7. Gaussian Processes (contd.)

Prediction with a Gaussian Process

In the case of only one test point $\mathbf{x}^*$ we have

$$K(X, \mathbf{x}^*) = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_*) \\ \vdots \\ k(\mathbf{x}_N, \mathbf{x}_*) \end{pmatrix} = \mathbf{k}_*$$

Now we compute the conditional distribution

$$p(y^* \mid \mathbf{x}^*, X, \mathbf{y}) = \mathcal{N}(y_* \mid \mu_*, \Sigma_*)$$

where

$$\mu_* = \mathbf{k}_*^T K^{-1} \mathbf{t}$$

$$\Sigma_* = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T K^{-1} \mathbf{k}_*$$

This defines the **predictive distribution.**

Implementation

Algorithm 1: GP regression

Data: training data $(X, \mathbf{y})$, test data $\mathbf{x}_*$
Input: Hyper parameters σ_f^2, l, σ_n^2

$K_{ij} \leftarrow k(\mathbf{x}_i, \mathbf{x}_j)$
$L \leftarrow \texttt{cholesky}(K + \sigma_n^2 I)$
$\boldsymbol{\alpha} \leftarrow L^T \backslash (L \backslash \mathbf{y})$

Precomputed during Training

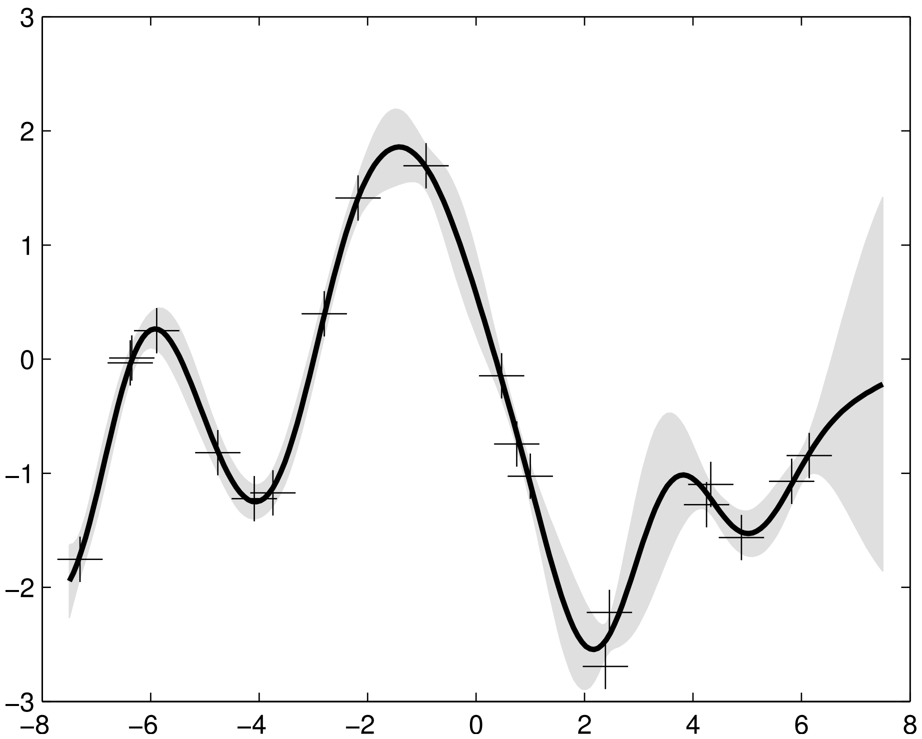
$\mathbb{E}[f_*] \leftarrow \mathbf{k}_*^T \boldsymbol{\alpha}$
$\mathbf{v} \leftarrow L \backslash \mathbf{k}_*$
$\texttt{var}[f_*] \leftarrow k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^T \mathbf{v}$

Test Phase

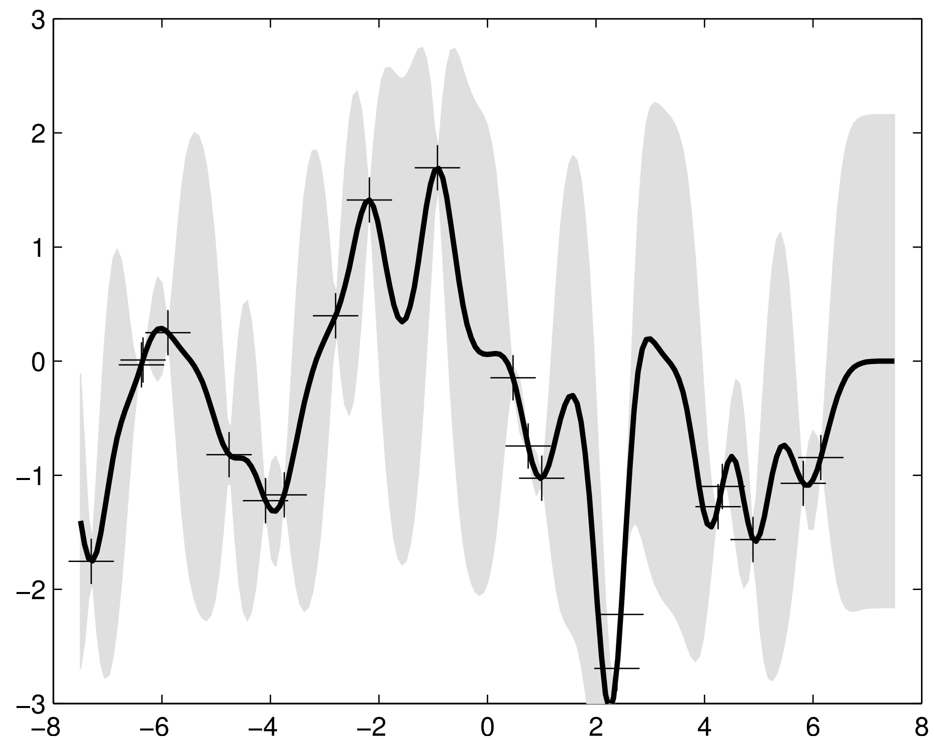
$\log p(\mathbf{y} \mid X) \leftarrow -\frac{1}{2}\mathbf{y}^T \boldsymbol{\alpha} - \sum_i \log L_{ii} - \frac{N}{2} \log(2\pi)$

- Cholesky decomposition is numerically stable
- Can be used to compute inverse efficiently

Varying the Hyperparameters



$$l = \sigma_f = 1, \quad \sigma_n = 0.1$$

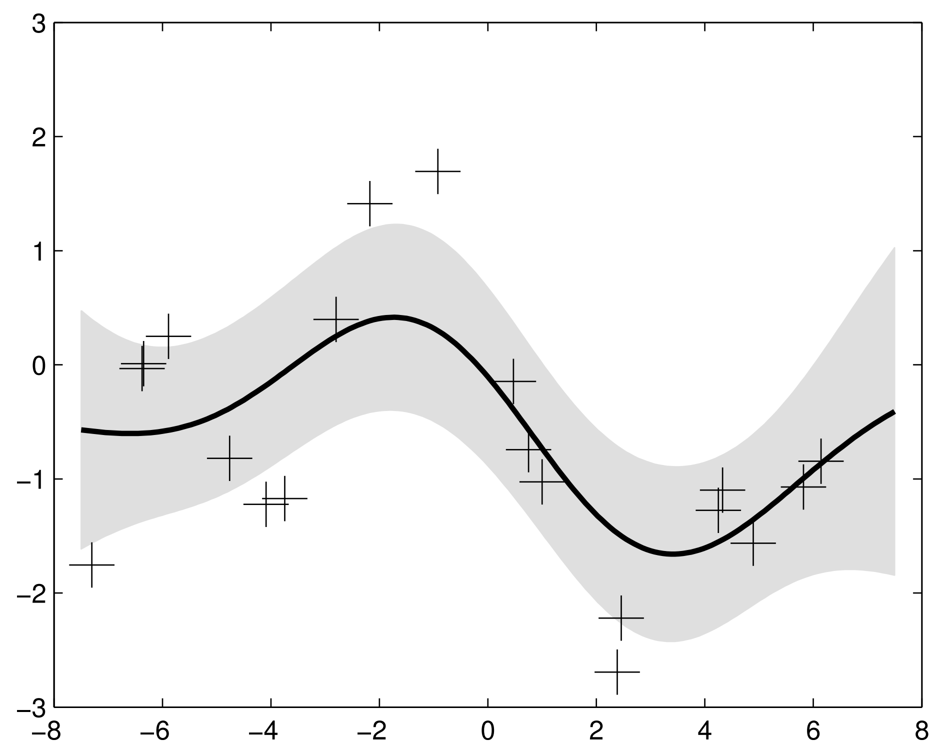


$$l = 0.3,$$

$$\sigma_f = 1.08,$$

$$\sigma_n = 0.0005$$

- 20 data samples
- GP prediction with different kernel hyper parameters



$$l = 3$$

$$\sigma_f = 1.16$$

$$\sigma_n = 0.89$$

Varying the Hyperparameters

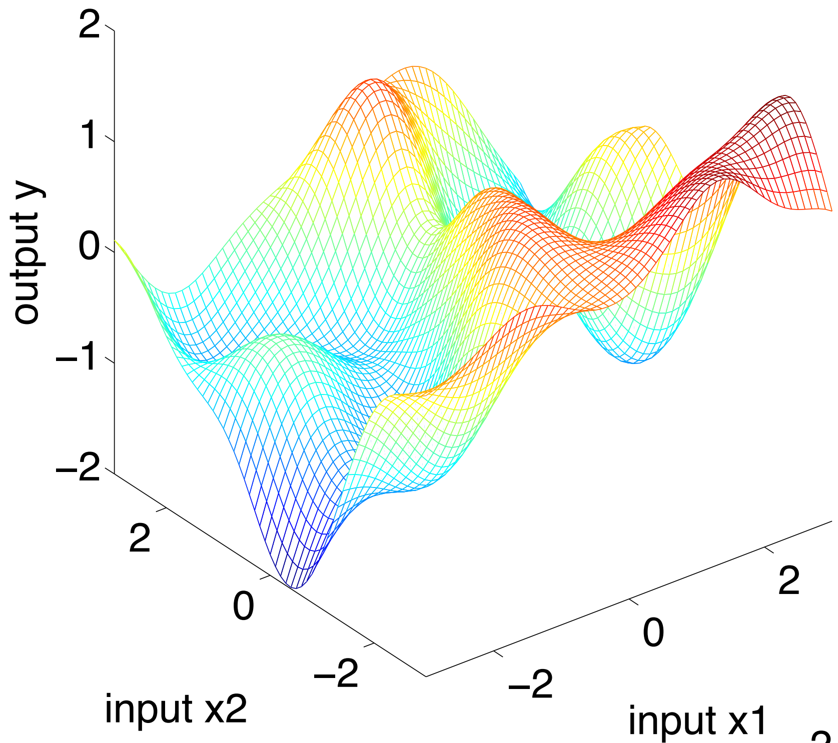
The squared exponential covariance function can be generalized to

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp(-\frac{1}{2}(\mathbf{x}_p - \mathbf{x}_q)^T M(\mathbf{x}_p - \mathbf{x}_q)) + \sigma_n^2 \delta_{pq}$$

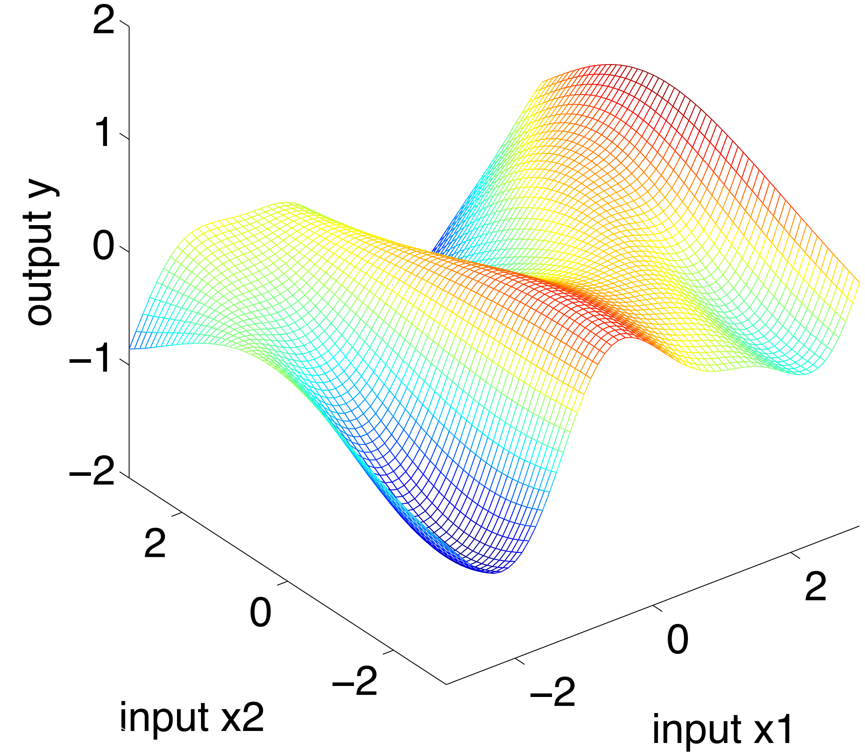
where M can be:

- $M = l^{-2} I$: this is equal to the above case
- $M = \mathrm{diag}(l_1, \dots, l_D)^{-2}$: every feature dimension has its own length scale parameter
- $M = \Lambda\Lambda^T + \mathrm{diag}(l_1, \dots, l_D)^{-2}$: here Λ has less than D columns

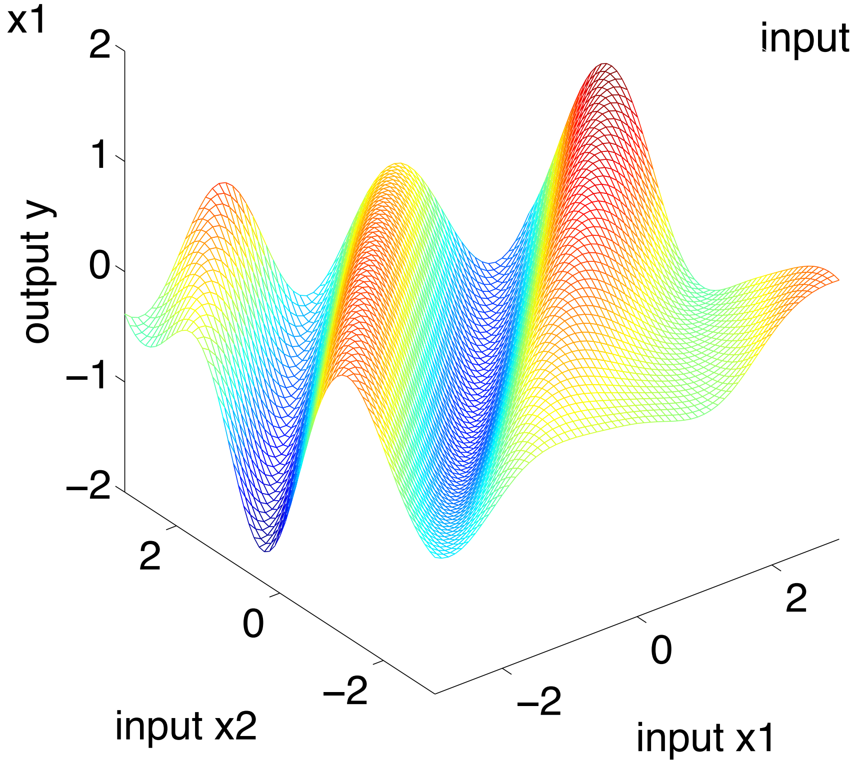
Varying the Hyperparameters



$$M = I$$



$$M = \mathrm{diag}(1,3)^{-2}$$



$$M = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \mathrm{diag}(6,6)^{-2}$$

Estimating the Hyperparameters

To find optimal hyper parameters we need the **marginal likelihood:**

$$p(\mathbf{y} \mid X) = \int p(\mathbf{y} \mid \mathbf{f}, X) p(\mathbf{f} \mid X) d\mathbf{f}$$

This expression implicitly depends on the hyper parameters, but $\mathbf{y}$ and X are given from the training data. It can be computed in closed form, as all terms are Gaussians.

We take the logarithm, compute the derivative and set it to 0. This is the **training** step.

Estimating the Hyperparameters

To find optimal hyper parameters we need the **marginal likelihood:**

$$p(\mathbf{y} \mid X) = \frac{1}{\sqrt{(2\pi)^n |K|}} \exp\left(-\frac{1}{2}\mathbf{y}^T K^{-1} \mathbf{y}\right)$$

Estimating the Hyperparameters

To find optimal hyper parameters we need the **marginal likelihood:**

$$p(\mathbf{y} \mid X) = \frac{1}{\sqrt{(2\pi)^n |K|}} \exp\left(-\frac{1}{2}\mathbf{y}^T K^{-1}\mathbf{y}\right)$$

$$\log p(\mathbf{y} \mid X) = -\frac{1}{2}\log((2\pi)^n |K|) - \frac{1}{2}\mathbf{y}^T K^{-1}\mathbf{y}$$

Estimating the Hyperparameters

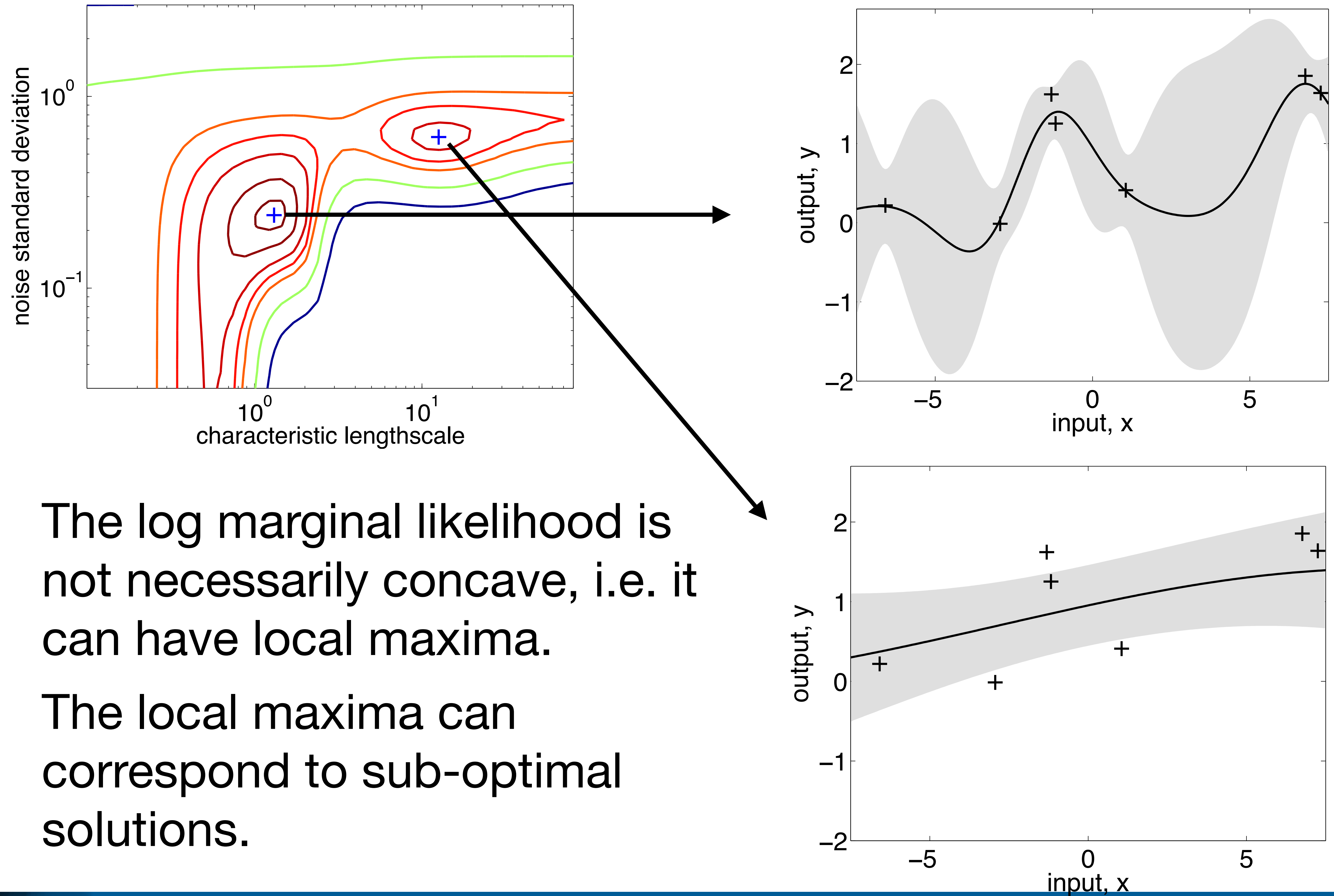
To find optimal hyper parameters we need the **marginal likelihood:**

$$p(\mathbf{y} \mid X) = \frac{1}{\sqrt{(2\pi)^n |K|}} \exp\left(-\frac{1}{2}\mathbf{y}^T K^{-1} \mathbf{y}\right)$$

$$\log p(\mathbf{y} \mid X) = -\frac{1}{2}\log((2\pi)^n |K|) - \frac{1}{2}\mathbf{y}^T K^{-1} \mathbf{y}$$

$$\frac{\partial \log p(\mathbf{y} \mid X)}{\partial \theta_i} = \frac{1}{2}\mathbf{y}^T K^{-1} \frac{\partial K}{\partial \theta_i} \mathbf{y} - \frac{1}{2}\mathrm{tr}\left(K^{-1}\frac{\partial K}{\partial \theta_i}\right)$$

Estimating the Hyperparameters



The log marginal likelihood is not necessarily concave, i.e. it can have local maxima.

The local maxima can correspond to sub-optimal solutions.

Automatic Relevance Determination

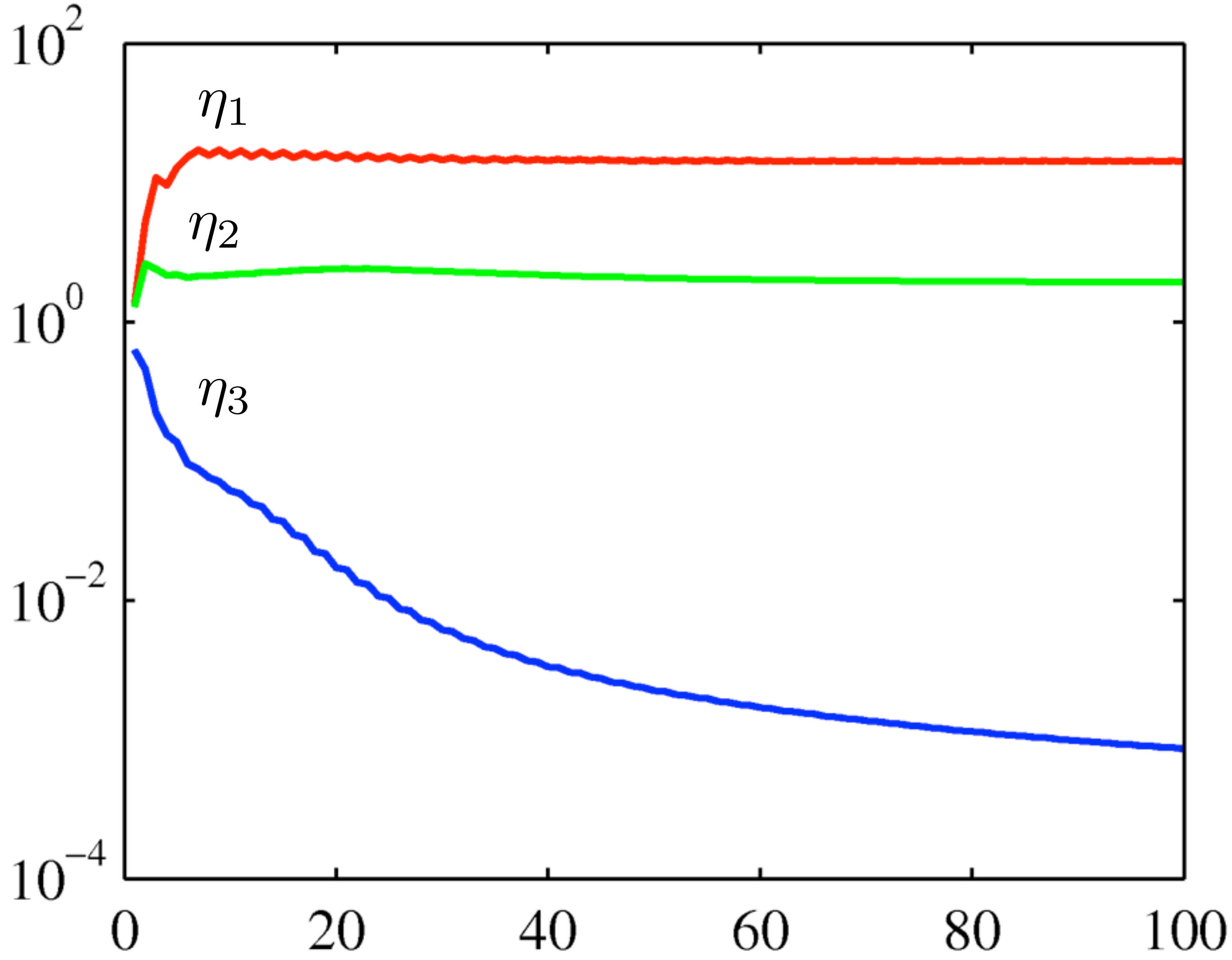
- We have seen how the covariance function can be generalized using a matrix M
- If M is diagonal this results in the kernel function

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f \exp\left(\frac{1}{2}\sum_{i=1}^{D} \eta_i (x_i - x_i')^2\right)$$

- We can interpret the η_i as weights for each feature dimension
- Thus, if the length scale $l_i = 1/\eta_i$ of an input dimension is large, the input is less relevant
- During training this is done automatically

Automatic Relevance Determination

3-dimensional data, parameters η_1 η_2 η_3 as they evolve during training

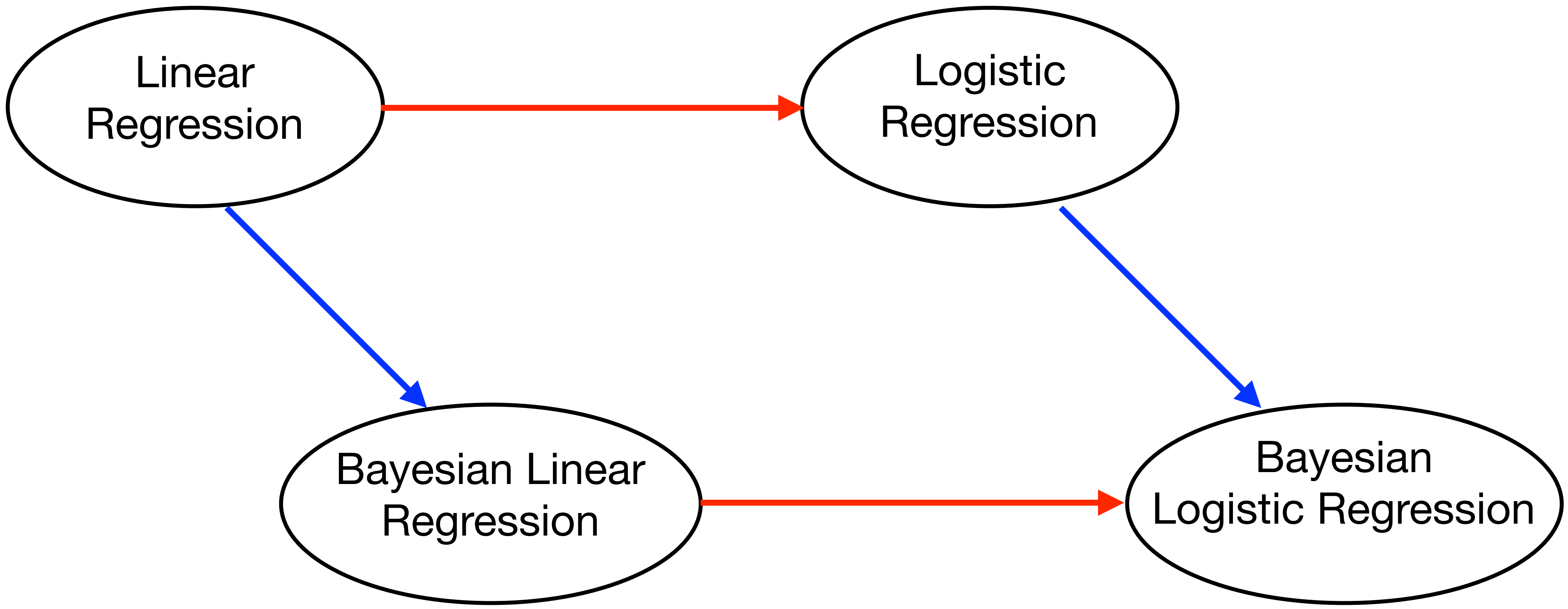


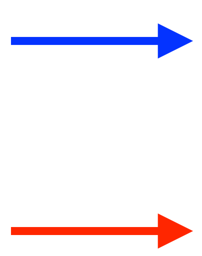
During the optimization process to learn the hyper-parameters, the reciprocal length scale for one parameter decreases, i.e.:

This hyper parameter is not very relevant!

Gaussian Processes - Classification

Remember the Visualisation

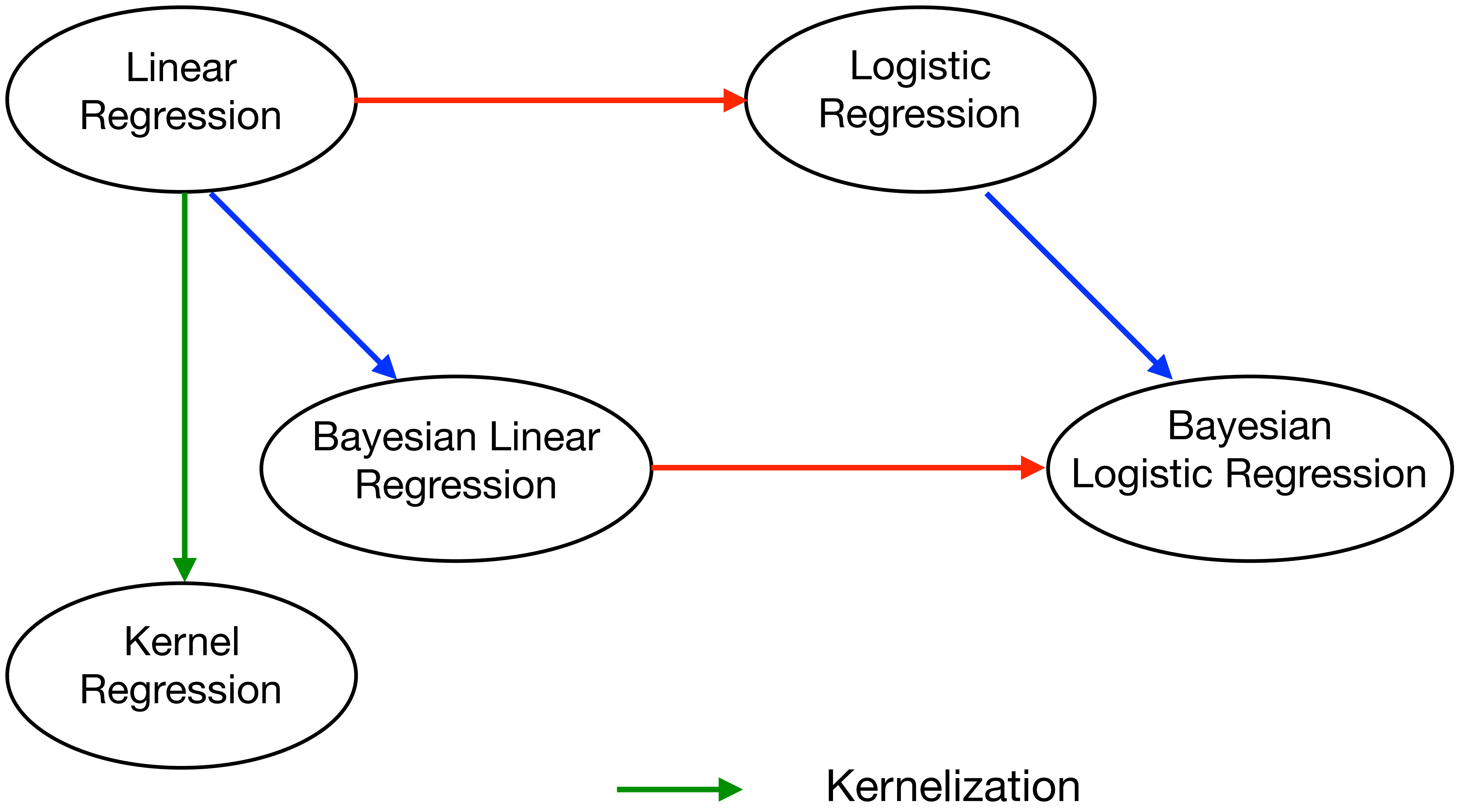




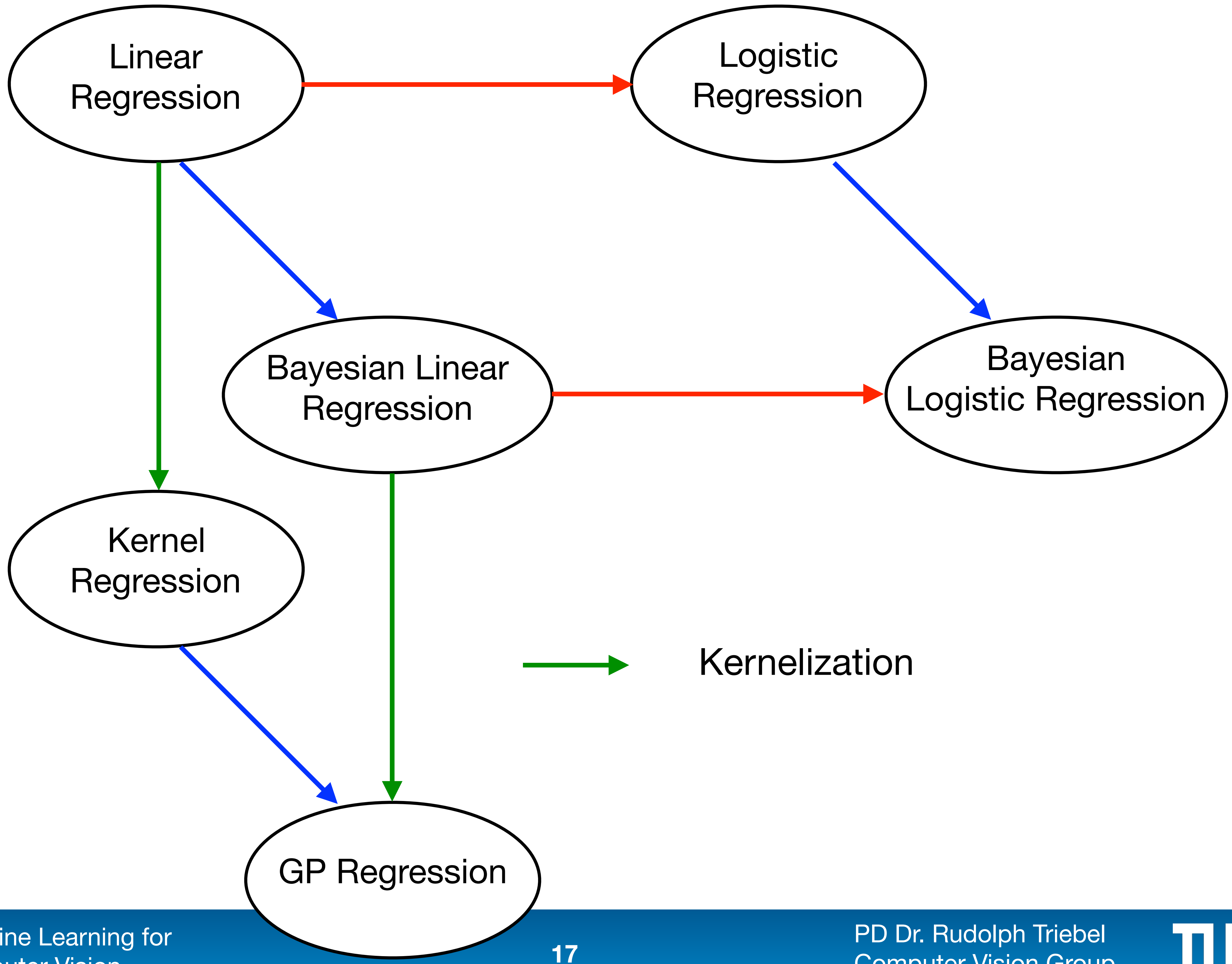
probabilistic reasoning

from regression to classification

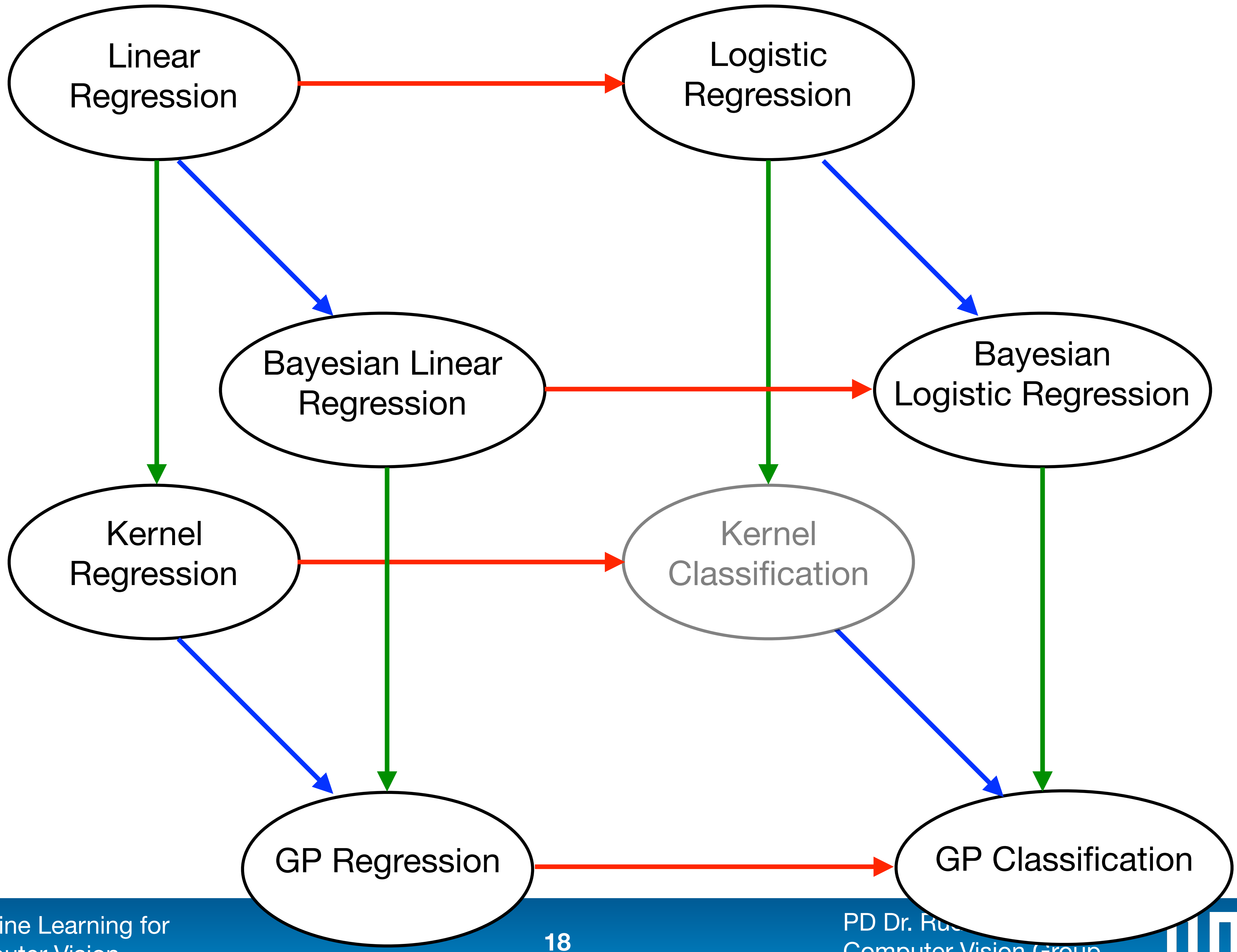
Kernelization as a New Dimension



Kernelization as a New Dimension



Kernelization as a New Dimension



Gaussian Processes For Classification

In regression we have $y \in \mathbb{R}$, in binary classification we have $y \in \{-1; 1\}$

To use a GP for classification, we can apply a **sigmoid** function to the posterior obtained from the GP and compute the class probability as:

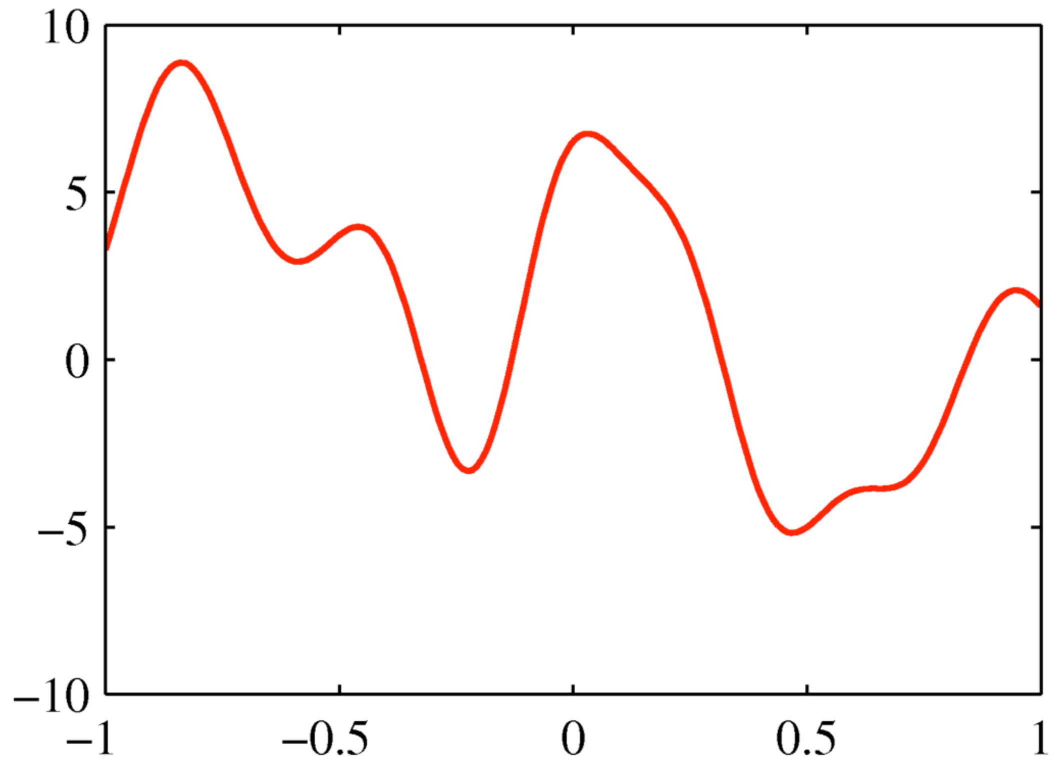
$$p(y = +1 \mid \mathbf{x}) = \sigma(f(\mathbf{x}))$$

If the sigmoid function is symmetric: $\sigma(-z) = 1 - \sigma(z)$ then we have $p(y \mid \mathbf{x}) = \sigma(yf(\mathbf{x}))$.

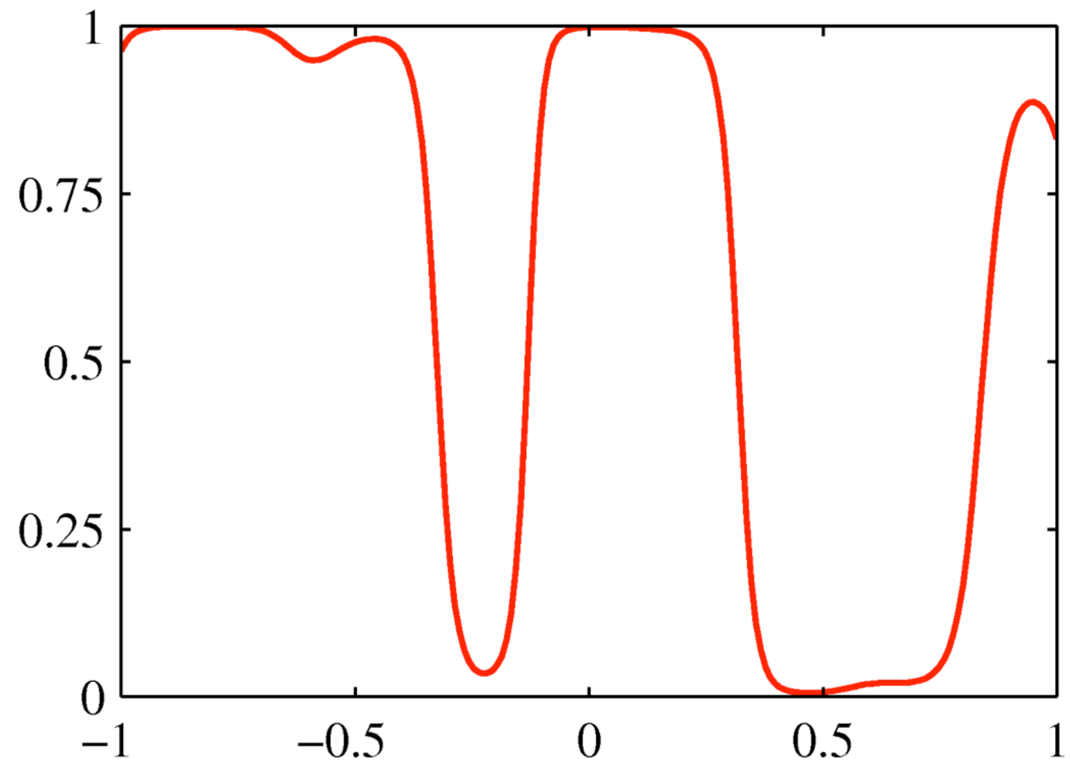
A typical type of sigmoid function is the logistic sigmoid:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Application of the Sigmoid Function



Function sampled from a Gaussian Process

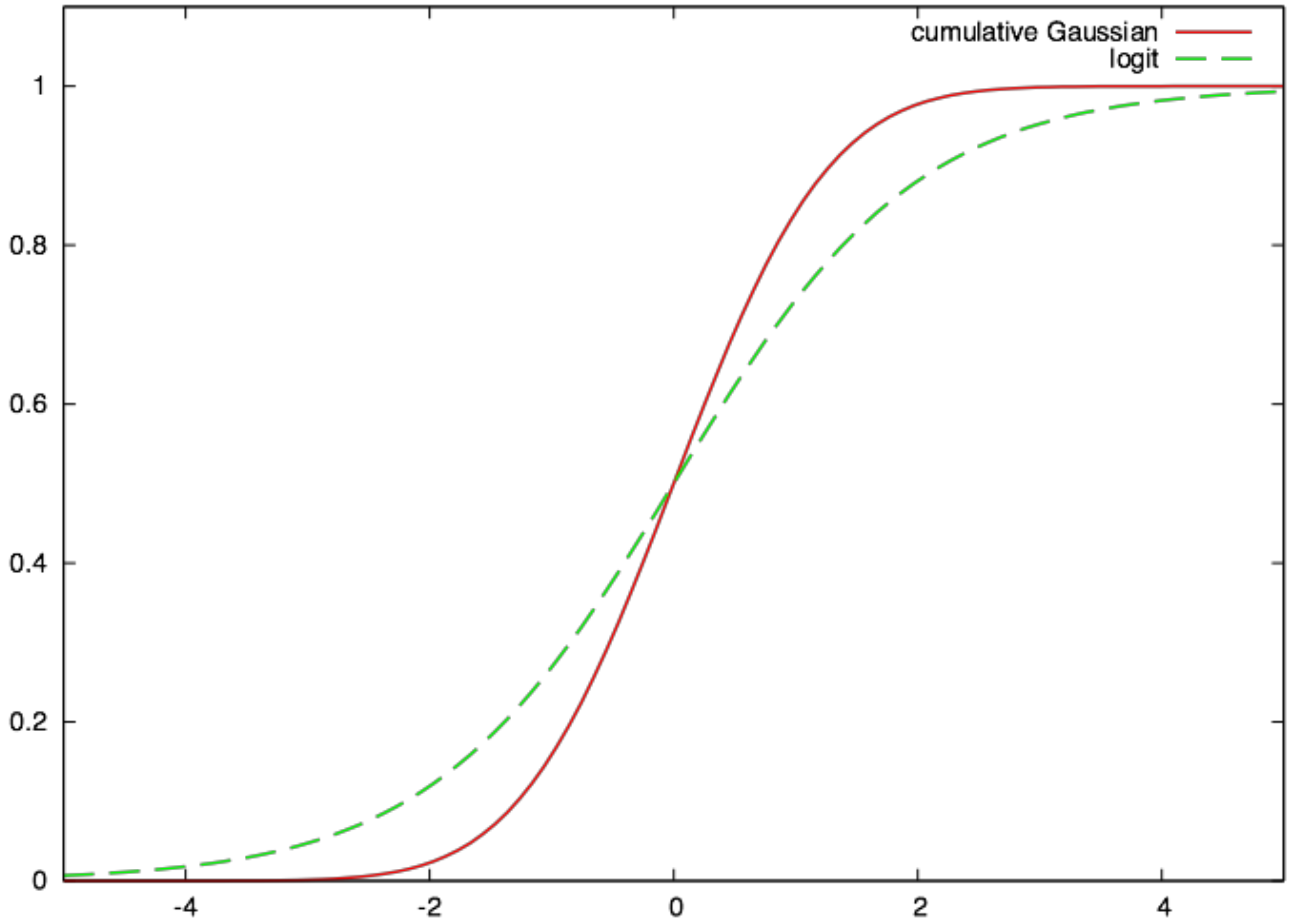


Sigmoid function applied to the GP function

Another symmetric sigmoid function is the **cumulative Gaussian:**

$$\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(x \mid 0, 1) dx$$

Visualization of Sigmoid Functions



The cumulative Gaussian is slightly steeper than the logistic sigmoid

The Latent Variables

In regression, we directly estimated f as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

and values of f where observed in the training data. Now only labels +1 or -1 are observed and f is treated as a set of **latent variables.**

A major advantage of the Gaussian process classifier over other methods is that it **marginalizes** over all latent functions rather than maximizing some model parameters.

Class Prediction with a GP

The aim is to compute the predictive distribution

$$p(y_* = +1 \mid X, \mathbf{y}, \mathbf{x}_*) = \int p(y_* \mid f_*) p(f_* \mid X, \mathbf{y}, \mathbf{x}_*) df_*$$

$\sigma(f_*)$

Class Prediction with a GP

The aim is to compute the predictive distribution

$$p(y_* = +1 \mid X, \mathbf{y}, \mathbf{x}_*) = \int p(y_* \mid f_*) p(f_* \mid X, \mathbf{y}, \mathbf{x}_*) df_*$$

we marginalize over the latent variables from the training data:

$$p(f_* \mid X, \mathbf{y}, \mathbf{x}_*) = \int p(f_* \mid X, \mathbf{x}_*, \mathbf{f}) p(\mathbf{f} \mid X, \mathbf{y}) d\mathbf{f}$$

predictive distribution of the latent variable (from regression)

Class Prediction with a GP

The aim is to compute the predictive distribution

$$p(y_* = +1 \mid X, \mathbf{y}, \mathbf{x}_*) = \int p(y_* \mid f_*) p(f_* \mid X, \mathbf{y}, \mathbf{x}_*) df_*$$

we marginalize over the latent variables from the training data:

$$p(f_* \mid X, \mathbf{y}, \mathbf{x}_*) = \int p(f_* \mid X, \mathbf{x}_*, \mathbf{f}) p(\mathbf{f} \mid X, \mathbf{y}) d\mathbf{f}$$

we need the posterior over the latent variables:

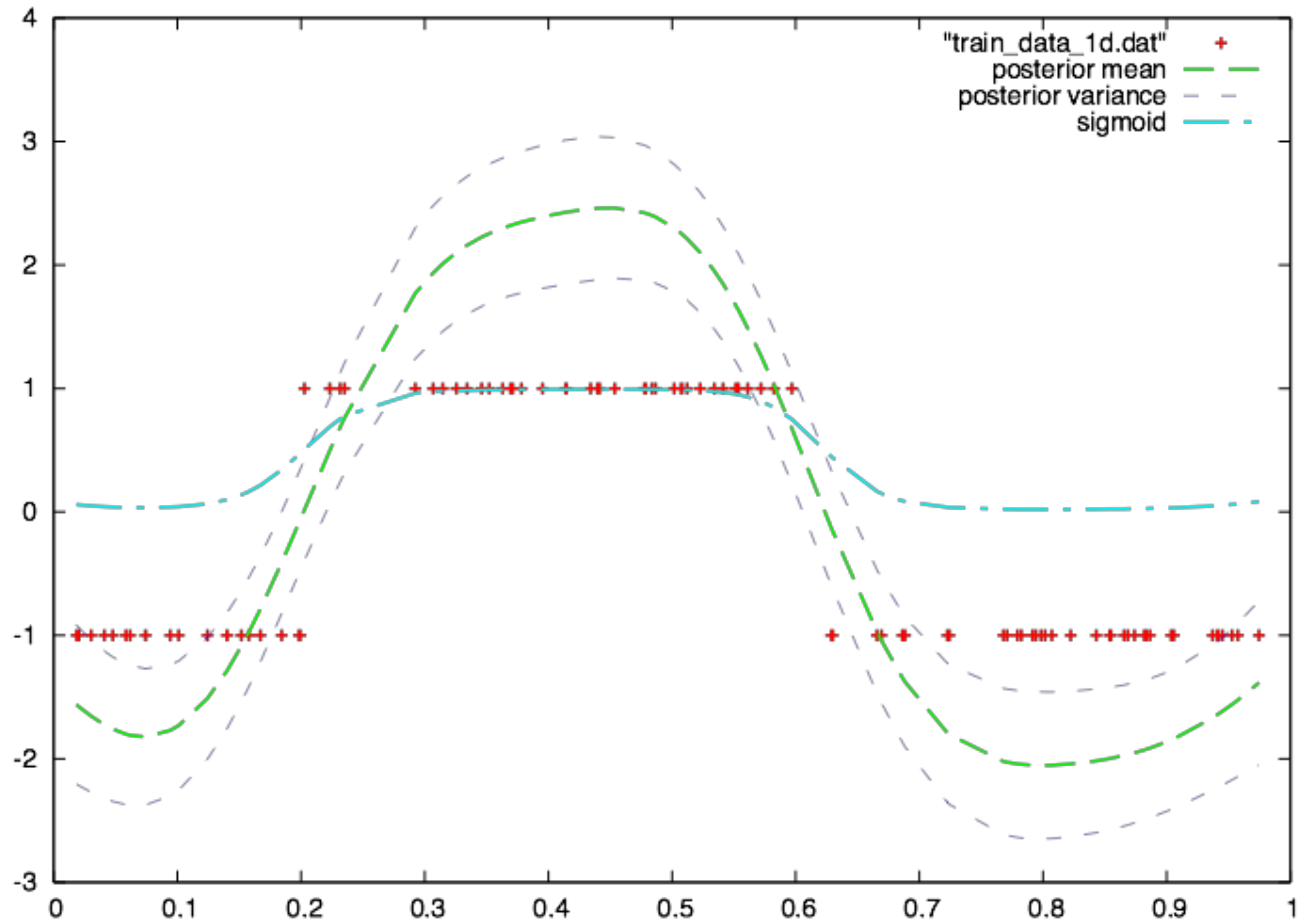
$$p(\mathbf{f} \mid X, \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid X)}{p(\mathbf{y} \mid X)}$$

likelihood (sigmoid)

prior

normalizer

A Simple Example



- Red: Two-class training data
- Green: mean function of $p(\mathbf{f} \mid X, \mathbf{y})$
- Light blue: sigmoid of the mean function

But There Is A Problem...

$$p(\mathbf{f} \mid X, \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid X)}{p(\mathbf{y} \mid X)}$$

- The likelihood term is not a Gaussian!
- This means, we can not compute the posterior in closed form.
- There are several different solutions in the literature, e.g.:
    - Laplace approximation
    - Expectation Propagation
    - Variational methods

Laplace Approximation

Consider a general distribution

$$p(z) = \frac{1}{Z} f(z) \quad \text{where} \quad Z = \int f(z) dz$$

Laplace Approximation

Consider a general distribution

$$p(z) = \frac{1}{Z} f(z) \quad \text{where} \quad Z = \int f(z) dz$$

Aim: approximate this with a normal distribution

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \hat{\mathbf{f}}, A^{-1})$$

$$\mathbf{f}_{new} = \mathbf{f} - (\nabla\nabla\Psi)^{-1}\nabla\Psi$$

$$A = K^{-1} + W$$

Laplace Approximation

$$p(\mathbf{f} \mid X, \mathbf{y}) \approx q(\mathbf{f} \mid X, \mathbf{y}) = \mathcal{N}(\mathbf{f} \mid \hat{\mathbf{f}}, A^{-1})$$

where $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f} \mid X, \mathbf{y})$

and $A = -\nabla\nabla \log p(\mathbf{f} \mid X, \mathbf{y})|_{\mathbf{f}=\hat{\mathbf{f}}}$

second-order Taylor expansion

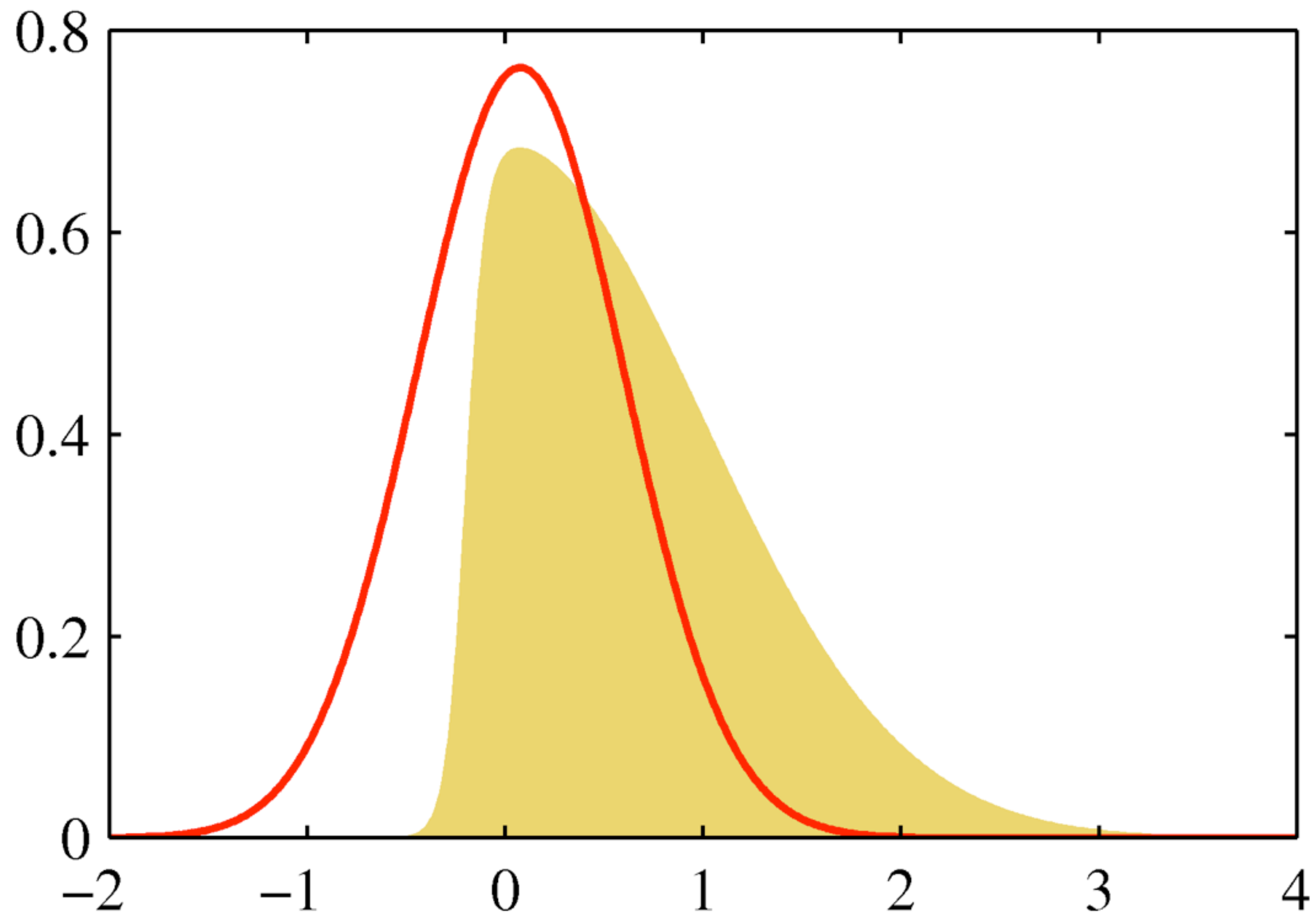
To compute $\hat{\mathbf{f}}$ an iterative approach using Newton's method has to be used.

The Hessian matrix A can be computed as

$$A = K^{-1} + W$$

where $W = -\nabla\nabla \log p(\mathbf{y} \mid \mathbf{f})$ is a diagonal matrix which depends on the sigmoid function.

Laplace Approximation



- Yellow: a non-Gaussian posterior
- Red: a Gaussian approximation, the mean is the mode of the posterior, the variance is the negative second derivative at the mode

Predictions

Now that we have $p(\mathbf{f} \mid X, \mathbf{y})$ we can compute:

$$p(f_* \mid X, \mathbf{y}, \mathbf{x}_*) = \int p(f_* \mid X, \mathbf{x}_*, \mathbf{f}) p(\mathbf{f} \mid X, \mathbf{y}) d\mathbf{f}$$

From the regression case we have:

$$p(f_* \mid X, \mathbf{x}_*, \mathbf{f}) = \mathcal{N}(f_* \mid \mu_*, \Sigma_*)$$

where $\mu_* = \mathbf{k}_*^T K^{-1} \mathbf{f}$ $\qquad \Sigma_* = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T K^{-1} \mathbf{k}_*$

Linear in $\mathbf{f}$

This reminds us of a property of Gaussians that we saw earlier!

Gaussian Properties (Rep.)

If we are given this:

I. $$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \mu, \Sigma_1)$$

II. $$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}(\mathbf{y} \mid A\mathbf{x} + \mathbf{b}, \Sigma_2)$$

Then it follows (properties of Gaussians):

III. $$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid A\mu + \mathbf{b}, \Sigma_2 + A\Sigma_1 A^T)$$

IV. $$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}(\mathbf{x} \mid \Sigma(A^T\Sigma_2^{-1}(\mathbf{y} - \mathbf{b}) + \Sigma_1^{-1}\mathbf{y}), \Sigma)$$

where

$$\Sigma = (\Sigma_1^{-1} + A^T\Sigma_s^{-1}A)^{-1}$$

Applying this to Laplace

$$\mathbb{E}[f_* \mid X, \mathbf{y}, \mathbf{x}_*] = \mathbf{k}(\mathbf{x}_*)^T K^{-1} \hat{\mathbf{f}}$$

$$\mathbb{V}[f_* \mid X, \mathbf{y}, \mathbf{x}_*] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (K + W^{-1})^{-1} \mathbf{k}_*$$

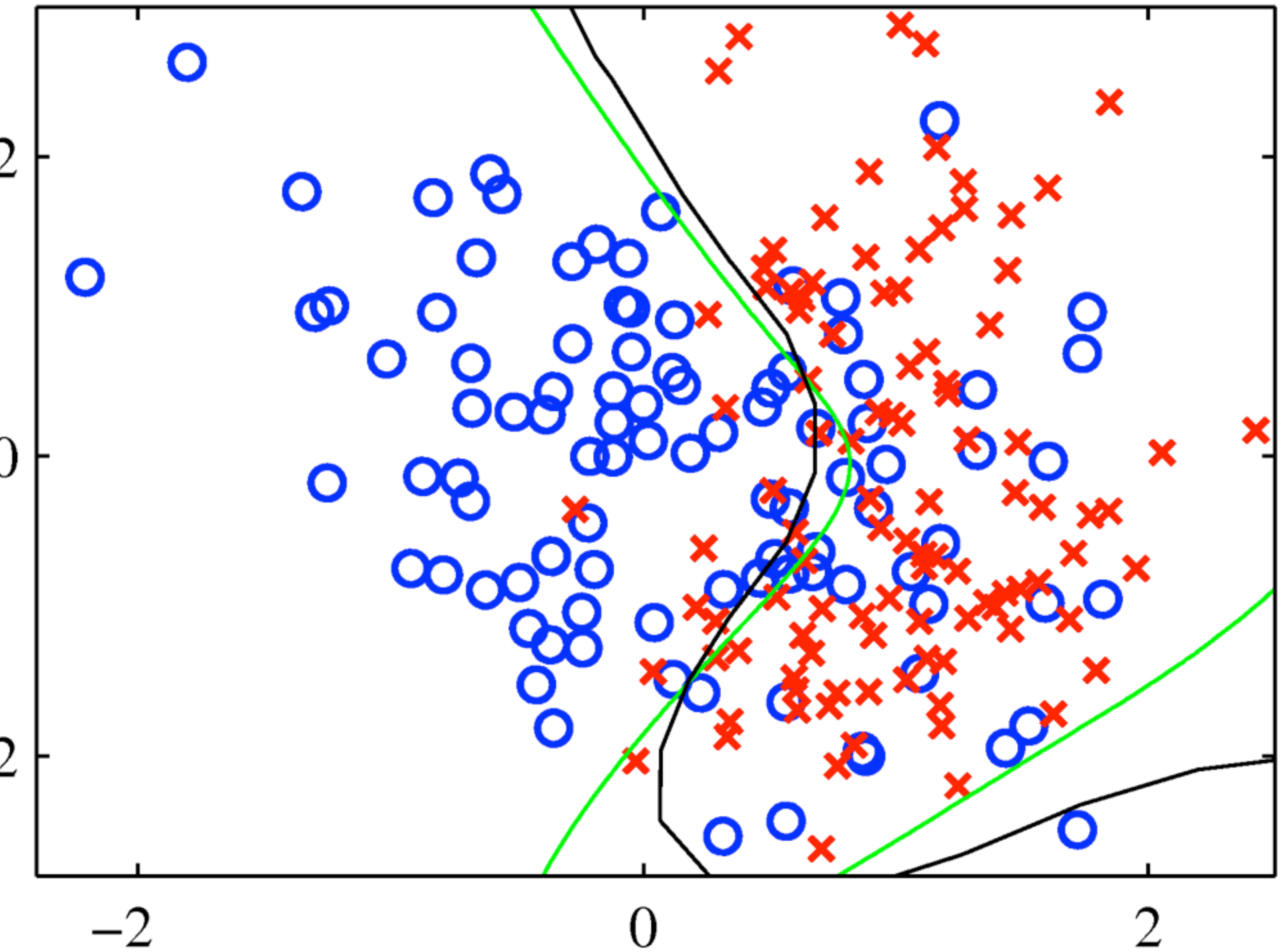
It remains to compute

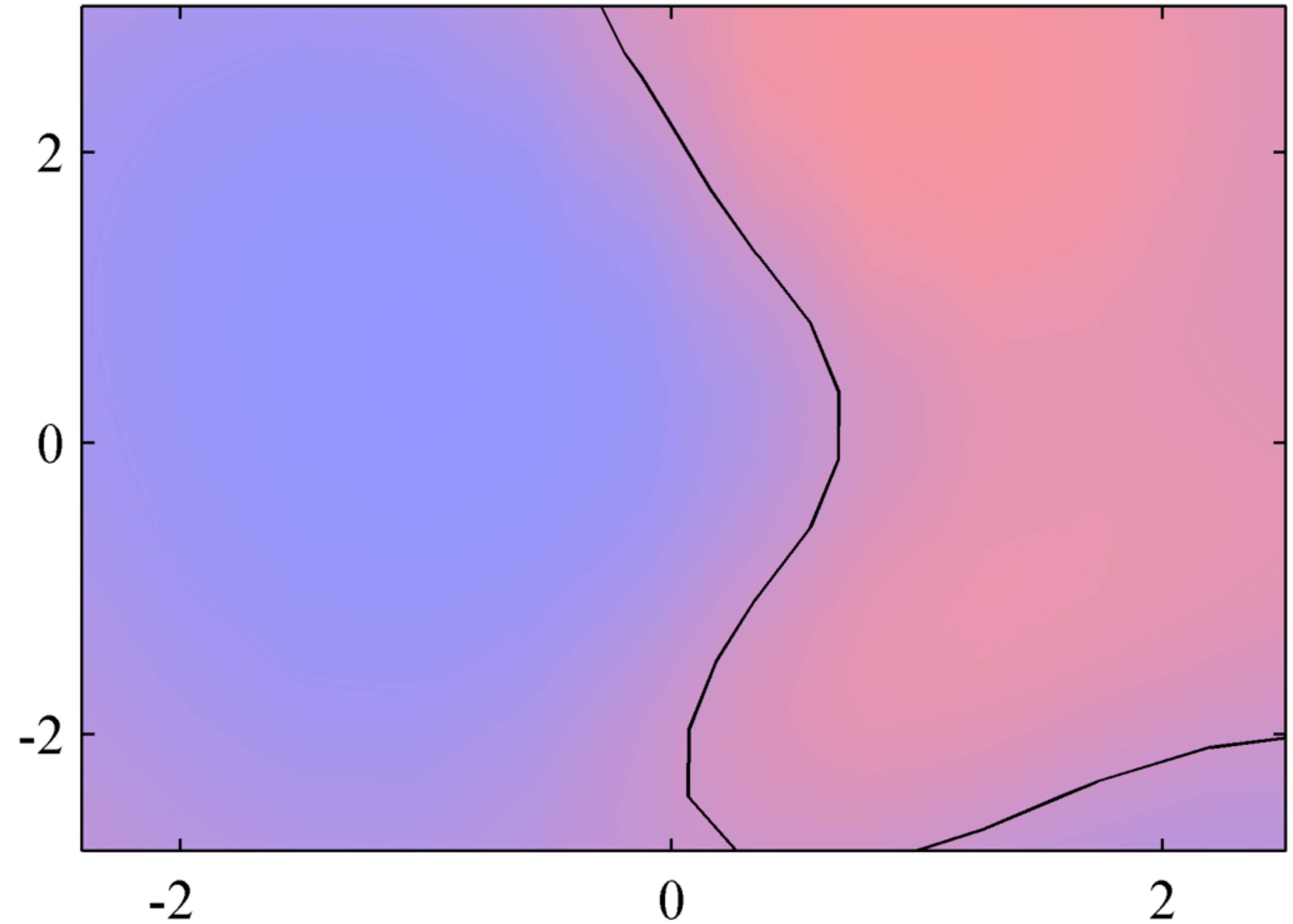
$$p(y_* = +1 \mid X, \mathbf{y}, \mathbf{x}_*) = \int p(y_* \mid f_*) p(f_* \mid X, \mathbf{y}, \mathbf{x}_*) df_*$$

Depending on the kind of sigmoid function we

- can compute this in closed form (cumulative Gaussian sigmoid)
- have to use sampling methods or analytical approximations (logistic sigmoid)

A Simple Example





- Two-class problem (training data in red and blue)
- Green line: optimal decision boundary
- Black line: GP classifier decision boundary
- Right: posterior probability

Summary

- Kernel methods solve problems by implicitly mapping the data into a (high-dimensional) feature space
- The feature function itself is not used, instead the algorithm is expressed in terms of the kernel
- Gaussian Processes are Normal distributions over functions
- To specify a GP we need a covariance function (kernel) and a mean function
- More on Gaussian Processes: http://videolectures.net/epsrcws08_rasmussen_lgp/