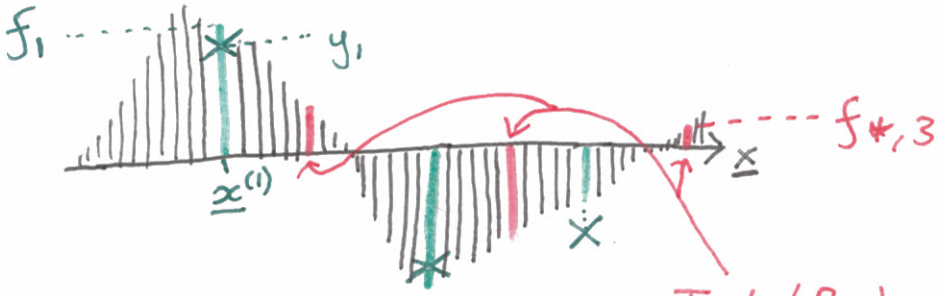
Gaussian Processes

"Function viewpoint"

Function values are parameters of model



Test / Probe locations X_*

Prior on function

$$GP(k)$$

$$\Rightarrow p(\underline{f}, \underline{f}_*) = \mathcal{N}\left(\begin{bmatrix}\underline{f} \\ \underline{f}_*\end{bmatrix}; \underline{0}, \begin{bmatrix}K_{xx} & K_{x*} \\ K_{*x} & K_{**}\end{bmatrix}\right)$$

Posterior

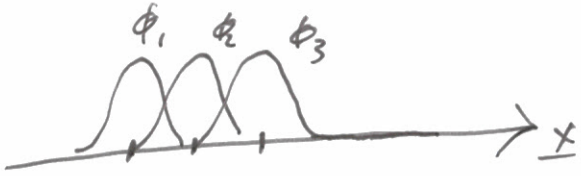
$$p(\underline{f}_* | \underline{y}) = \mathcal{N}(\cdots) \quad \text{(see notes)}$$

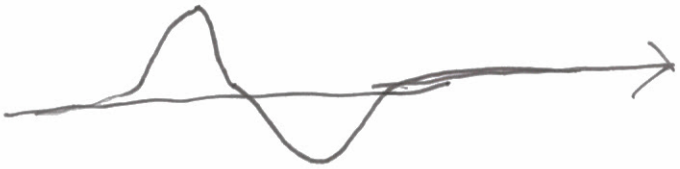
"Weight Space view"

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

$$f(\underline{x}) = \underline{w}^T \underline{\phi}(\underline{x})$$

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





If we put RBF <u>everywhere</u>

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

P84 GPML

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

<u>"Kernel trick"</u>

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

"Mercer kernel"

Aside Kernel Logistic Regression

(Non-exam.)

cost / $O(DN)$

Linear case, SGD, "one epoch"

$\underline{w} \leftarrow \underline{0}$ $\quad D \times 1$ $\quad$ step size λ

for $n = 1 \ldots N$:

$\underline{w} \leftarrow \underline{w} + u^{(n)} \underline{x}^{(n)} - \lambda \underline{w}$

($u^{(n)} \underline{x}^{(n)}$: gradient involves $y^{(n)}$ & confidence; $-\lambda \underline{w}$: regularizer $\times$ step size)

$\underline{w}$ is in span of $\{\underline{x}^{(n)}\}$

$\underline{w} = \sum_n a_n \underline{x}^{(n)} = X^T \underline{a}$

$\underline{a} \leftarrow \underline{0}$ $\quad N \times 1$

for $n = 1 \ldots N$:

$\underline{a} \leftarrow (1-\lambda)\, \underline{a}$

$a_n \leftarrow a_n + u^{(n)}$

$O(N^2 D)$ or is dim.

To get $u^{(n)}$ or to predict

need $\underline{w}^T \underline{x}^{(*)} = \sum_n a_n \underline{x}^{(n)T} \underline{x}^{(*)}$

$= \sum_n a_n k(\underline{x}^{(n)}, \underline{x}^{(*)})$

For 1 prediction:

$$p(f_* | \underline{y}) = N(f_*; m, s^2)$$

Can show: ($N \times N$; Noise variance)

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

($\underline{y}$: training labels)

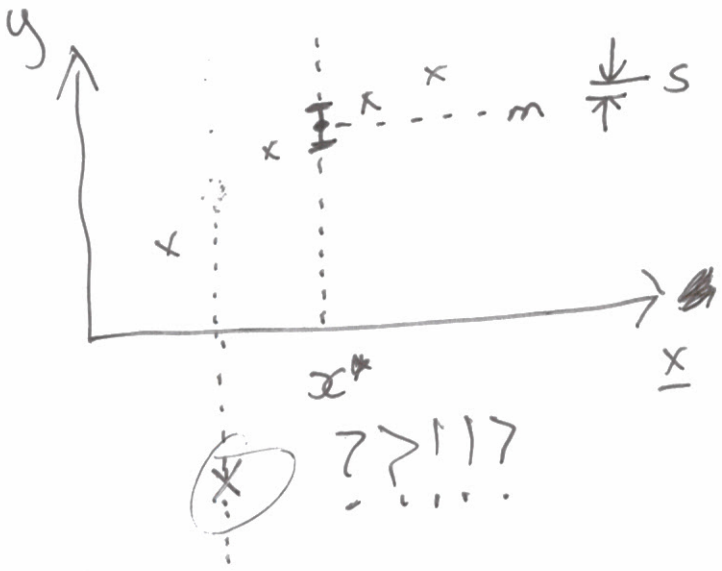
$$(\underline{k}^*)_i = k(\underline{x}^*, \underline{x}^{(i)})$$

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

($\left(k(X,X) + \sigma_y^2 \mathbb{I} \right)$: +ve definite)

Has no $\underline{y}$ dependence.

+ve



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

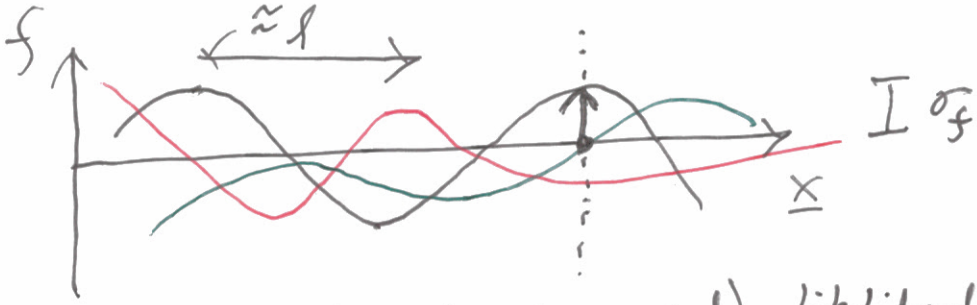
→ Change the GP kernel function.

⇒ Learn the kernel... how?

Learning the kernel

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

Also learn noise level σ_y^2



Pick parameters by (marginal) likelihood:

$$p(y \mid X, \theta = \{\sigma_y^2, \sigma_f^2, \{\ell_d\}\}) = \mathcal{N}(y; \underline{0}, k(X,X) + \sigma_y^2 \mathbb{I})$$