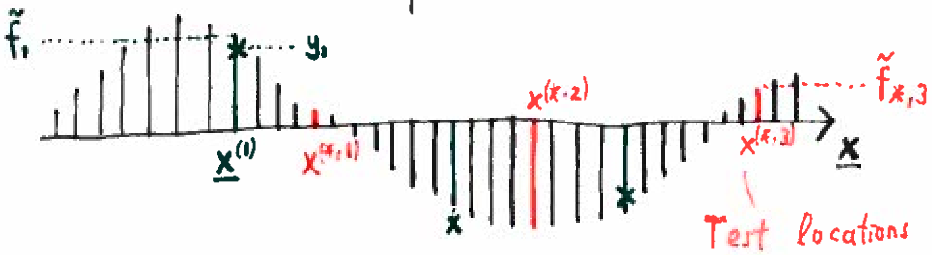
Gaussian Processes

"Function view":

Function values are parameters of the model



Prior on function

$$GP(k)$$

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

Posterior

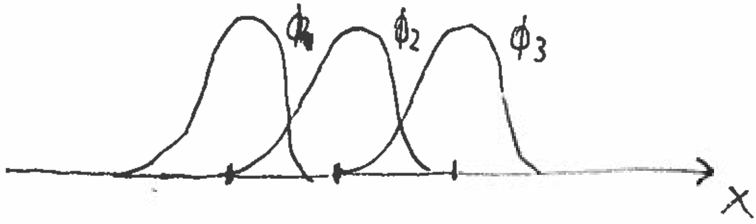
$$p(\tilde{\underline{f}}_* | \underline{y}) = \mathcal{N}(\ldots)$$: see notes

"Weight space view":

$$f(\underline{x}; \underline{w}) = \underline{w}^T \underline{\phi}(\underline{x}) \quad ; \quad \text{Prior } \underline{w} \sim \mathcal{N}(\underline{0}, \sigma_w^2 \mathbb{I})$$

$$\Rightarrow k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_w^2 \, \underline{\phi}(\underline{x}^{(i)})^T \underline{\phi}(\underline{x}^{(j)})$$

RBFs in weight space and function view



IF we put RBFs <u>everywhere</u>

→ Analytically derive $\underline{\phi}^T \underline{\phi}$

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

$$\downarrow \ \#\text{RBFs} \to \infty$$

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

"Kernel trick"

- Rewrite algorithm so it only depends on inner products
- Set $\underline{\phi}(\underline{x}^{(i)})^T \underline{\phi}(\underline{x}^{(j)}) = \overbrace{k(\underline{x}^{(i)}, \underline{x}^{(j)})}^{\text{Mercer kernel}}$

For 1 prediction:

$$p(\tilde{f}_* | \underline{y}) = \mathcal{N}(\tilde{f}_*; m, s^2)$$

One can show:

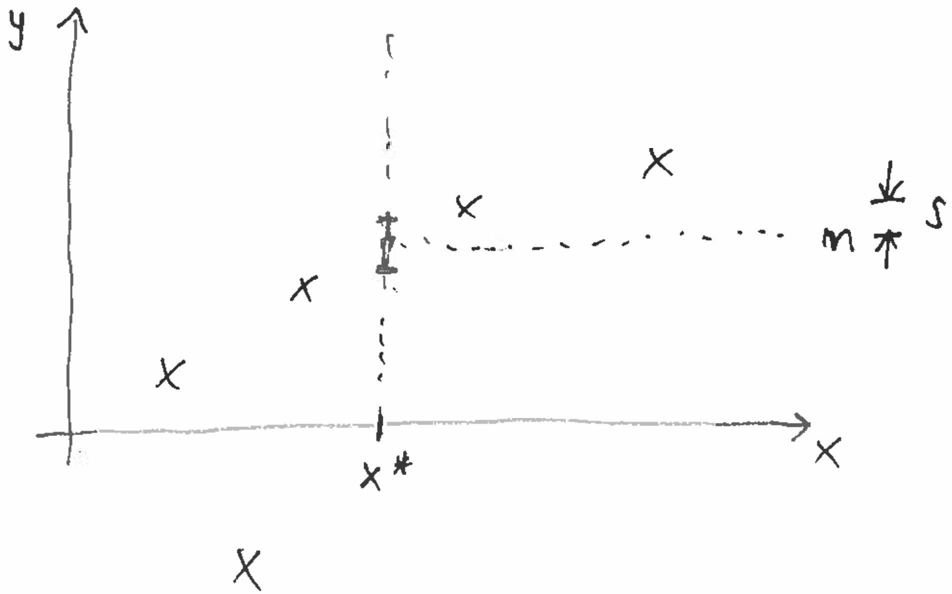
$$m = \underline{k}^{*T} \left(K(X,X) + \sigma_y^2 \mathbb{I} \right)^{-1} \underline{y}$$

training labels

where $(\underline{k}^*)_i = k(\underline{x}^*, \underline{x}^{(i)})$

$$s^2 = k(\underline{x}^*, \underline{x}^*) - \underbrace{\underline{k}^{*T} \underbrace{\left(K(X,X) + \sigma_y^2 \mathbb{I} \right)^{-1}}_{\text{pos. def}} \underline{k}^*}_{\text{pos. def.}}$$

Effect of surprising data point



Can we be more uncertain about x^* in response to surprises?

⇝ Change GP kernel function

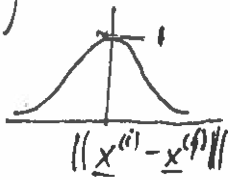
⇒ Learn the kernel

How can we learn the kernel?

Kernel function examples

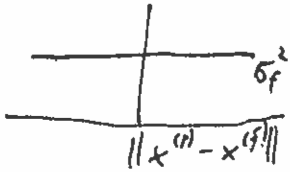
- Squared exponential:

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



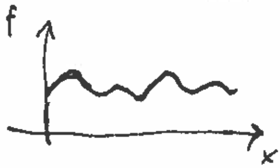
- Constant:

$$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_f^2$$



- Periodic: don't memorize

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



- Kernel can be combined:

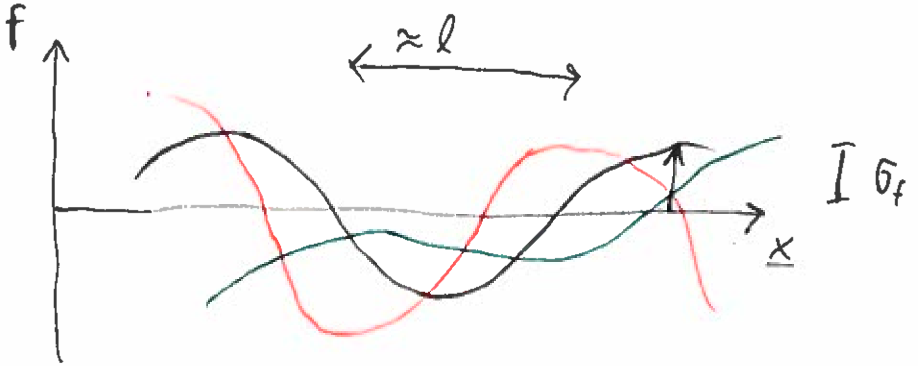
$$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \alpha\, k_1(\underline{x}^{(i)}, \underline{x}^{(j)}) + \beta\, k_2(\underline{x}^{(i)}, \underline{x}^{(j)})$$

$$\alpha > 0, \quad \beta > 0$$

is a kernel function if k_1 and k_2 are kernel functions

Learn the kernel

$$k(\underline{x}^{(i)}, \underline{x}^{(j)}) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_d (x_d^{(i)} - x_d^{(j)})^2 / \ell_d^2\right)$$



Pick parameters by (marginal) likelihood:

$$p(\underline{y} \mid X, \theta = \{\sigma_y^2, \sigma_f^2, \{\ell_d\}\})$$

$$= \mathcal{N}(\underline{y}; \underline{0}, K(X, X) + \sigma_y^2 \mathbb{I})$$