

CS 480/680

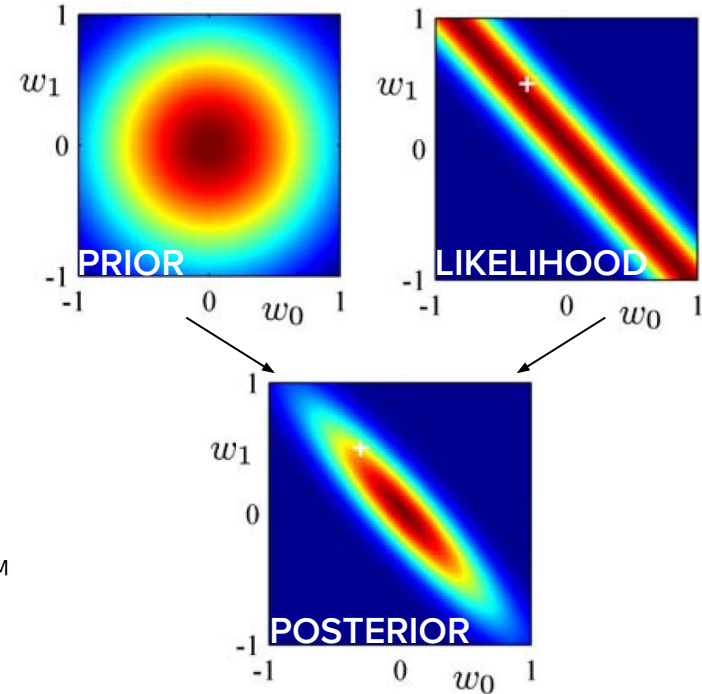
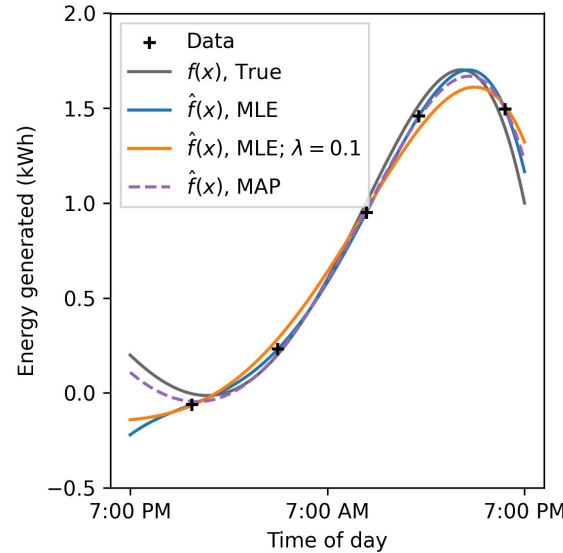
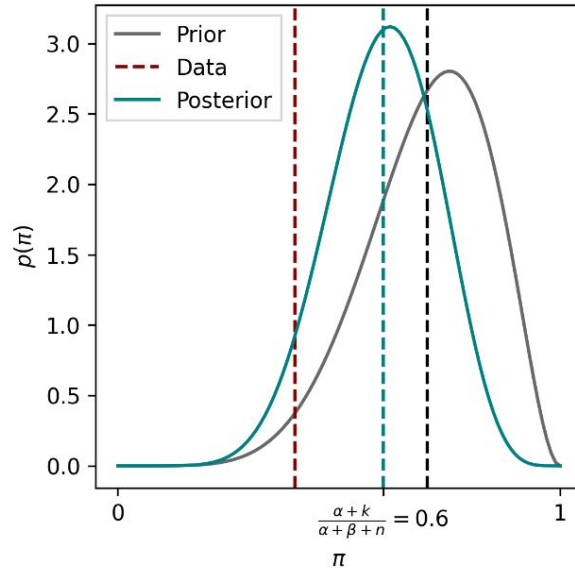
Introduction to Machine Learning

Lecture 9b

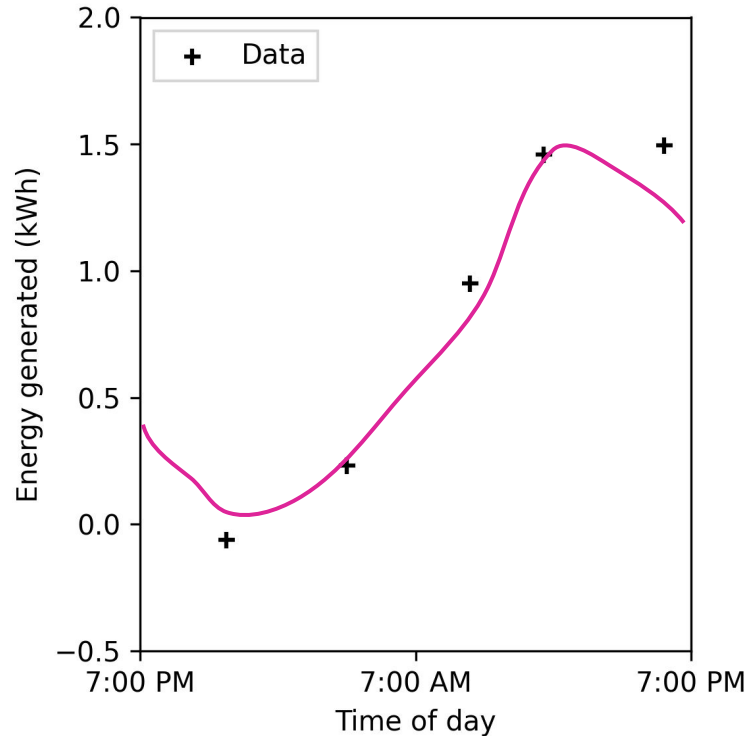
Gaussian Processes

Kathryn Simone
10 October 2024

Incorporating prior knowledge as a form of regularization



An assumed set of basis functions can be limiting



Key Questions

9a

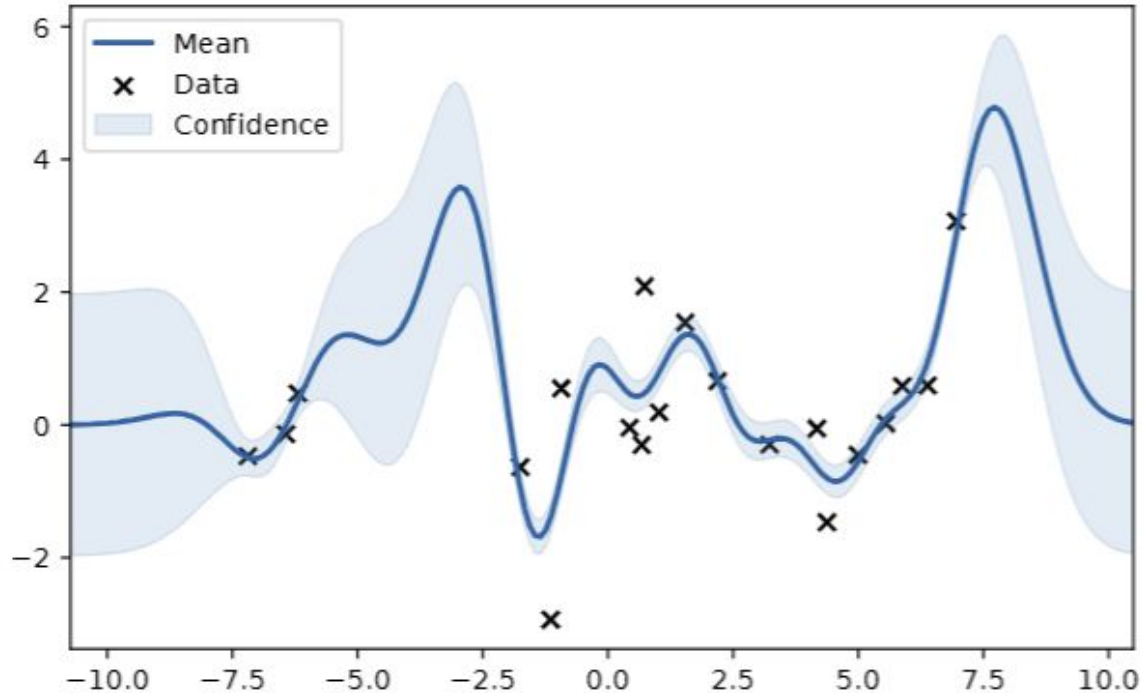
I. How can we incorporate prior knowledge into a model?

II. How can we account for uncertainty in parameters?

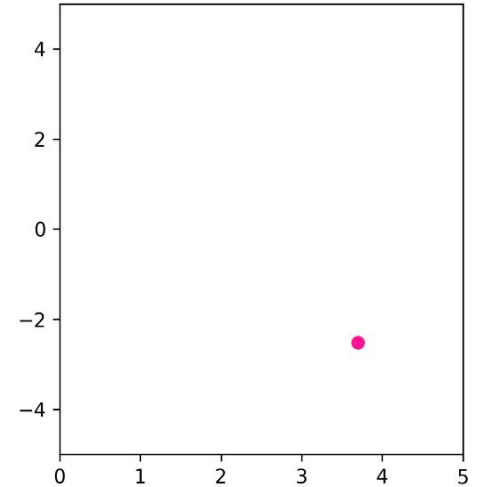
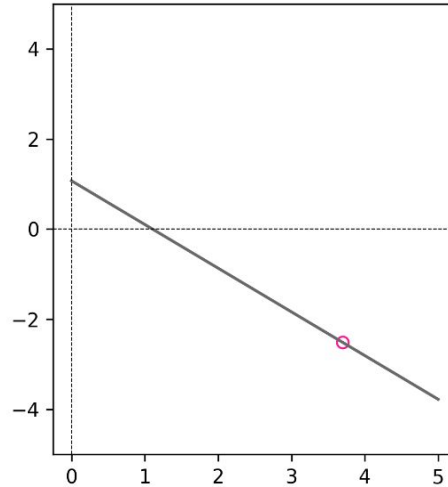
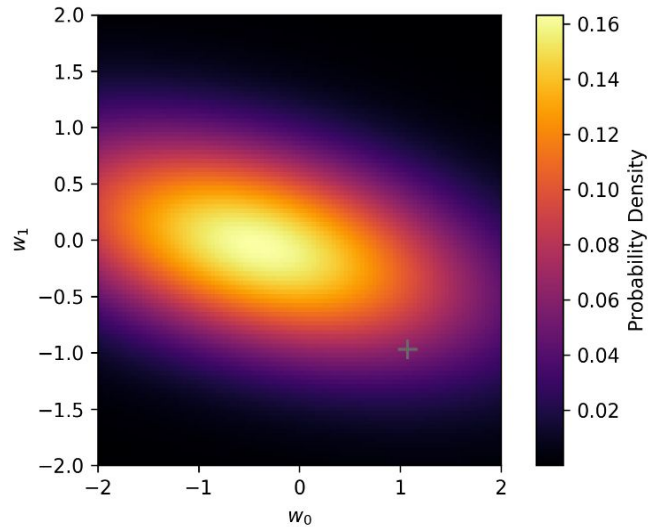
9b

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

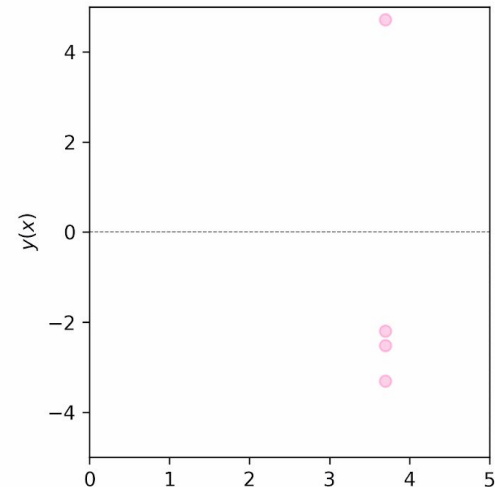
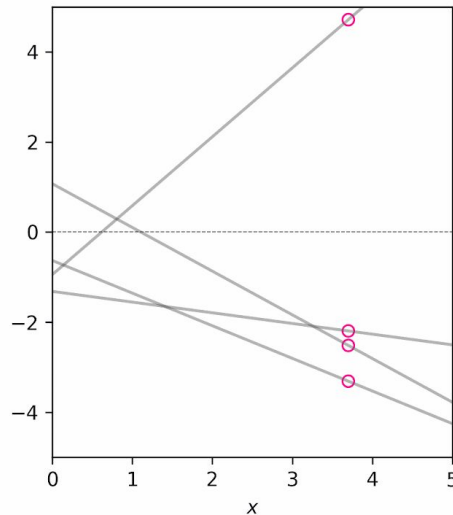
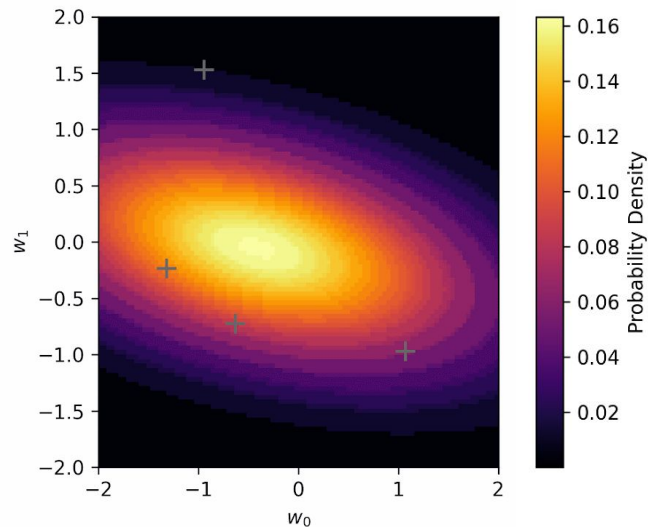
Behavior of the Gaussian Process model



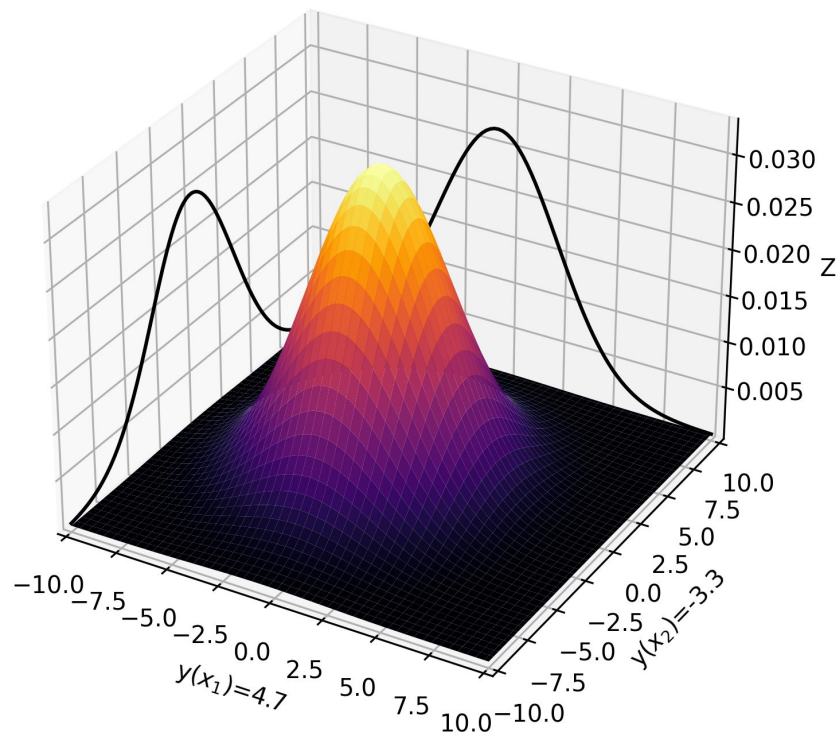
A sample from the posterior distribution defines one function



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

$$y(x) = \mathbf{w}^\top \phi(x)$$

$$\mathbf{y} = \Phi \mathbf{w}$$

$$\mathbb{E}[\mathbf{y}] = \Phi \mathbb{E}[\mathbf{w}] = \mathbf{0}$$

$$\text{Cov}[\mathbf{y}] = \mathbb{E}[\mathbf{y}\mathbf{y}^\top]$$

$$= \Phi \mathbb{E}[\mathbf{w}\mathbf{w}^\top] \Phi^\top$$

$$= \alpha^{-1} \Phi \Phi^\top$$

$$= \alpha^{-1} \mathbf{K}$$

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

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

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

A Gaussian Process model for regression

$$t_i = y_i + \epsilon_i$$

$$p(t_i \mid y_i) = \mathcal{N}(t_i \mid y_i, \beta^{-1})$$

$$p(\mathbf{t} \mid \mathbf{y}) = \mathcal{N}(\mathbf{t} \mid \mathbf{y}, \beta^{-1} \mathbf{I}_N)$$

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

$$\begin{aligned} p(\mathbf{t}) &= \int p(\mathbf{t} \mid \mathbf{y}) p(\mathbf{y}) d\mathbf{y} \\ &= \mathcal{N}(\mathbf{t} \mid \mathbf{0}, \mathbf{C}) \end{aligned}$$

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

Augmenting the covariance matrix for prediction

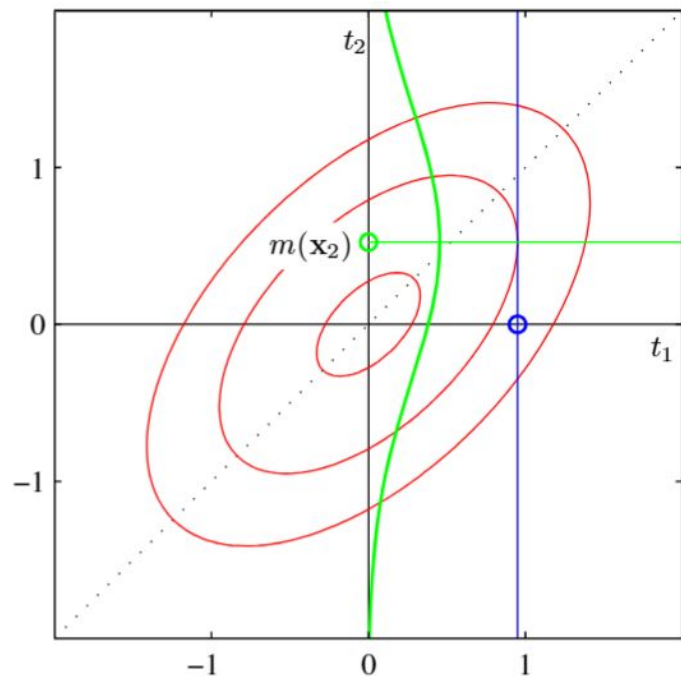
$$p(\mathbf{t}_{N+1}) \sim \mathcal{N}(\mathbf{t}_{N+1} \mid 0, C_{N+1})$$

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

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

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

The predictive distribution for Gaussian process regression



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

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

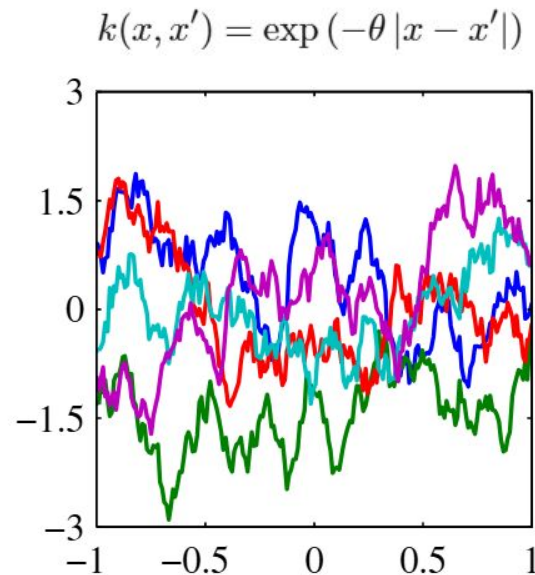
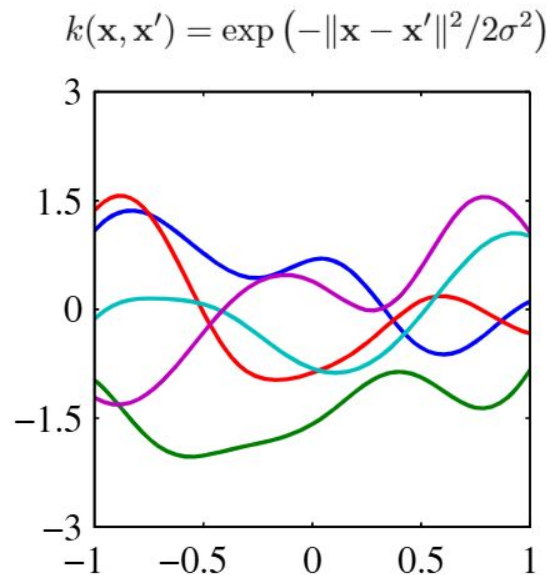
$$\mu_{x_{N+1}} = \mathbf{k}^\top \mathbf{C}_N^{-1} \mathbf{t}$$

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

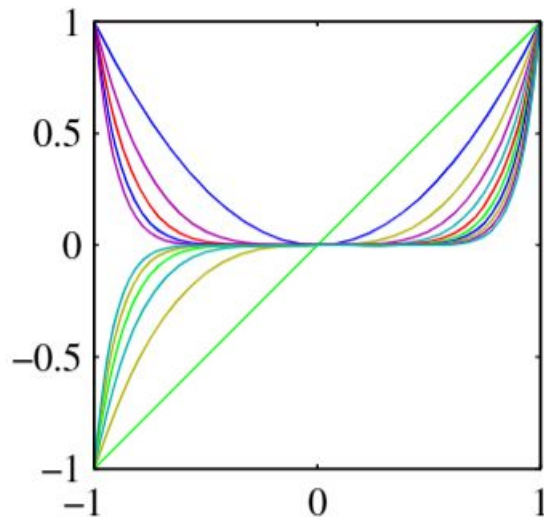
where a_i are the elements of the N -vector $\mathbf{a} = \mathbf{C}_N^{-1} \mathbf{t}$.



The kernel function itself need not be Gaussian

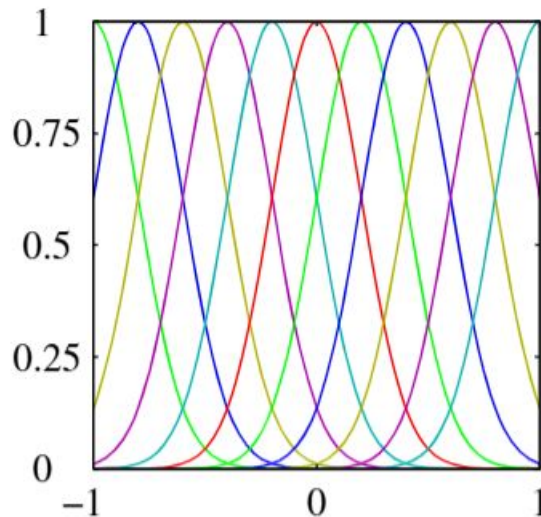


Fixed nonlinear basis functions in regression are special cases of Gaussian Processes



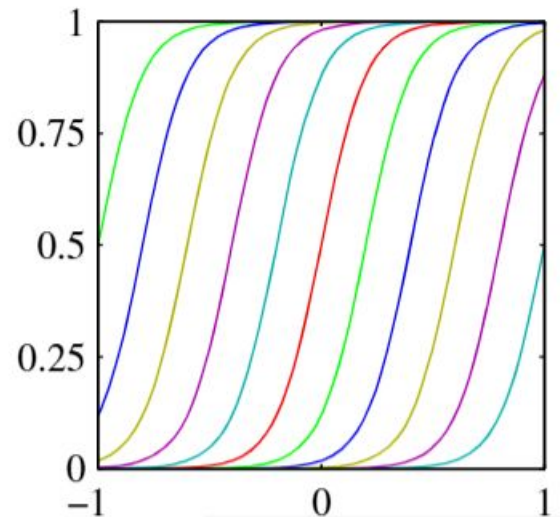
Polynomial:

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



Gaussian:

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



Sigmoidal:

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

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 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.