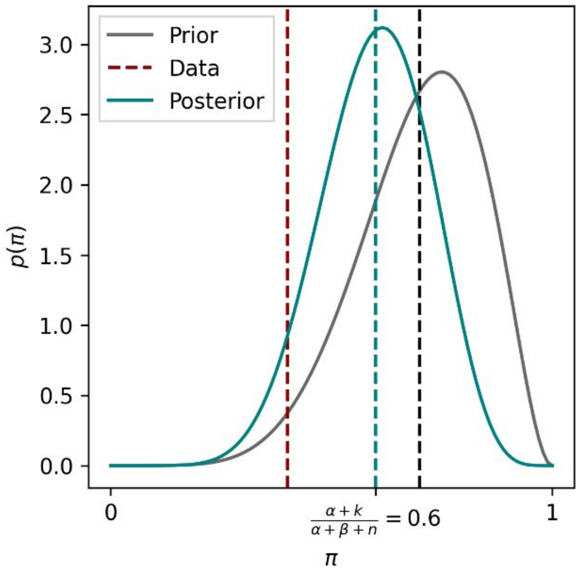
CS 480/680 Introduction to Machine Learning

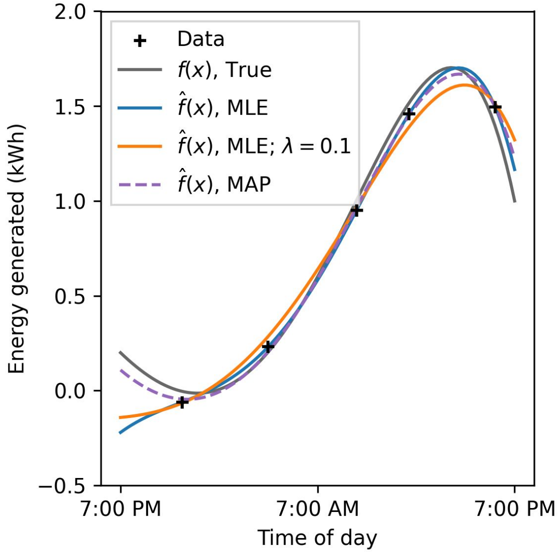
Lecture 9b
Gaussian Processes

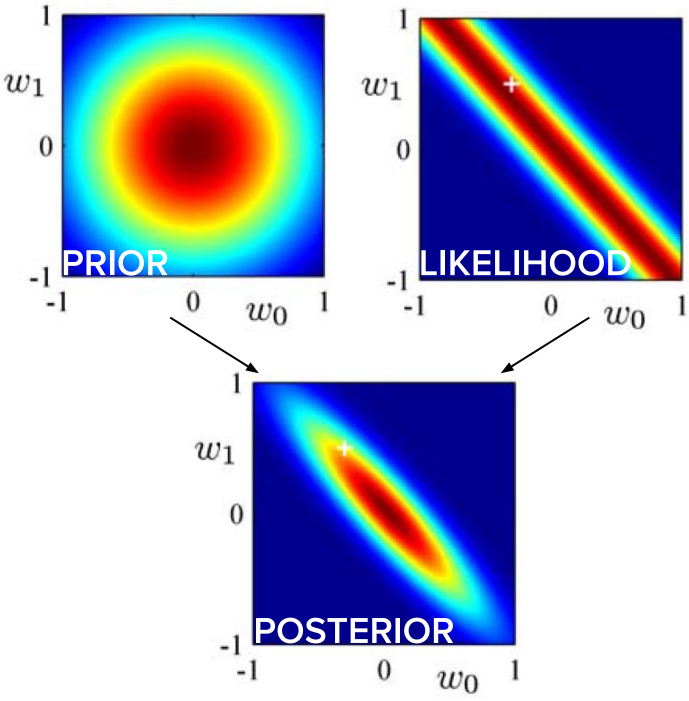
Kathryn Simone
10 October 2024

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

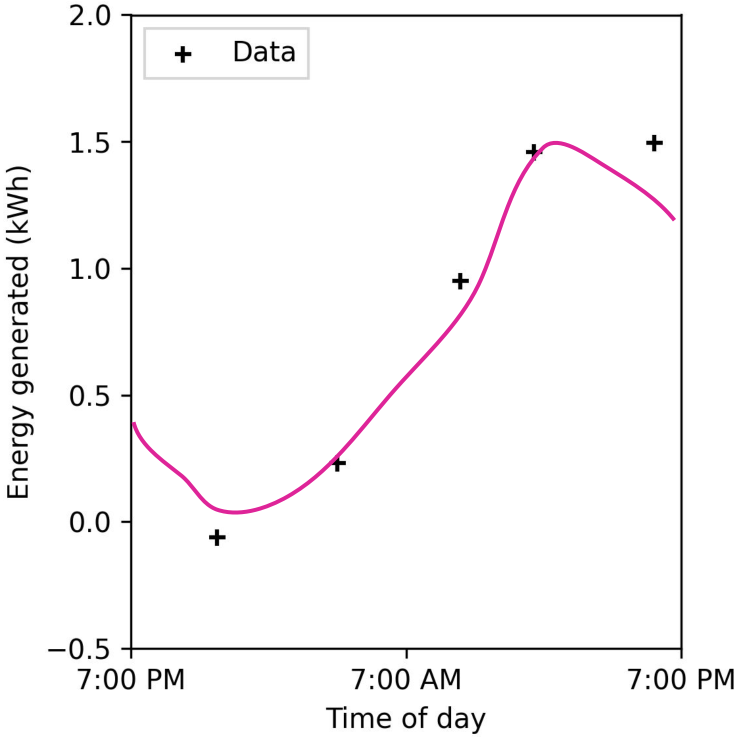
Incorporating prior knowledge as a form of regularization







An assumed set of basis functions can be limiting



Key Questions

9a

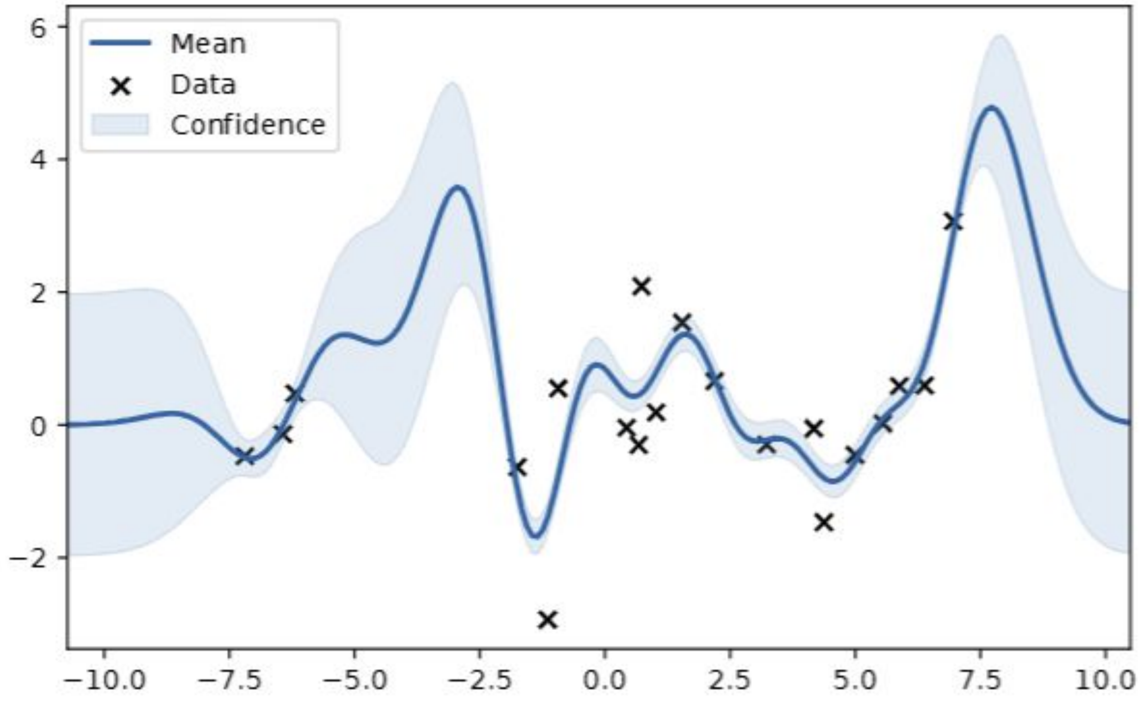
I. How can we incorporate prior knowledge into a model?

II. How can we account for uncertainty in parameters?

9b

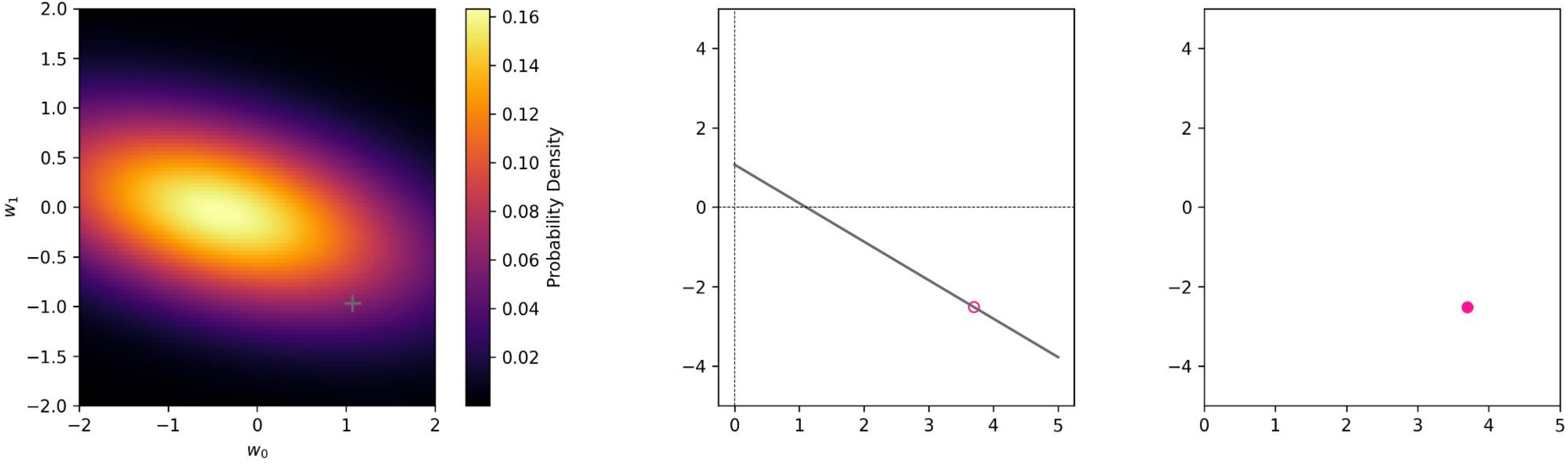
III. What if we don't even know the structure of a model?

Behavior of the Gaussian Process model

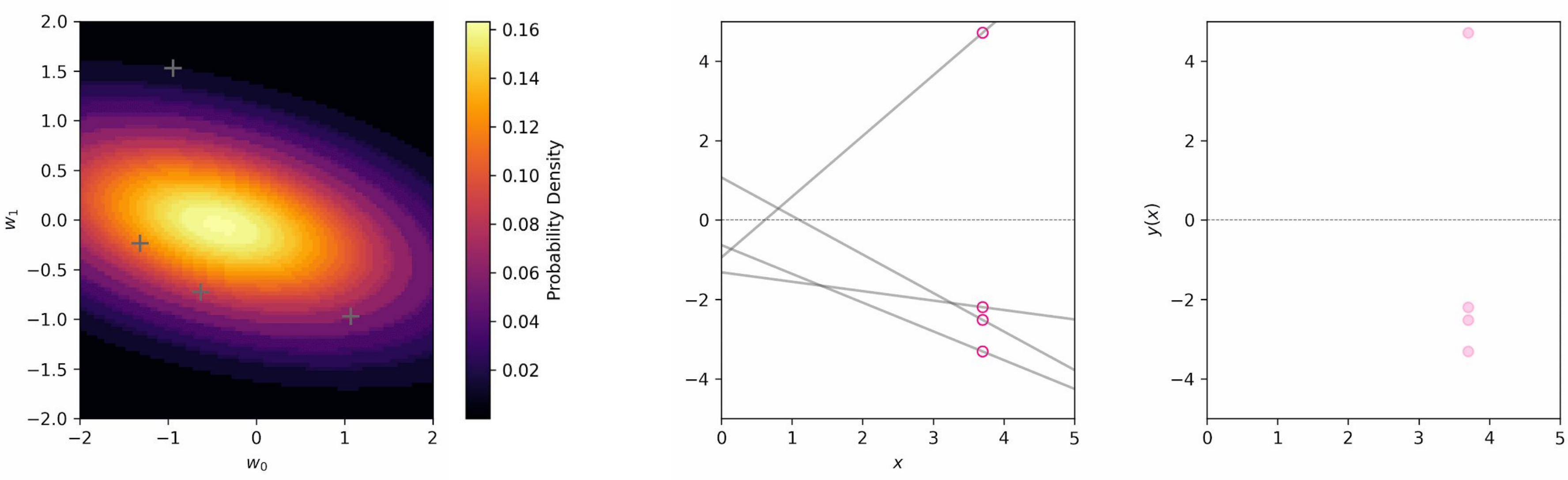


PML, Section 17.2

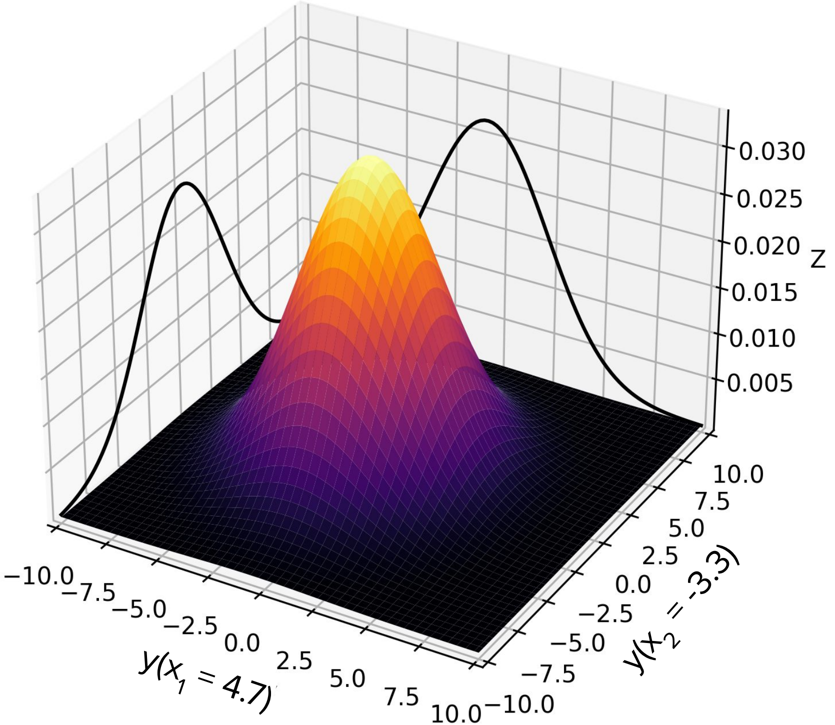
A sample from the posterior distribution defines one function



A Gaussian distribution over the *weights* induces a Gaussian distribution over *functions*



The *joint* distribution over *functions* is a Gaussian



Gaussian Processes are specified completely by the mean and covariance

$$
\begin{aligned}
y(x) &= \mathbf{w}^\top \boldsymbol{\phi}(x) \\
\mathbf{y} &= \boldsymbol{\Phi}\mathbf{w} \\
\mathbb{E}[\mathbf{y}] &= \boldsymbol{\Phi}\mathbb{E}[\mathbf{w}] = \mathbf{0} \\
\operatorname{Cov}[\mathbf{y}] &= \mathbb{E}\left[\mathbf{y}\mathbf{y}^\top\right] \\
&= \boldsymbol{\Phi}\mathbb{E}\left[\mathbf{w}\mathbf{w}^\top\right]\boldsymbol{\Phi}^\top \\
&= \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top \\
&= \alpha^{-1}\mathbf{K}
\end{aligned}
$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$K_{ij} = k(x_i, x_j) = \boldsymbol{\phi}(x_i)^\top \boldsymbol{\phi}(x_j)$$

$$\boxed{\mathbb{E}\left[y(x_i)y(x_j)\right] = \alpha^{-1}k(x_i, x_j)}$$

PRML, Section 6.4

A Gaussian Process model for regression

$$t_i = y_i + \epsilon_i$$

$$\begin{aligned} p(\mathbf{t} \mid \mathbf{y}) &= \mathcal{N}(\mathbf{t} \mid \mathbf{y}, \beta^{-1}\mathbf{I}_N) \\ p(\mathbf{y}) &= \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \alpha^{-1}\mathbf{K}) \\ p(\mathbf{t}) &= \int p(\mathbf{t} \mid \mathbf{y}) p(\mathbf{y}) \, d\mathbf{y} \\ &= \mathcal{N}(\mathbf{t} \mid \mathbf{0}, \mathbf{C}) \end{aligned}$$

$$\mathbf{C}_{ij} = \alpha^{-1} k(x_i, x_j) + \beta^{-1} \delta_{ij}$$

PRML, Section 6.4

Augmenting the covariance matrix for prediction

$$p(\boldsymbol{t}_{N+1}) \sim \mathcal{N}(\boldsymbol{t}_{N+1} \mid 0, C_{N+1})$$

$$C_{N+1} = \begin{pmatrix} C_N & \mathbf{k} \\ \mathbf{k}^\top & c \end{pmatrix}$$

$$\mathbf{k} = \begin{bmatrix} \alpha^{-1}k(x_1, x_{N+1}) \\ \alpha^{-1}k(x_2, x_{N+1}) \\ \vdots \\ \alpha^{-1}k(x_N, x_{N+1}) \end{bmatrix} \quad \text{for } i = 1, \ldots, N$$

$$c = \alpha^{-1}k(x_{N+1}, x_{N+1}) + \beta^{-1}$$

PRML, Section 6.4

The predictive distribution for Gaussian process regression

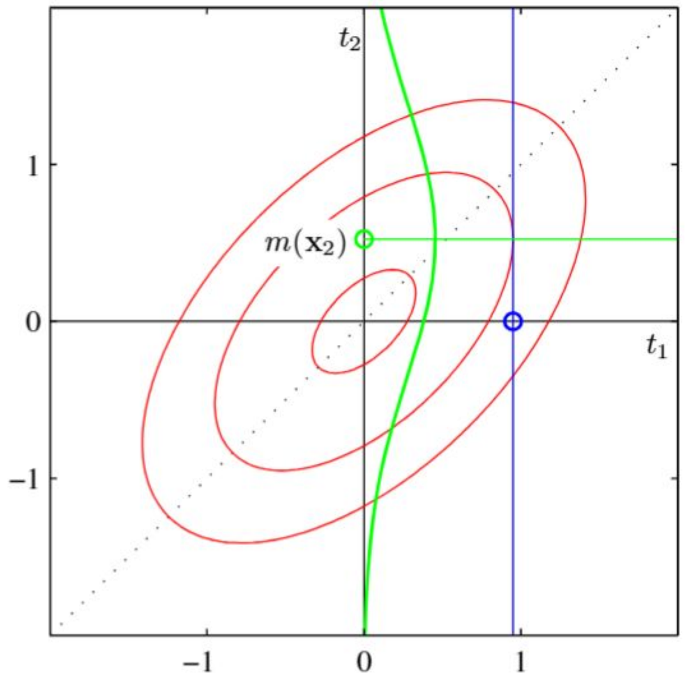
$$p(t_{N+1} \mid \mathbf{t}) \sim \mathcal{N}(\mu_{x_{N+1}}, \sigma^2_{x_{N+1}}),$$

$$\mu_{x_{N+1}} = \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{t}$$

$$\mu_{x_{N+1}} = \sum_{i=1}^{N} a_i k(x_i, x_{N+1})$$

where a_i are the elements of the N-vector $\boldsymbol{a} = \mathbf{C}_N^{-1} \boldsymbol{t}$.

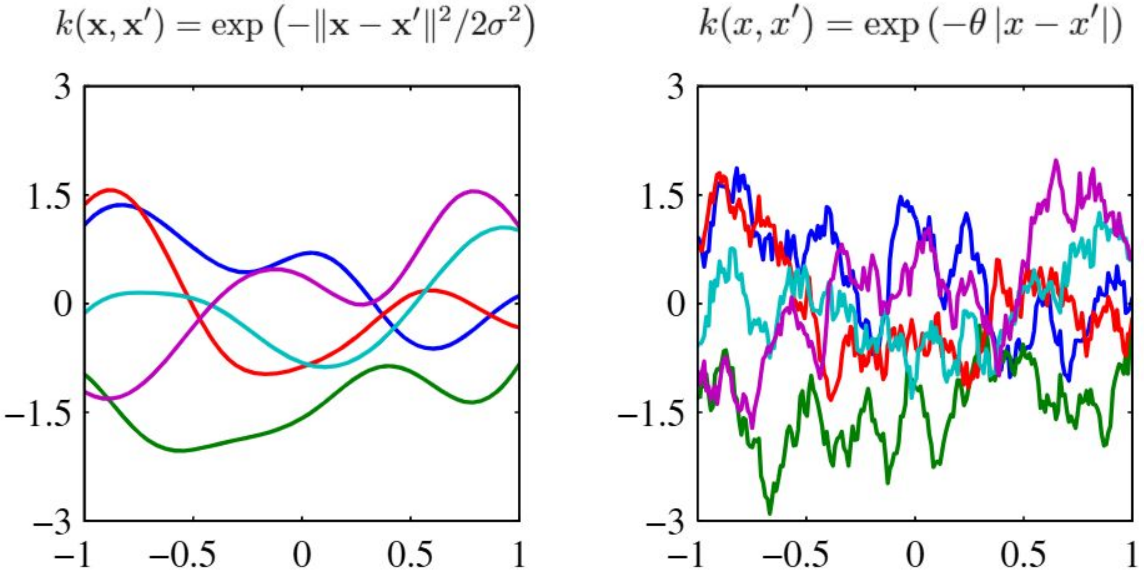
$$\sigma^2_{x_{N+1}} = c - \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{k}$$



PRML, Section 6.4

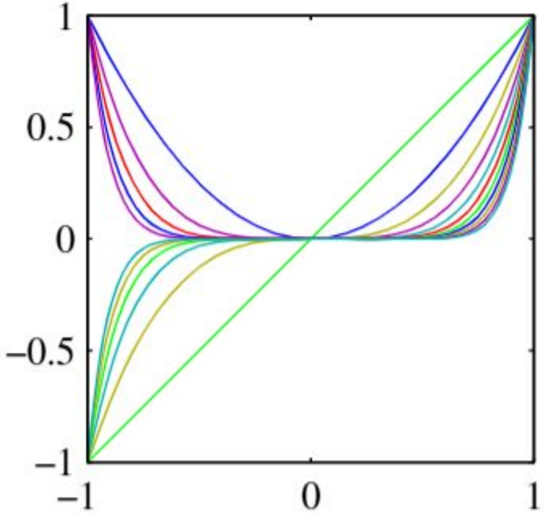

THE POSTERIOR FOLLOWS A GAUSSIAN
THE FUNCTION VALUES FOLLOW A GAUSSIAN
THE NOISE MODEL IS A GAUSSIAN
EVERYTHING FOLLOWS A GAUSSIAN
imgflip.com

The kernel function itself need not be Gaussian



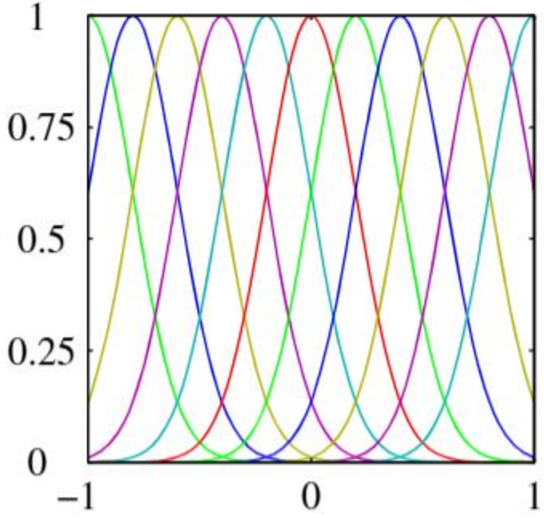
PRML, Section 6.4

Fixed nonlinear basis functions in regression are special cases of Gaussian Processes



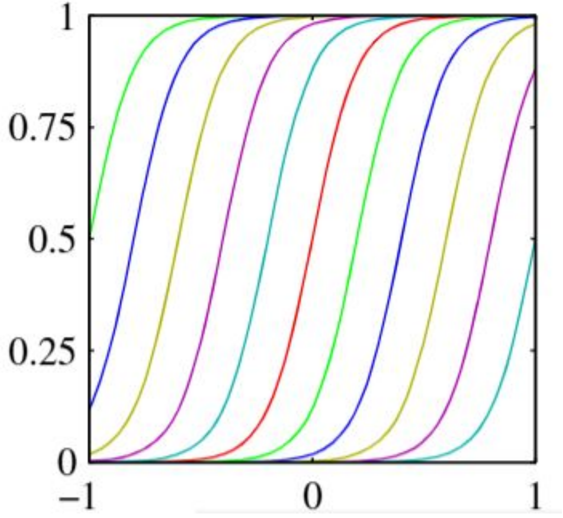
Polynomial:

$$\{\phi_j(\boldsymbol{x})\} = \{x^j\}$$



Gaussian:

$$\{\phi_j(\boldsymbol{x})\} = \{e^{-\frac{(x-\mu_j)^2}{2\sigma^2}}\}$$



Sigmoidal:

$$\{\phi_j(\boldsymbol{x})\} = \{\frac{1}{1+e^{\frac{-(x-\mu_j)}{\sigma}}}\}$$

Pattern Recognition and Machine Learning, Section 3.1

Computational considerations for regression

	Train	Test (one new point)
Fixed Basis Functions (M features)	$O(M^3)$	$O(M^2)$
Gaussian Process (N data points)	$O(N^3)$	$O(N^2)$

Now that we're at the end of the lecture, you should be able to...

- ★ Apply **Bayesian updating** to determine the **posterior distribution** of parameters, from the likelihood and a given prior.
- ★ **Design suitable priors** to reflect domain knowledge and serve as a form of regularization.
- ★ Use **maximum a posteriori** to incorporate priors on the weights in regression in a data-scarce applications involving domain knowledge.
- ★ **Interpret ridge regression** as a imposing a prior on the distribution of weights.
- ★ Given the expression for the mean and covariance of the predictive distribution, and a particular kernel function, **compute and/or sketch the prediction** of the GP solution for a test input.
- ★ Select and defend the choice of using either Gaussian Process regression or Bayesian Linear Regression, taking into account the **tradeoff between computational complexity and flexibility** of the model.

Errata

- On slide 8, the labels on the x- and y-axes appeared as $\mathbf{y(x_1) = 4.7}$ and $\mathbf{y(x_2) = -3.3}$, respectively. These have been corrected to $\mathbf{y(x_1 = 4.7\ )}$ and $\mathbf{y(x_2 = -3.3)}$.