# CS559 Lecture 11:
# Gaussion Process

Reading: Chapter 6, Bishop book

# Constructing kernels

- Kernels allow us to work with the dual representation
- Many linear models have dual representation involving kernels (perception, Gaussian Process)
- We can construct kernels directly from basis functions $\phi(x)$
- Or we can construct kernels directly, as long as it is a valid kernel corresponding to a scalar product in some feature space.
- function k(x,x') is a valid kernel ←→ the Gram matrix K with elements k(x_n,x_m) is pos. def. for all sets of {x_i}.

# Kernels

$$k(x, x') = \phi^T(x)\phi(x) = \sum_i \phi_i(x)\phi_i(x')$$

An example,

$$k(x, z) = (x^T z)^2, \quad \phi(x) = ?$$

$$\begin{aligned} k(x, z) &= (x^t z)^2 = (x_1 z_1 + x_2 z_2)^2 \\ &= x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\ &= (x_1^2, \sqrt{2} x_1 x_2, x_2^2)(z_1^2, \sqrt{2} z_1 z_2, z_2^2)^T \end{aligned}$$

Thus, in this case,

$$\phi(x) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$$

and $k(x, z)$ computes the dot product of $\Re^3$ in $\Re^2$!

## Popular Kernels

- Linear kernel

$$K(x, x') = x^T x'$$

- Polynomial kernel

$$K(x, x') = (1 + x^T x')^p$$

- Radial basis kernel

$$K(x, x') = \exp\left(-\frac{1}{2}||x - x'||^2\right)$$

Here the feature space corresponds to infinite dimensional space (infinite terms in the expansion of $\exp(x)$).

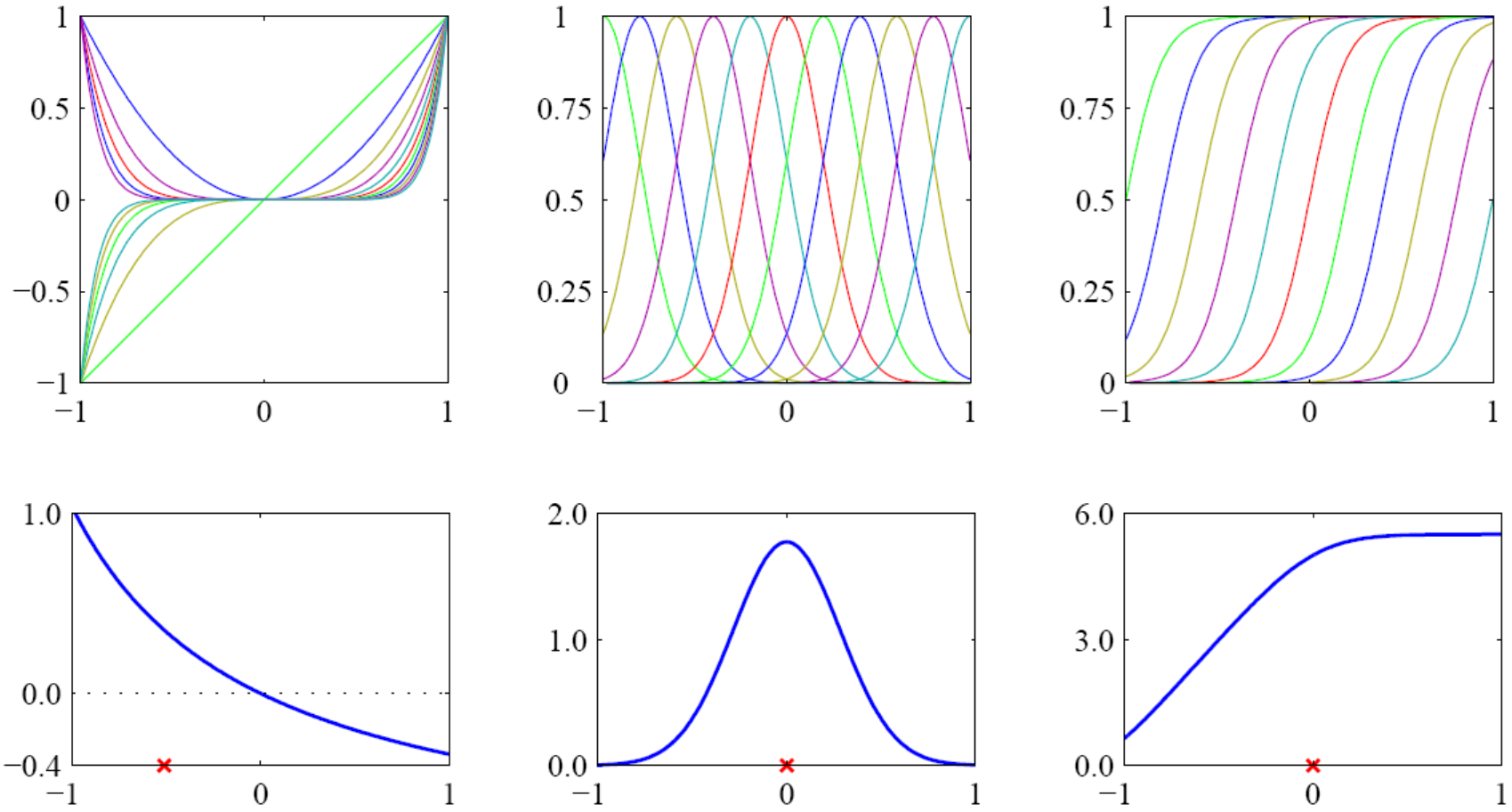



**Figure 6.1** Illustration of the construction of kernel functions starting from a corresponding set of basis functions. In each column the lower plot shows the kernel function $k(x, x')$ defined by (6.10) plotted as a function of $x$, where $x'$ is given by the red cross ($\times$), while the upper plot shows the corresponding basis functions given by polynomials (left column), 'Gaussians' (centre column), and logistic sigmoids (right column).

# Kernel of Kernels

$$
\begin{aligned}
k(x, x') &= c k_1(x, x') \\
k(x, x') &= f(x) k_1(x, x') f(x') \\
k(x, x') &= q(k_1(x, x')) \\
k(x, x') &= \exp(k_1(x, x')) \\
k(x, x') &= k_1(x, x') + k_2(x, x') \\
k(x, x') &= k_1(x, x') k_2(x, x') \\
k(x, x') &= k_1(x, x') k_2(x, x') \\
&\ldots
\end{aligned}
$$

where $c > 0$, $f()$ is any function, and $q()$ is any polynomial with nonnegative coefficients

# More kernels of kernels

$$k(x, x') = k_3(\phi(x), \phi(x'))$$

$$k(x, x') = x^T A x'$$

$$k(x, x') = k_a(x_a, x_a') + k_b(x_b, x_b')$$

$$k(x, x') = k_a(x_a, x_a') k_b(x_b, x_b')$$

Kernels are not limited to Euclidean distance. We can replace $x^T x$

With a nonlinear kernel $\kappa(x^T x)$

# Kernel from Generative Models

$$
\begin{aligned}
k(x, x') &= p(x)p(x') \\
k(x, x') &= \sum_i p(x|i)p(x'|i)p(i)
\end{aligned}
$$

where $i$ could be mixture components or hidden states sequence
Another interesting kernel is Fisher kernel

$$
\begin{aligned}
k(x, x') &= g(\theta, x)F^{-1}g(\theta, x) \\
g(\theta, x) &= \nabla_\theta \ln p(x|\theta) \\
F^{-1} &= E_x[g(\theta, x), g(\theta, x)^T] \\
&\approx \frac{1}{N}\sum_{i=1}^{N} g(\theta, x_i), g(\theta, x_i)^T]
\end{aligned}
$$

For more details, see Jaakkola and Haussler, 1999.

# Kernels

- **Kernels define** a **similarity measure** :
  - define a distance in between two objects
- **Design criteria:** we want kernels to be
  - **valid** – Satisfy **Mercer condition** of positive semi-definiteness
  - **good** – embody the "true similarity" between objects
  - **appropriate** – generalize well
  - **efficient** – the computation of K(x,x') is feasible
    - NP-hard problems abound with graphs

# Kernels

- Research have proposed kernels for comparison of variety of objects:
  - Strings
  - Trees
  - Graphs

# Gaussian Process

- Extension of kernels to probabilistic discriminant models
- Define a prior probability distribution over the functions directly.
- Only need to consider function values at training and test data sets. – work in finite space.
- Kriging, ARMA, Kalman filters, RBFs are all Gaussian Processes.
- http://www.gaussianprocess.org/

## Gaussian Stochastic Process

Consider linear model

$$y(x) = w^T \phi(x)$$

with prior over $w$

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

Now, if $\mathbf{y}$ is vector of samples from a stochastic process, then it is a Gaussian stochastic process with

$$\begin{aligned} E[\mathbf{y}] &= \Phi E[w] = 0 \\ E[\mathbf{y}] &= \Phi E[ww^T]\Phi^T = \frac{1}{\alpha}\Phi\Phi^T = K \end{aligned}$$

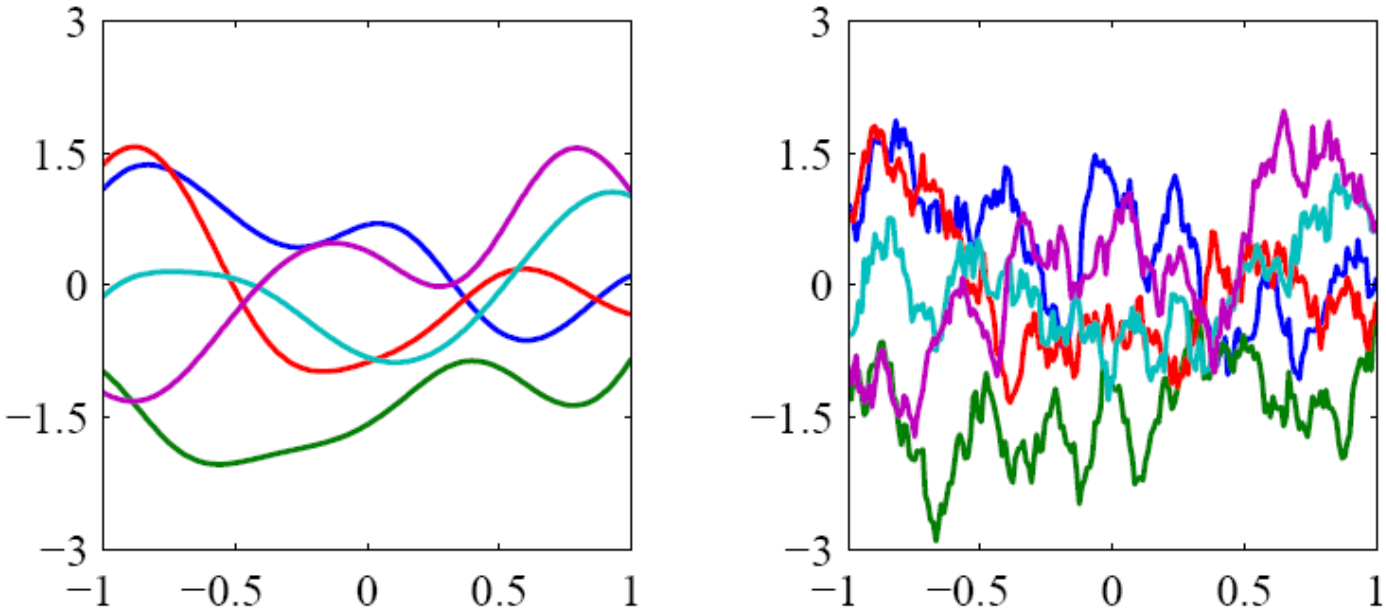
where $K$ is the Gram matrix

$$K_{nm} = k(x_n, x_m) = \frac{1}{\alpha}\phi(x_n)^T \phi(x_m)$$

# Gaussion process

- Gaussion process is a probability distribution over function $y(x)$ such that the values of y's at points x1,… x_N have a joint Gaussion distribution.

- 2D -> Gaussian random field.



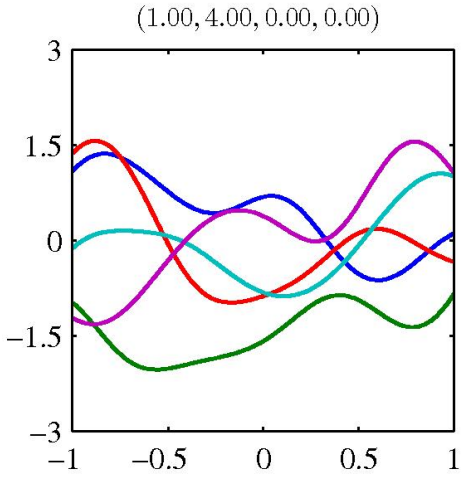

$$k_1(x_n, x_m) = \exp(-||x_n - x_m||^2/2\sigma^2)$$
$$k_2(x_n, x_m) = \exp(-\theta|x_n - x_m|)$$

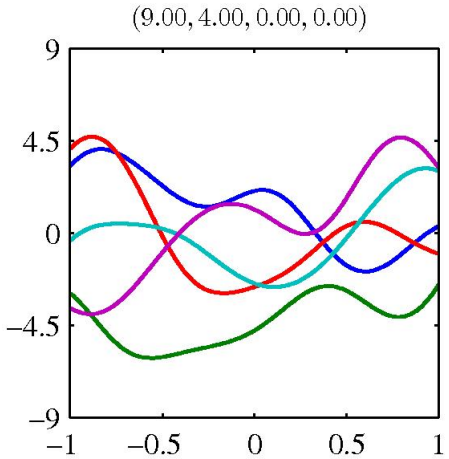
Gaussian stochastic processes are completely defined by the second order statistics!

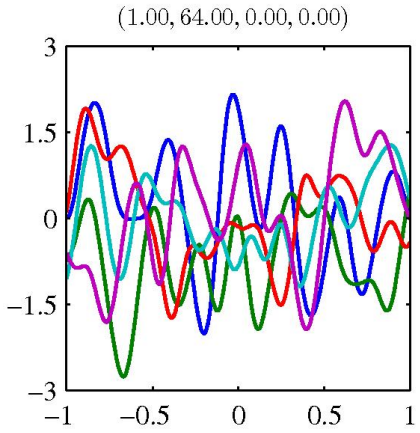
# Gaussian Process

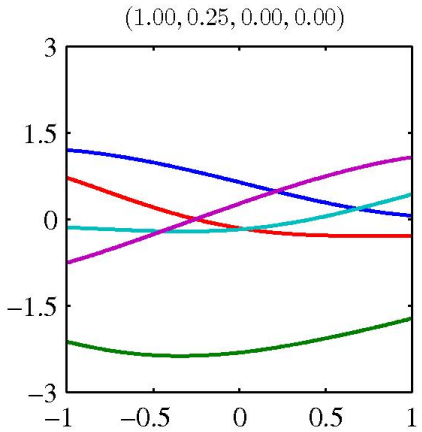
One popular choice of kernel in this case is

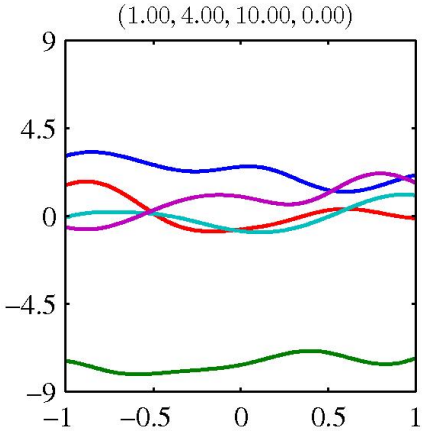
$$k(x_n, x_m) = \theta_0 \exp\left(-\frac{\theta_1}{2}||x_n - x_m||^2\right) + \theta_2 + \theta_3 x_n^T x_m$$

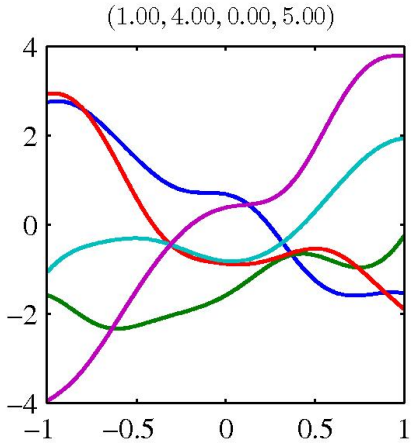


# Gaussian Process for Regression

$$t_n = y_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \beta^{-1})$$

Since $y_n$ and $\epsilon_n$ are independent,

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

and we know

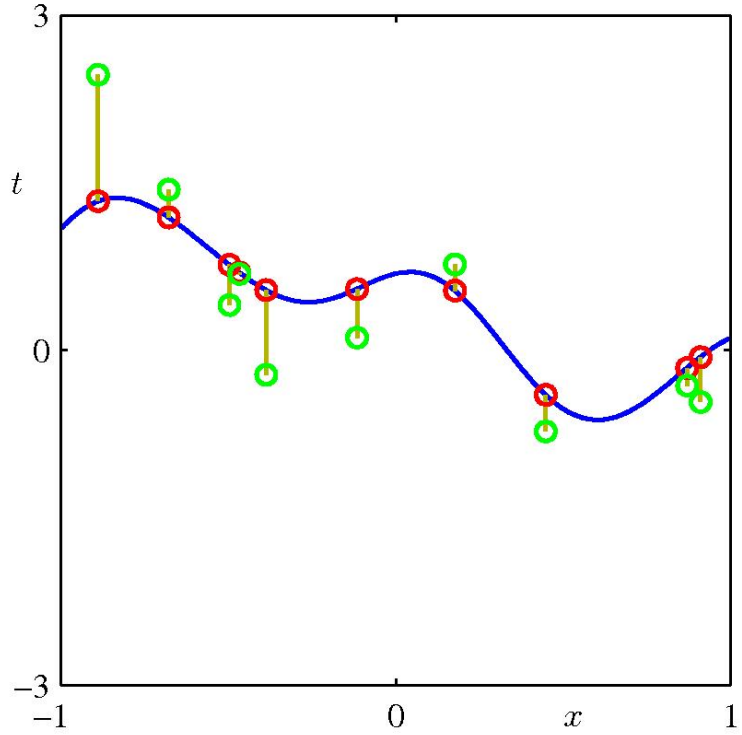
$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|0, K)$$

Therefore

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y} = \mathcal{N}(\mathbf{t}|0, \mathbf{C})$$

where

$$C(x_n, x_m) = k(x_n, x_m) + \beta^{-1}\delta_{nm}$$


3
0
−3
t
−1
0
1
x

# Gaussian Process for Regression

For regression, however, we need $p(t_{n+1}|\mathbf{t})$. We start with

$$p(\mathbf{t}_{n+1}) = \mathcal{N}(\mathbf{t}_{n+1}|0, \mathbf{C}_{n+1})$$

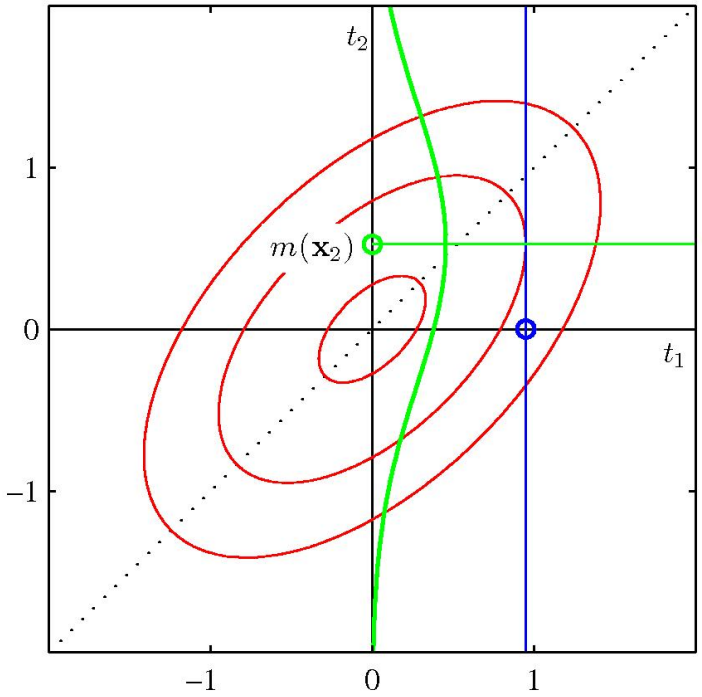
where $\mathbf{C}_{n+1}$ is a $(N+1) \times (N+1)$ covariance matrix.

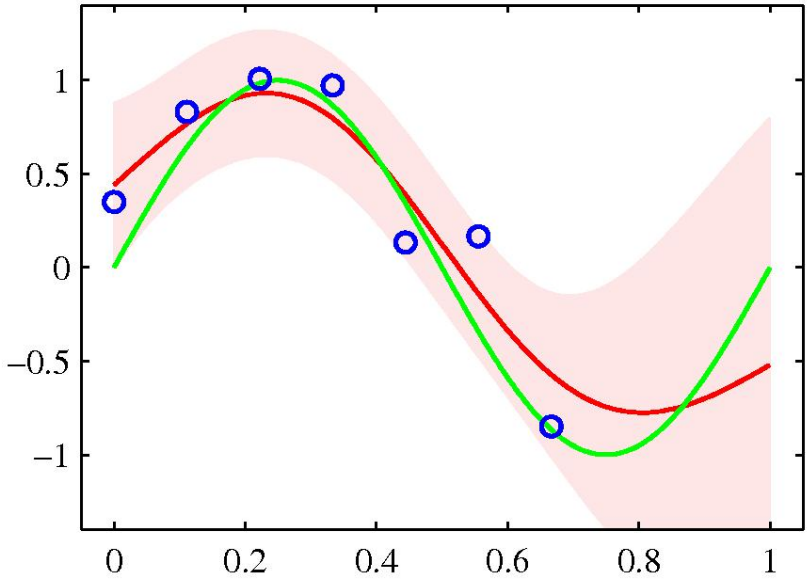
Since the underlying process is a Gaussian stochastic process, you can marginalize the distribution by partitioning the covariance (2.81,2.82), thus

$$\begin{aligned} E[t_{n+1}] &= m(x_{n+1}) = \mathbf{k}^T \mathbf{C}_n^{-1} \mathbf{t} \\ E[t_{n+1} t_{n+1}] &= \sigma^2(x_{n+1}) = c - \mathbf{k}^T \mathbf{C}_n^{-1} \mathbf{k} \end{aligned}$$

$m(x_{n+1})$ can also be written as

$$m(x_{n+1}) = \sum_{n=1}^{N} a_n k(x_n, x_{n+1})$$




If the kernel function is chose as a specific finite set of basis functions, it will lead to the Linear Regression case we talked about before.

Parametric space + linear regresss ←→ functional space + Gaussion Process