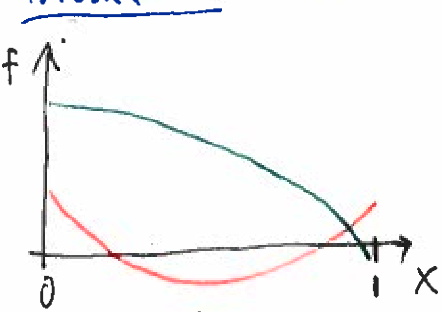
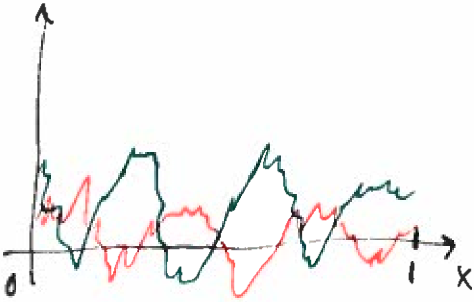
Bayesian Inference and Bandwidth

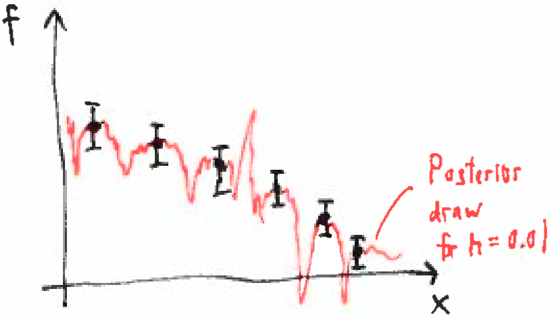
100 basis functions (RBFs), spaced between 0 and 1

Model 1: Bandwidth $h=1$ **Model 2:** $h=0.01$

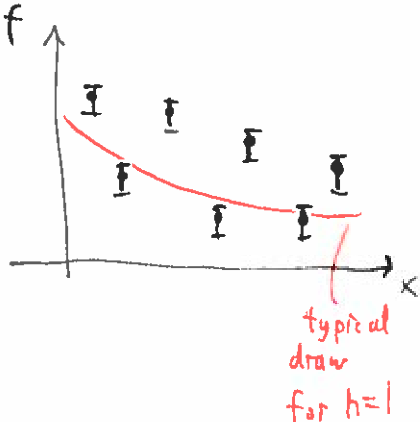




↓ Data from model 1 ↓ Data from model 2



$$p(y \mid X, h=1) \gg p(y \mid X, h=0.01)$$



$$p(y \mid X, h=1) < p(y \mid X, h=0.01)$$

$\Rightarrow$ Marginal likelihood can be used for model choice

Can compare model on training data!

Likelihood not suitable: $p(y \mid X, h=0.01, \hat{w})$ big for data sampled from model 1

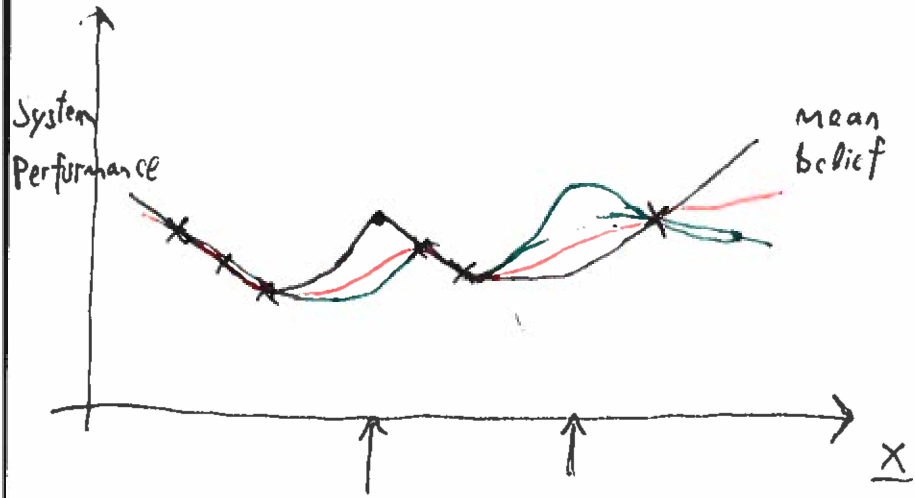
Likelihood of all parameters:

$$p(\underline{y} \mid \underline{w}, h, \ldots, X)$$

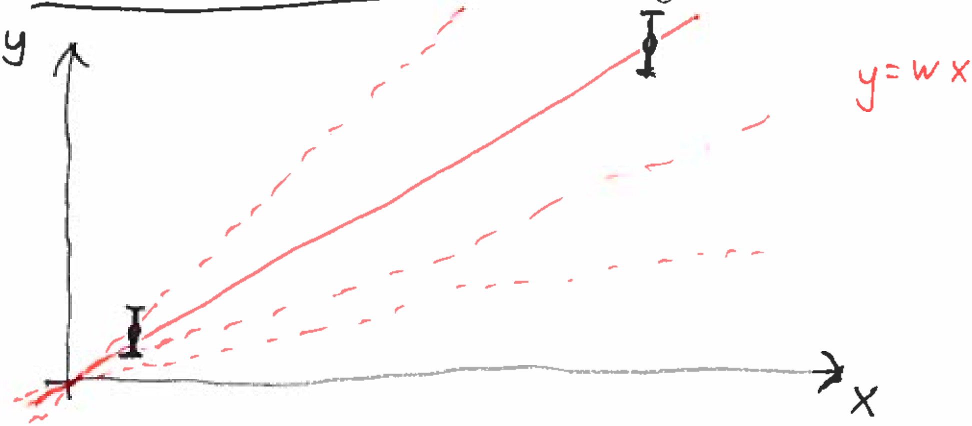
Marginal likelihood:

$$\int p(\underline{y} \mid \underline{w}, h, \ldots, X) \cdot p(w \mid h, \ldots X) \, d\underline{w}$$

Bayesian Optimisation



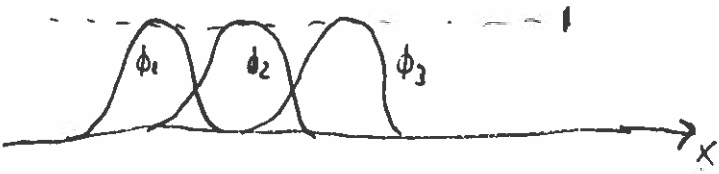
Limitations of linear regression



Active selection

$\Rightarrow$ Make x as big as possible

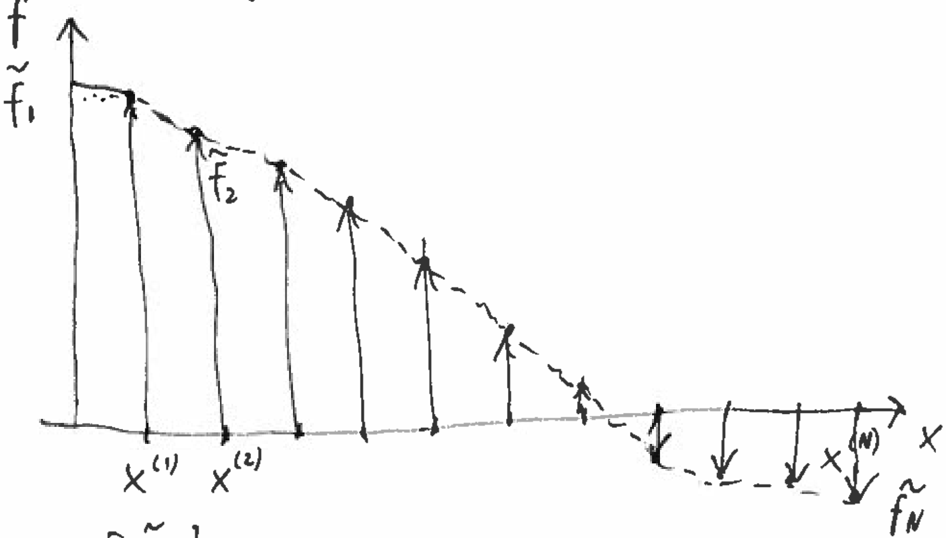
with basis functions



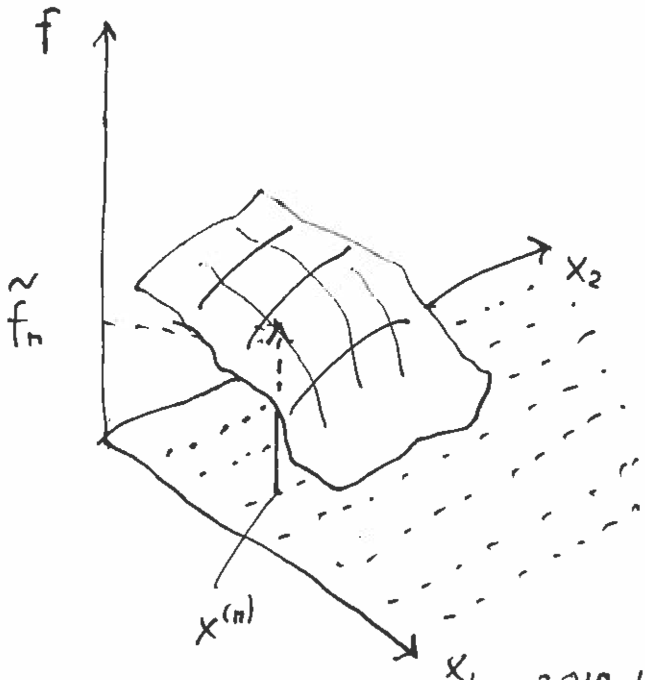
- Certain that outputs will be 0 if we move away from basis functions

Gaussian processes

Very big Gaussian distribution



$$\tilde{f} = \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \vdots \\ \tilde{f}_N \end{bmatrix}$$

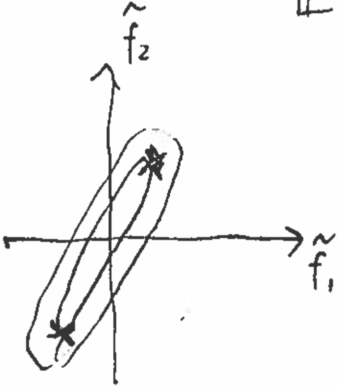


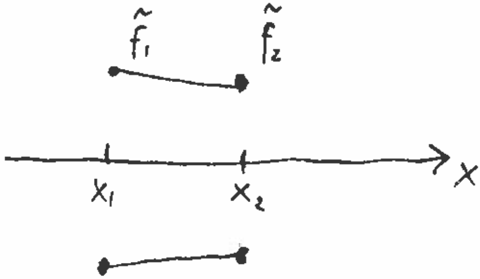
Gaussian process prior

$$p(\tilde{f}) = \mathcal{N}(\tilde{f}; \underline{0}, \Sigma)$$

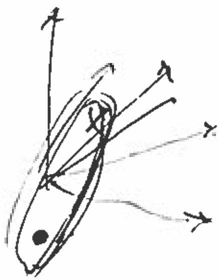
$$\Sigma_{ij} = \mathrm{cov}(\tilde{f}_i, \tilde{f}_j)$$

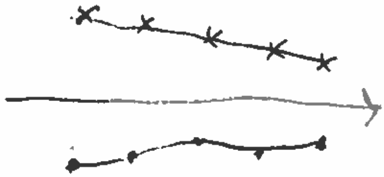
$$= \mathbb{E}[\tilde{f}_i \tilde{f}_j] - \cancel{\mathbb{E}[\tilde{f}_i]\,\mathbb{E}[\tilde{f}_j]}$$





5-dim Gaussian





Covariance function / kernel function

Function prior $f \sim GP$ with kernel k:

For any subset of values $\underline{\tilde{f}}$:

$$p(\underline{\tilde{f}}) = \mathcal{N}(\underline{\tilde{f}}; \underline{0}, K)$$

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

"kernel function" or "covariance function"

What functions can we use for k?

"Mercer kernels" / Positive definite kernels

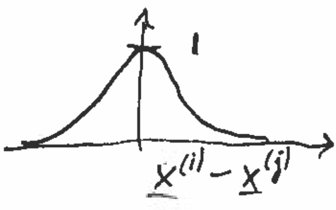
$\Rightarrow K$ positive semi-definite

Example:

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

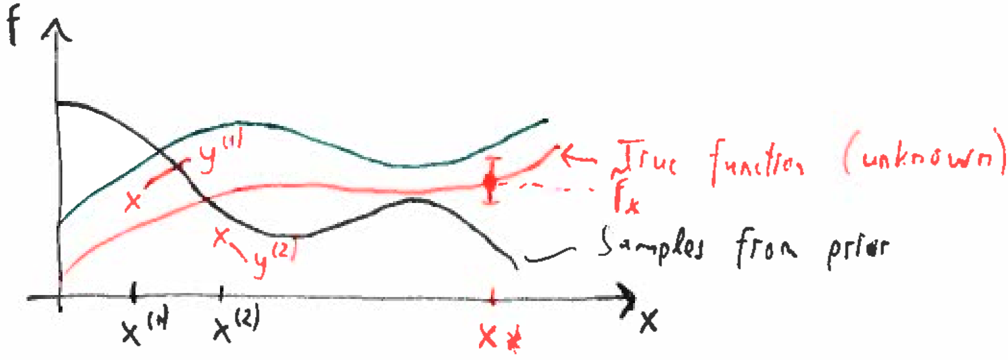
"Gaussian kernel" or "RBF kernel"

or "Squ. exponential kernel"



Regression model with GP

Prior on functions $f \sim GP(k)$



Observation model:

$$y_n \sim \mathcal{N}(\tilde{f}_n, \sigma_y^2)$$

$\Rightarrow$ Likelihood $p(y_n \mid \underline{\tilde{f}}) = \mathcal{N}(y_n; \hat{f}_n, \sigma_y^2)$

for n datapoints, product over $n = 1, \ldots, N$