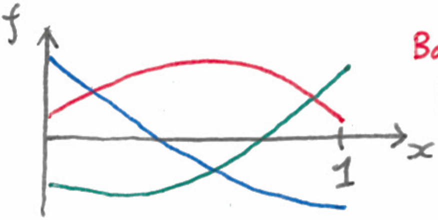
# Bayesian Inference & Bandwidth

Assume $k=100$ RBF basis functions

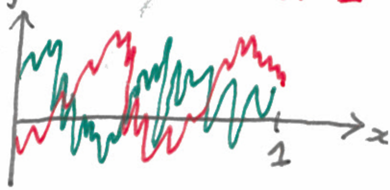
Centers evenly spaced between 0 and 1

Samples from prior $\underline{w} \sim \mathcal{N}(\underline{0}, \mathbb{I}_k)$

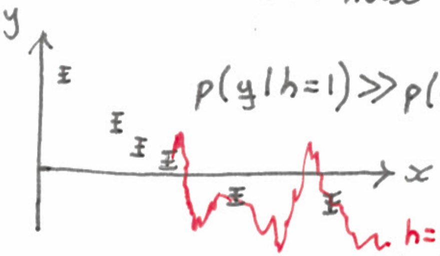



Bandwidth $h=1$

$h=0.01$



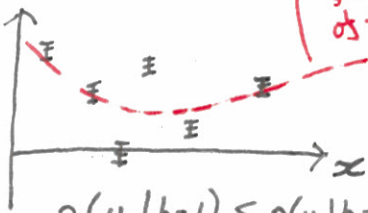

Typical observation
known noise level $\sigma_y$




$p(y|h=1) \gg p(y|h=0.01)$

$h=0.01$
Thinks this is typical

$h=1$ model is fairly sure of this




$p(y|h=1) \leq p(y|h=0.$

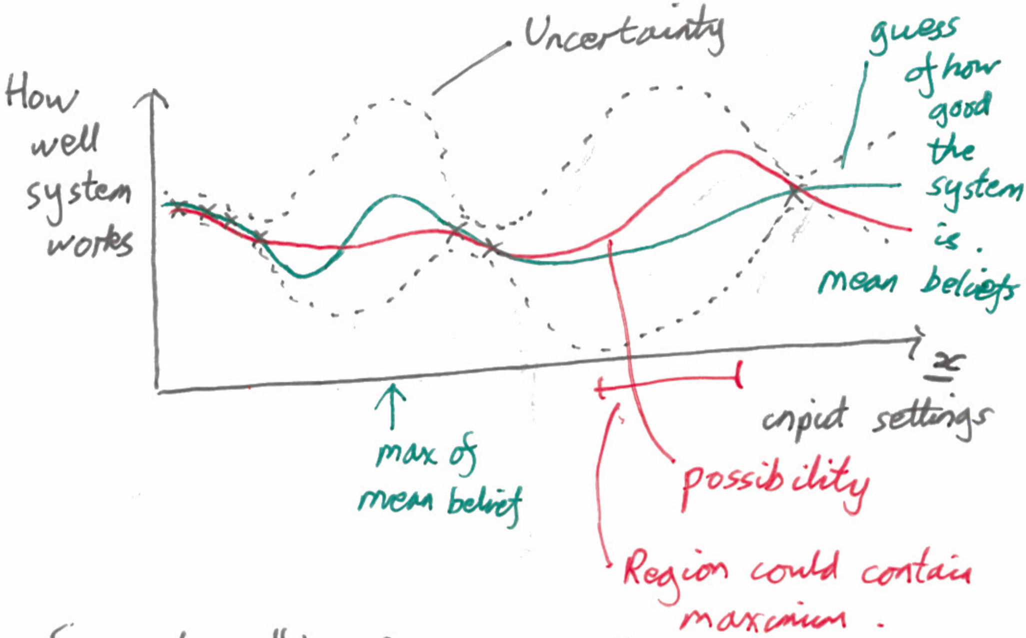
<u>Marginal likelihood</u>

$$p(\text{data} \mid \text{model}) = \int p(\text{data} \mid \underline{w}, \text{model})\, p(\underline{w} \mid \text{model})\, d\underline{w}$$

Can compare models on training set.

Could also choose $\sigma_y$, $\sigma_w$ in $\underline{w} \sim \mathcal{N}(\underline{0}, \sigma_w^2 \mathbb{I}_k)$, ...

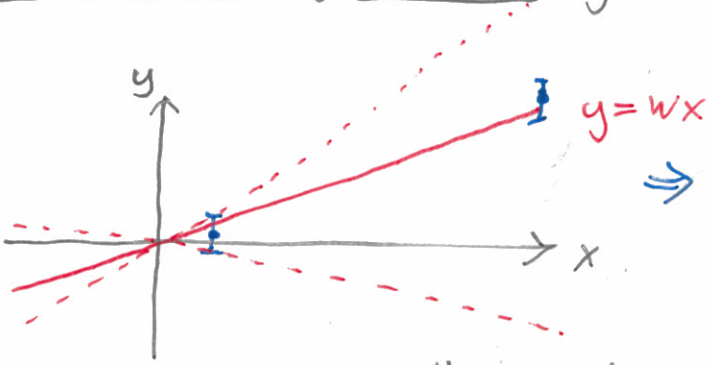
# Bayesian Optimization




See also "Active Learning"

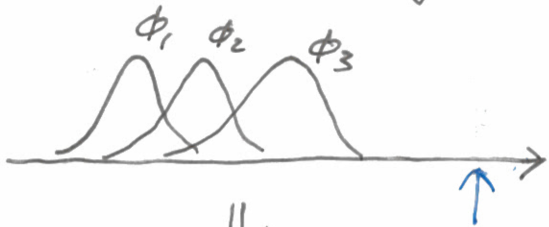
eg to ~~ex~~ build a good classifier.

# Limitations of linear regression



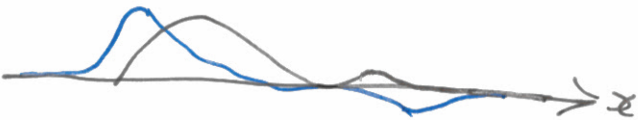

$\Rightarrow$ Go to big x.

$\Downarrow$ Go to basis $f^{\underline{n}}$s.



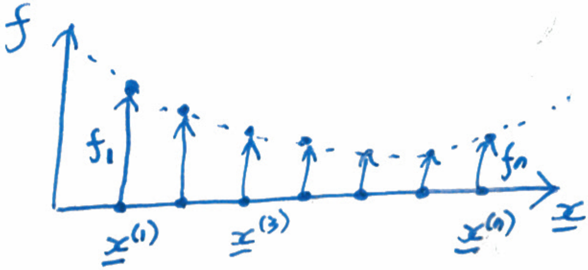

we know $f^{\underline{n}}$ & we know it's zero.

$\Downarrow$

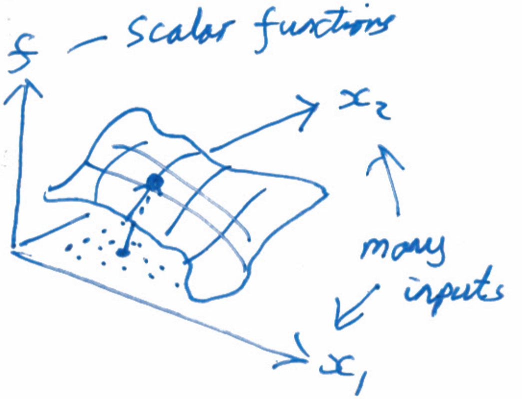


# Gaussian Processes

Really big Gaussian distribution




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




# Gaussian Process Prior

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

$$\Sigma_{ij} = \text{cov}[f_i f_j]$$

0 If zero mean

$$= \mathbb{E}[f_i f_j] - \cancel{\mathbb{E}[f_i]\mathbb{E}[f_j]}$$

# 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

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

# GP Regression

Function prior $f \sim GP$

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

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

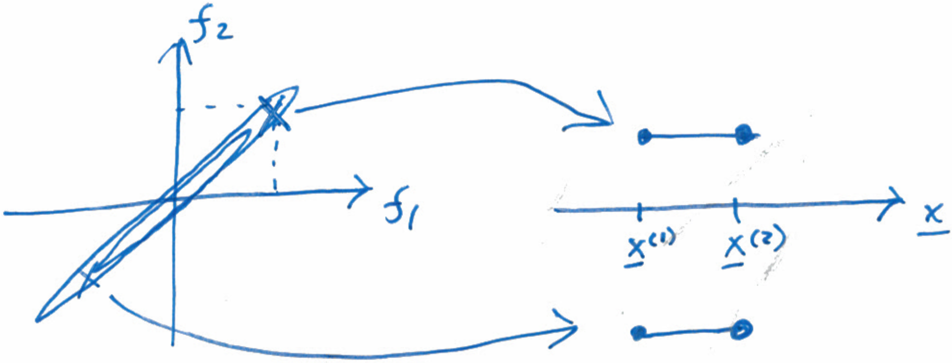
$$K_{ij} = k(\underline{x}^{(i)}, \underline{x}^{(j)})$$

kernel function
or covariance function.

"Mercer kernels" / Positive definite kernels

$\Rightarrow K$ will always
be positive semi-definite.

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




5-Dim Gaussian