y}, X) \propto \underbrace{p(\underline{y} \mid X, h)}_{\text{Marginal likelihood}} p(h)$$

Picking $h = 0.05$ vs $h = 1$ is just a Gaussian classifier!

Posterior for $h = 1$

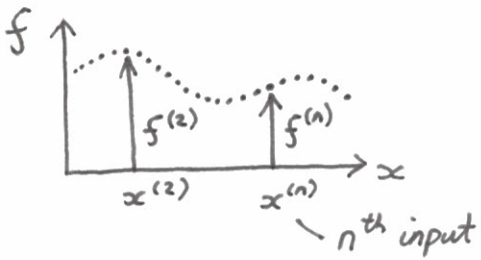


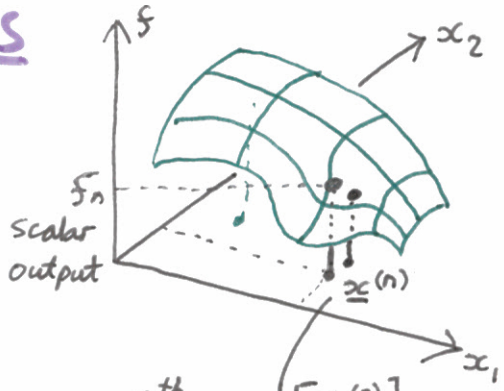
$$p(\underline{w} \mid \underline{y}, X, h = 1) = N(\underline{w}; \underline{w}_N, V_N)$$

} samples of $f(\underline{x}; \underline{w})$

(Variance of f at $\underline{x}$: $\underline{x}^T V_N \underline{x}$)

Gaussian Processes





n^{th} D-dimensional input $\begin{bmatrix} x_1^{(n)} \\ x_2^{(n)} \end{bmatrix}$

$$\underline{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix}$$

N outputs for $\{\underline{x}^{(n)}\}$ stored in X.

covariance / kernel function

If function $f \sim GP(m, k)$ $\Rightarrow$ $\boxed{p(\underline{f}) = N(\underline{f}; \underline{m}, K)}$

mean function

$m_i = m(\underline{x}^{(i)})$ usually 0

$K_{ij} = k(\underline{x}^{(i)}, \underline{x}^{(j)})$

Mercer kernels

Example:

$$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \exp\left(-\|\underline{x}^{(i)} - \underline{x}^{(j)}\|^2\right)$$

We need k to always give positive (semi-) definite K

Things we can do with Gaussians

For a joint Gaussian:

$$p(\underline{f}, \underline{g}) = \mathcal{N}\left(\begin{bmatrix} \underline{f} \\ \underline{g} \end{bmatrix}; \begin{bmatrix} \underline{a} \\ \underline{b} \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right)$$

Marginals

$$p(\underline{f}) = \int p(\underline{f}, \underline{g}) \, d\underline{g}$$

$$= \mathcal{N}(\underline{f}; \underline{a}, A)$$

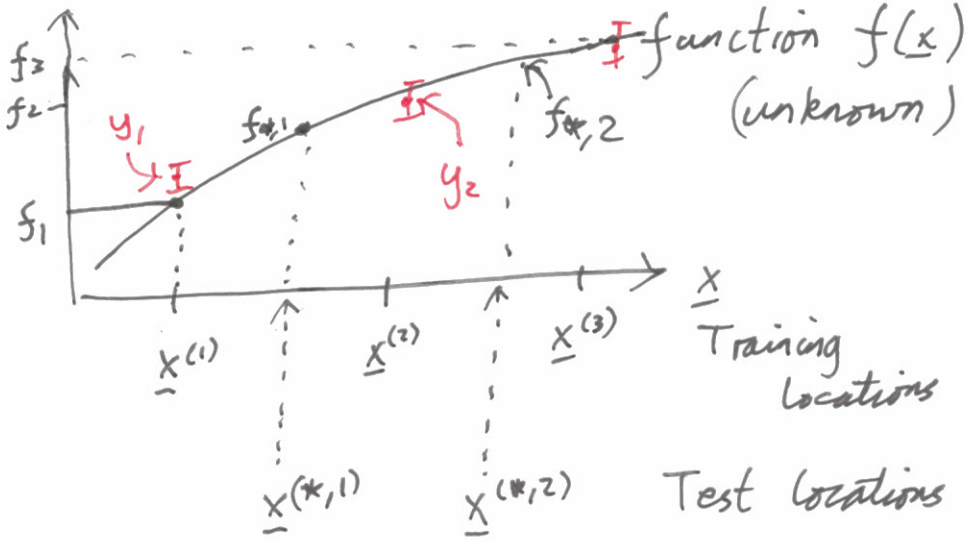
Conditionals

$$p(\underline{f} \mid \underline{g}) = \mathcal{N}(\underline{f}; \underline{a} + CB^{-1}(\underline{g} - \underline{b}), A - CB^{-1}C^T)$$

also

$$p(\underline{g} \mid \underline{f}) = \mathcal{N}(\underline{g}; \underline{b} + C^T A^{-1}(\underline{f} - \underline{a}), B - C^T A^{-1} C)$$

Apply to Regression



$$p(\underline{f}, \underline{f}_*) = \mathcal{N}\left(\begin{bmatrix}\underline{f} \\ \underline{f}_*\end{bmatrix}; \begin{bmatrix}\underline{0} \\ \underline{0}\end{bmatrix}, \begin{bmatrix}k(X,X) & k(X,X_*) \\ k(X_*,X) & k(X_*,X_*)\end{bmatrix}\right)$$

Notation

$$K(X, Z)_{ij} = k(\underline{x}^{(i)}, \underline{z}^{(j)})$$

$N \times D$ $M \times D$

$N \times M$

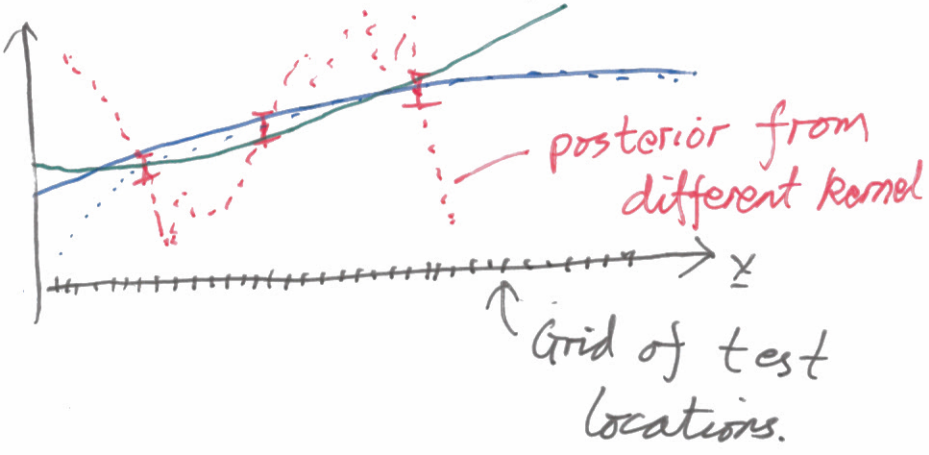
$$p(\underline{y}, \underline{f}_*)$$

$$= \mathcal{N}\left(\begin{bmatrix} \underline{y} \\ \underline{f}_* \end{bmatrix}; \begin{bmatrix} \underline{0} \\ \underline{0} \end{bmatrix}, \begin{bmatrix} k(X,X) + \sigma_y^2 \mathbb{I} & k(X, X_*) \\ k(X_*, X) & k(X_*, X_*) \end{bmatrix} \right)$$

Inference

$$p(\underline{f}_* | \underline{y}) = \mathcal{N}(\underline{f}_*; \_\_\_\_, \_\_\_\_)$$

From standard results for Gaussians.



Bayesian Linear Regression is a GP

Model

$$f_i = f(\underline{x}^{(i)}) = \underline{w}^T \underline{x}^{(i)} + b$$

Prior $\underline{w} \sim N(\underline{0}, \sigma_w^2 \mathbb{I})$, $b \sim N(0, \sigma_b^2)$

$$\text{cov}(f_i, f_j) = \mathbb{E}[f_i f_j] - \cancel{\mathbb{E}[f_i]\mathbb{E}[f_j]}$$ ($\mathbb{E}[f_i] = 0$)

$$= \mathbb{E}\left[(\underline{w}^T \underline{x}^{(i)} + b)^T (\underline{w}^T \underline{x}^{(j)} + b) \right]$$

$$= \mathbb{E}\left[\underline{x}^{(i)T} \underline{w}\, \underline{w}^T \underline{x}^{(j)} + b^2 + \dots \right]$$

$$= \underline{x}^{(i)T} \underbrace{\mathbb{E}[\underline{w}\,\underline{w}^T]}_{\sigma_w^2 \mathbb{I}} \underline{x}^{(j)} + \underbrace{\mathbb{E}[b^2]}_{\sigma_b^2} + \underbrace{\dots}_{0}$$

$$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_w^2 \underbrace{\underline{x}^{(i)T} \underline{x}^{(j)}}_{\underline{\phi}(\underline{x}^{(i)})^T \underline{\phi}(\underline{x}^{(j)})} + \sigma_b^2$$