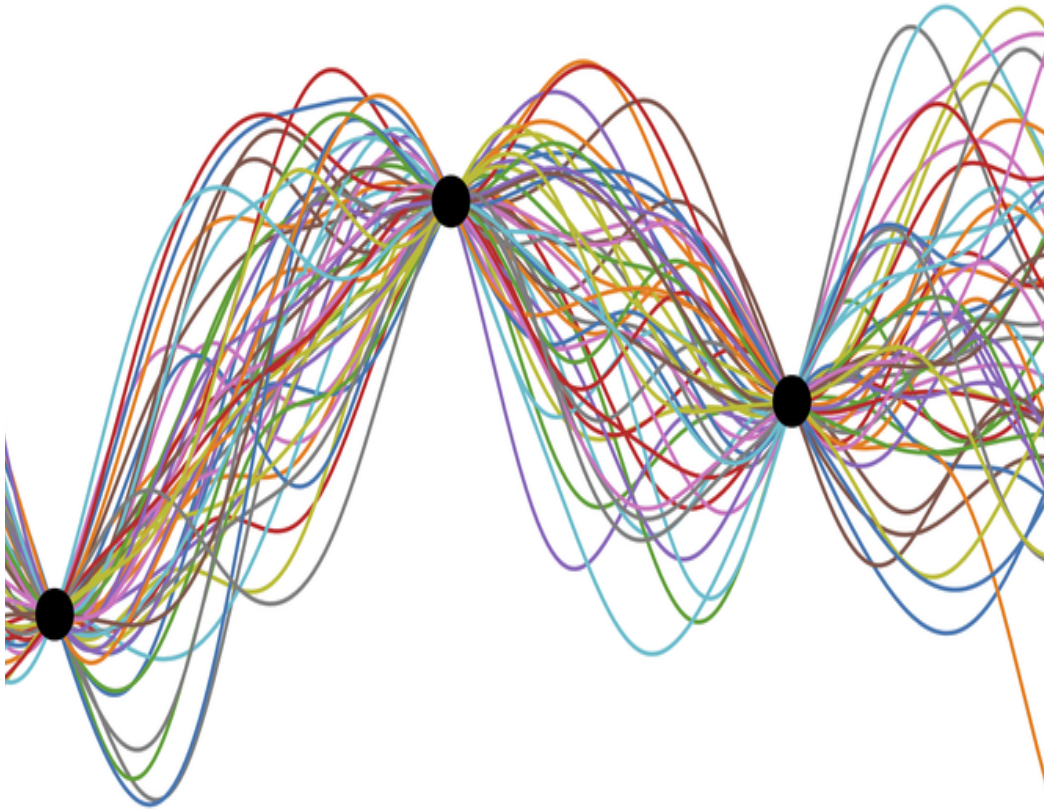
2.160 Identification, Estimation, and Learning

Part 4 Machine Learning and Nonlinear System Modeling

Lecture 23

Gaussian Processes

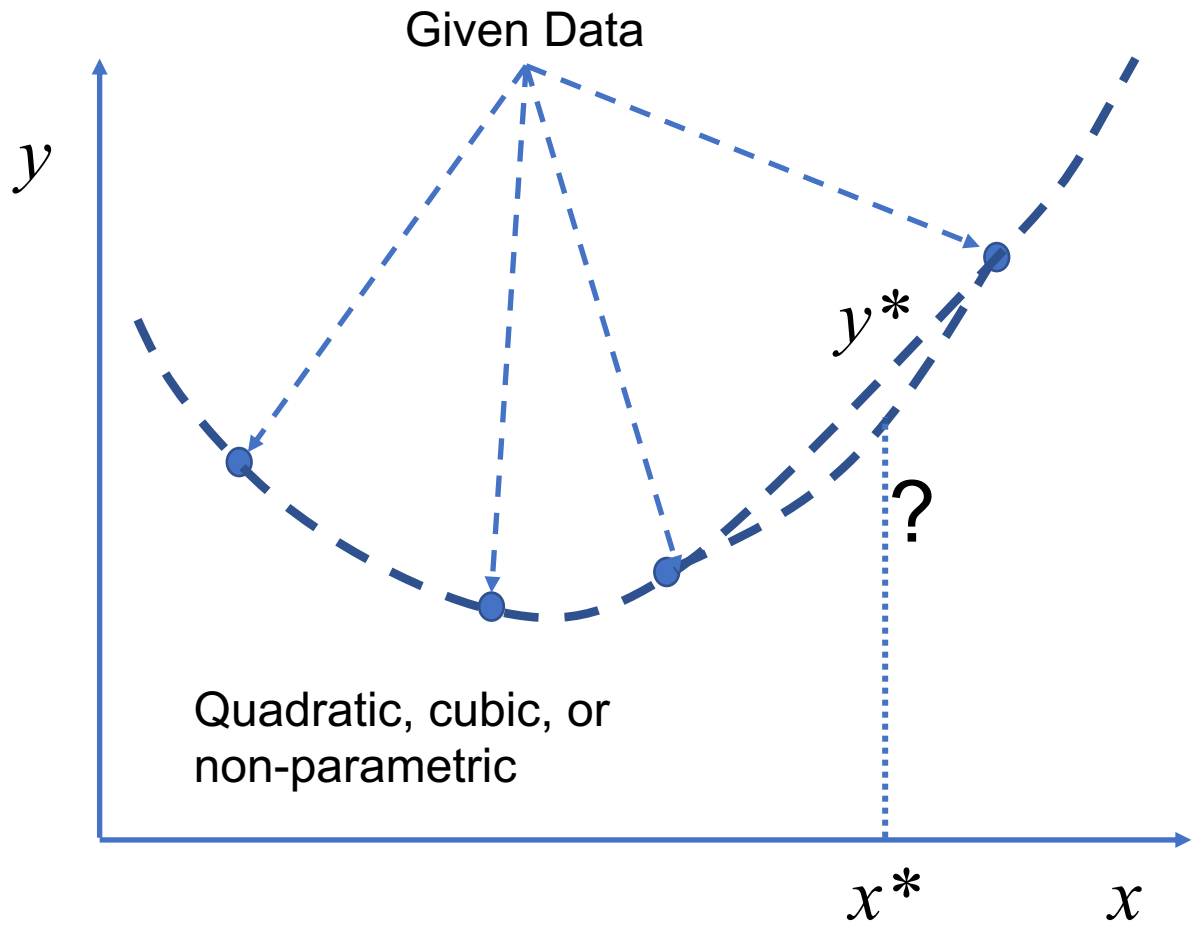


H. Harry Asada
Department of Mechanical Engineering
MIT

Gaussian Processes in a Nutshell

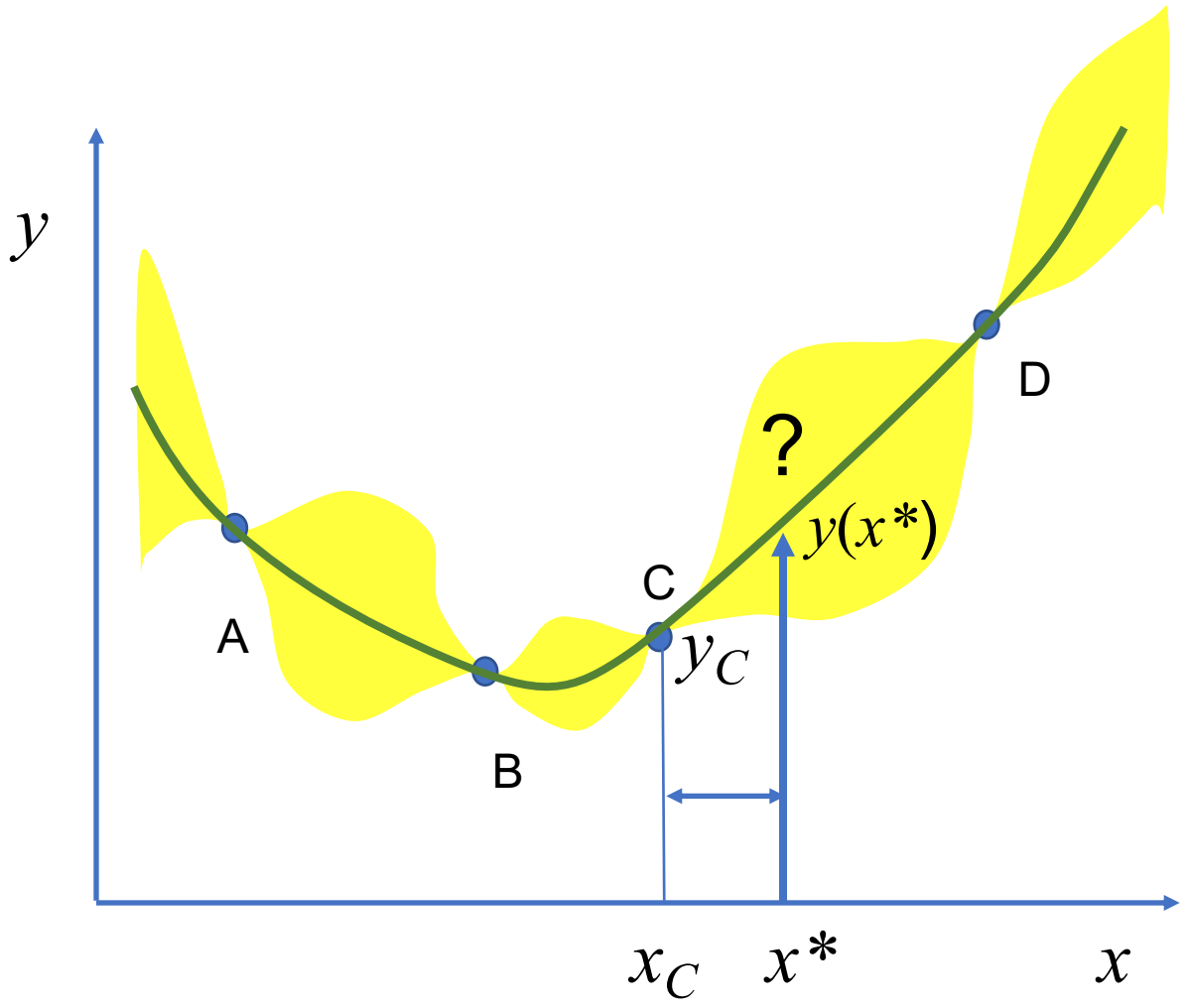
A Gaussian Process is a non-parametric, nonlinear regression, providing a stochastic model.

- A few data points are given.
- Want to find an output value y^* in response to a test point x^*



Gaussian Processes in a Nutshell

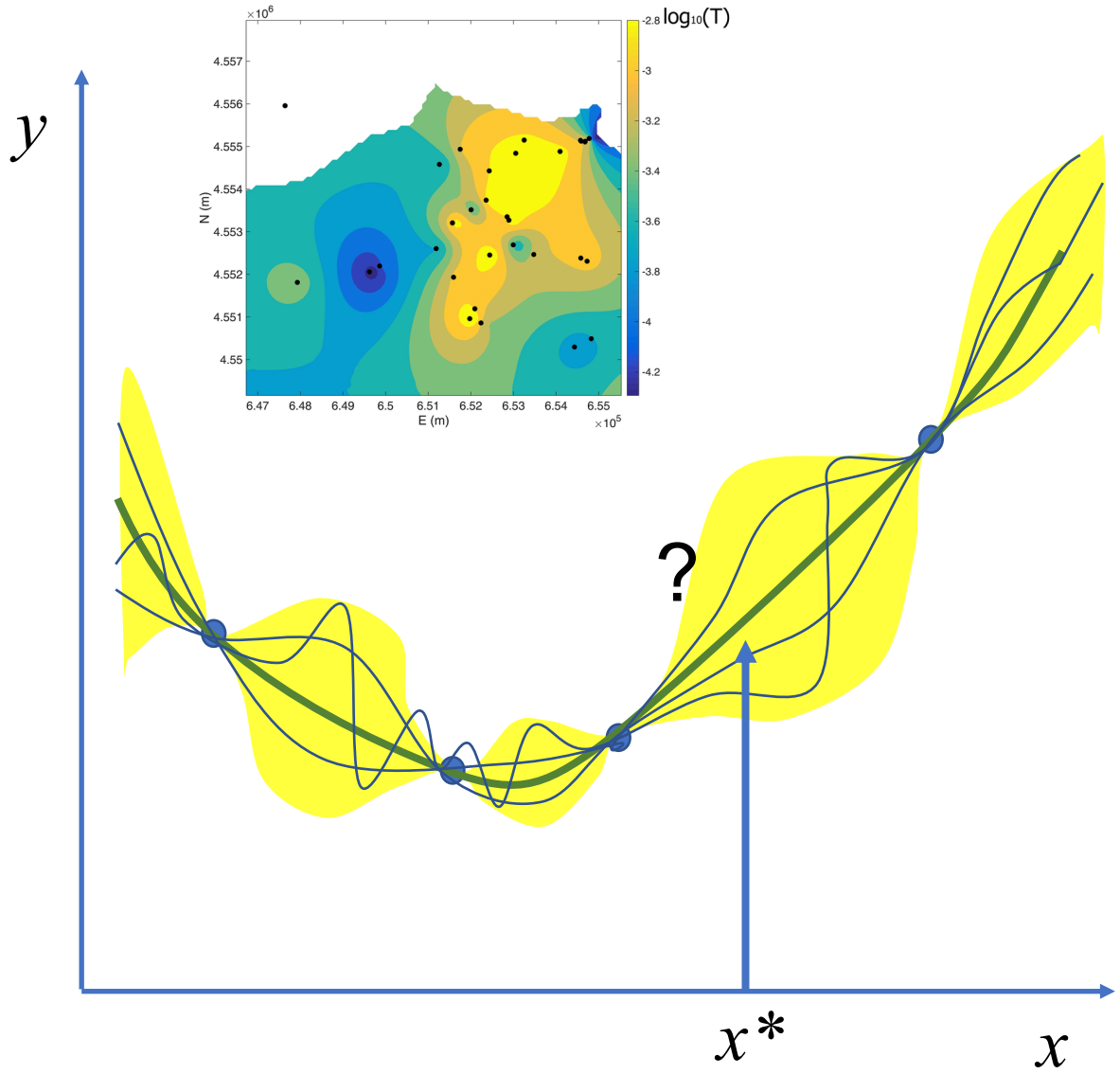
Nonlinear Regression



- If x^* is close to x_C, output $y(x^*)$ should be close to y_C.
- As x^* gets further from x_C, the output may differ from y_C, and become more uncertain (unpredictable).
- Distal point D as well as points A and B, too, influence the prediction of output in response to x^*.
- How can we find a likely $y(x^*)$? More importantly, its mean and variance: the distribution of $y(x^*)$?
- Gaussian Process regression can answer these questions.

Gaussian Processes in a Nutshell

Geostatistics



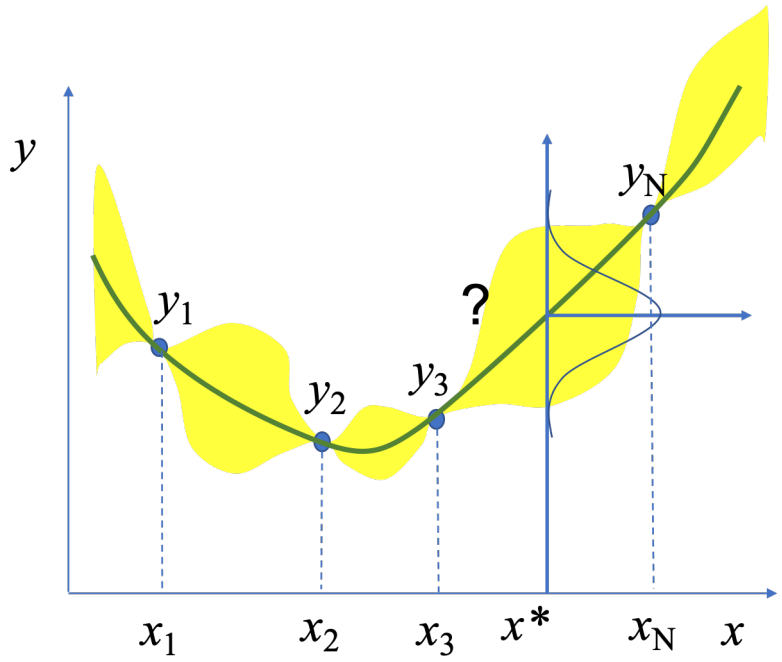
- Geostatistics is a branch of statistics dealing with spatial or spatiotemporal data.
- Historically, Gaussian Process came from geostatistics for mining.
- Now it is applied to petroleum geology, hydrogeology, oceanography, geochemistry, environmental control, and more.
- The basic theory and techniques are applied to machine learning and system dynamics & control.

Gaussian Processes

Machine Learning

- Given training sample data

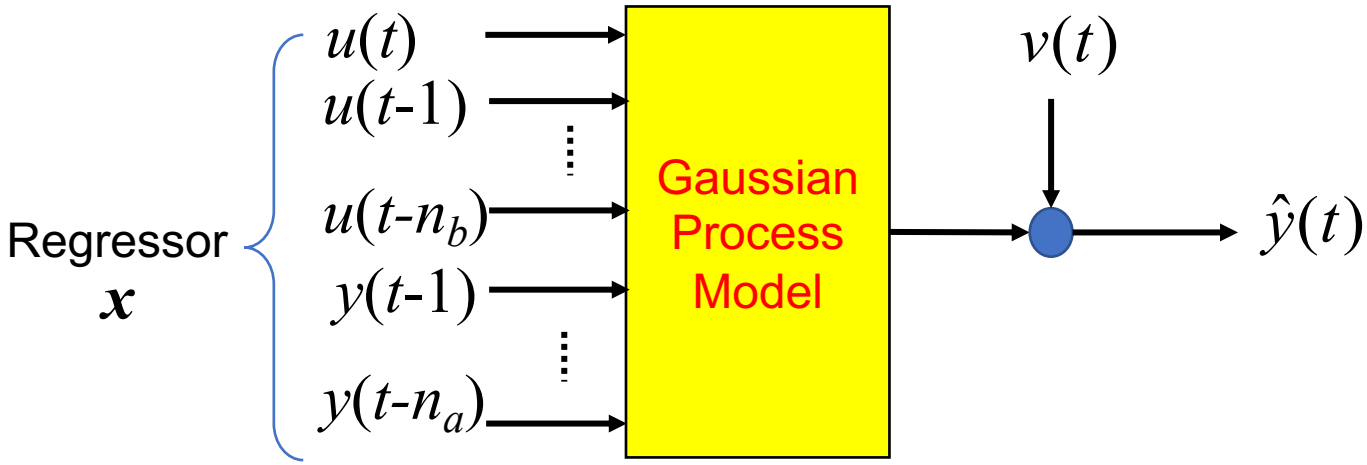
- Predict output mean and covariance for new points



System Identification and Control

- Combined with various model structures of dynamical systems, Gaussian Processes provide us with nonlinear system identification framework.

Nonlinear ARX Model (GP-NARX)

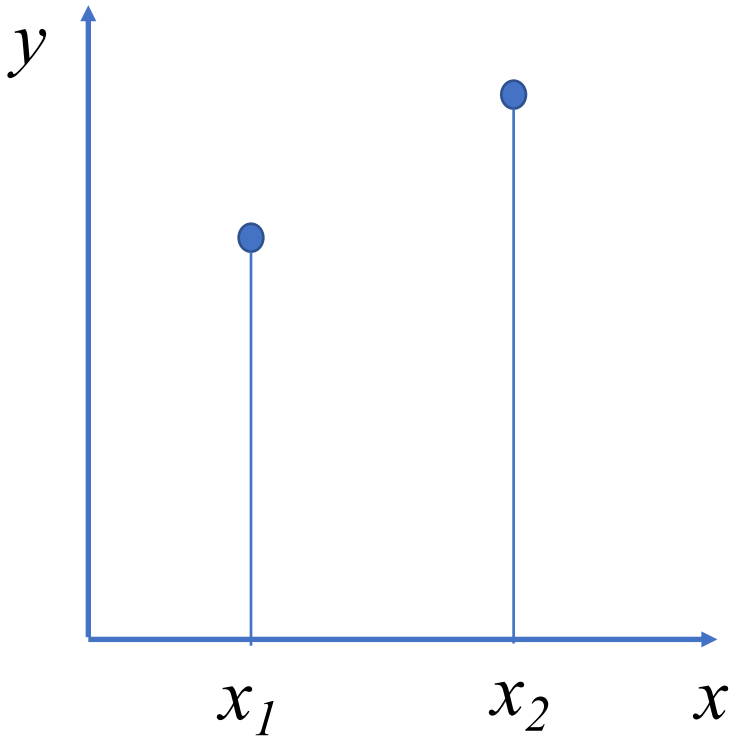


Gaussian Processes in a Nutshell

- Suppose that random variables x_1 and x_2 are jointly Gaussian.
- Note x_1 and x_2 are correlated.

$$p(x_1, x_2) = \frac{1}{(2\pi \det \Sigma)} \exp\left(-\frac{1}{2}(x_1, x_2)\Sigma^{-1}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right)$$

$$\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

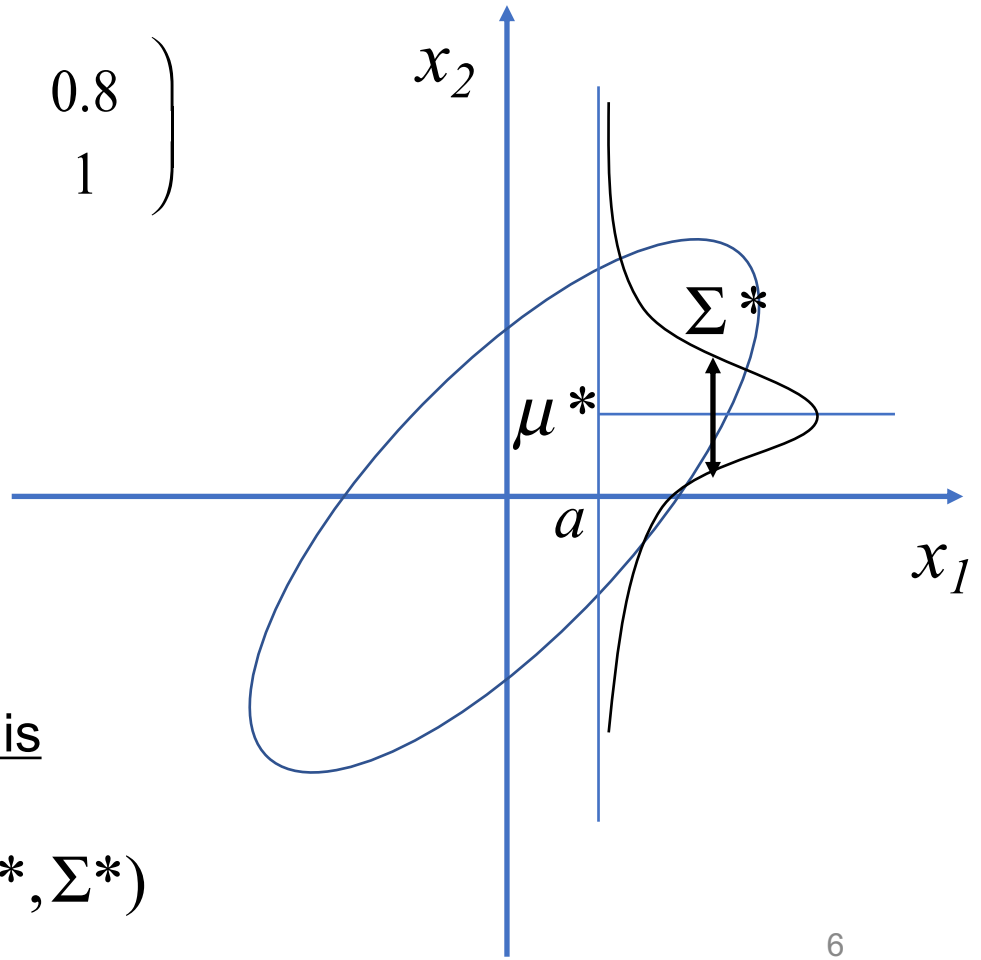


- Suppose that $x_1 = a$ is observed.
- Find the conditional probability distribution:

$$p(x_2 \mid x_1)$$

- It turns out that the conditional distribution is also Gaussian.

$$x_2 \mid x_1 \sim \mathbb{N}(\mu^*, \Sigma^*)$$

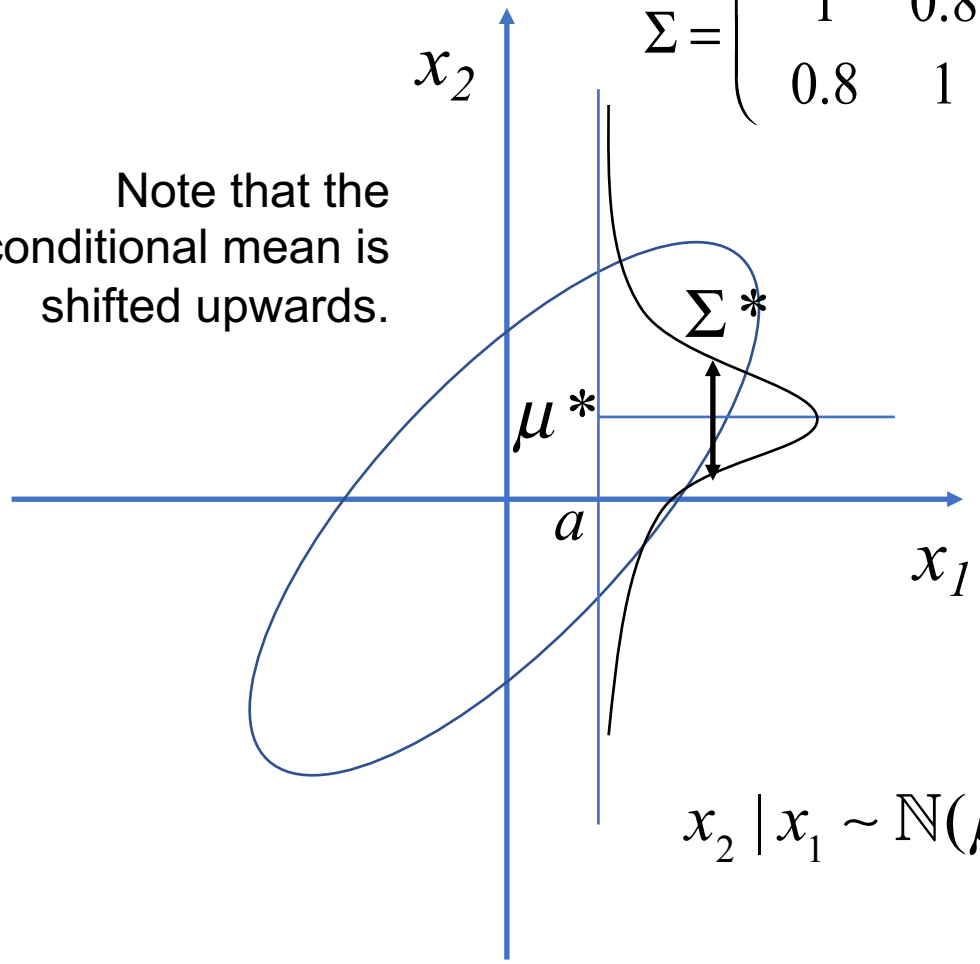


Gaussian Processes in a Nutshell

$$p(x_1,x_2)=\frac{1}{(2\pi\det\Sigma)}\exp\left(-\frac{1}{2}(x_1,x_2)\Sigma^{-1}\begin{pmatrix}x_1\\x_2\end{pmatrix}\right)$$

$$\Sigma=\begin{pmatrix}1 & 0.8\\ 0.8 & 1\end{pmatrix}$$

Note that the conditional mean is shifted upwards.



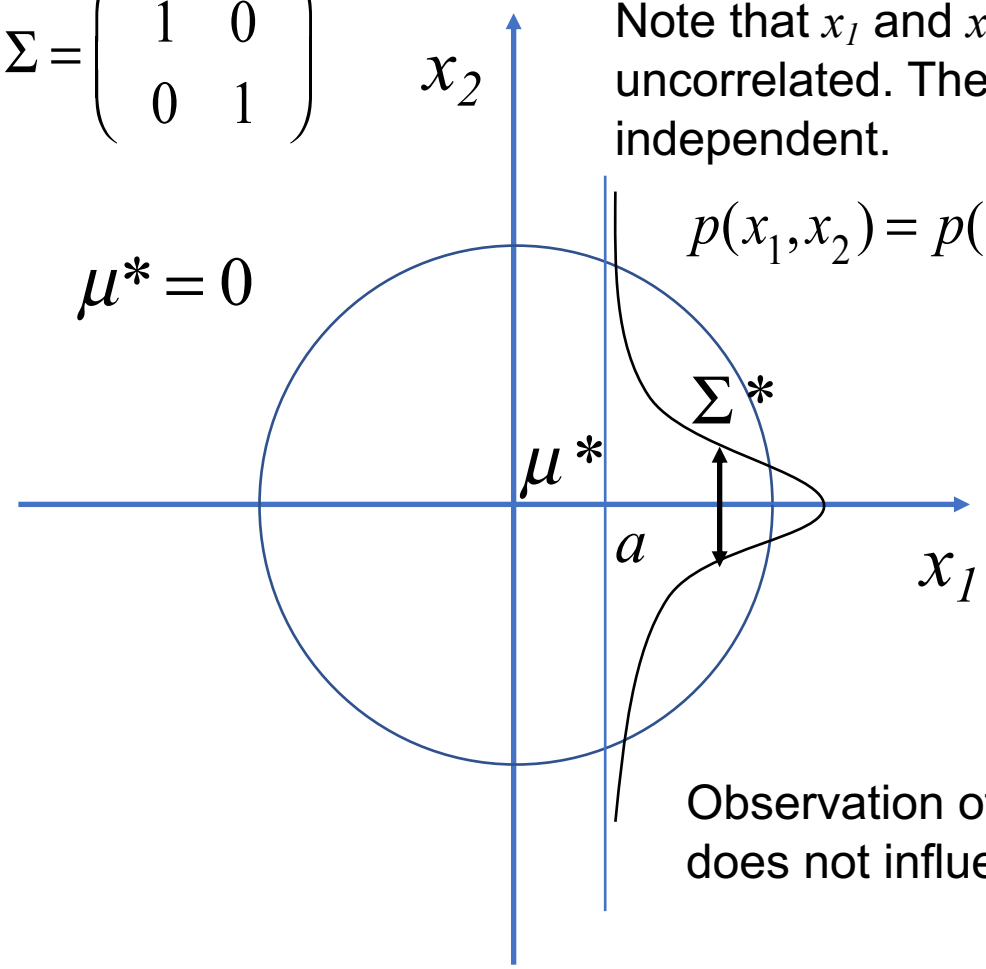
$$x_2 \mid x_1 \sim \mathbb{N}(\mu^*,\Sigma^*)$$

$$\Sigma=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

Note that x_1 and x_2 are uncorrelated. They are independent.

$$p(x_1,x_2)=p(x_1)p(x_2)$$

$$\mu^*=0$$

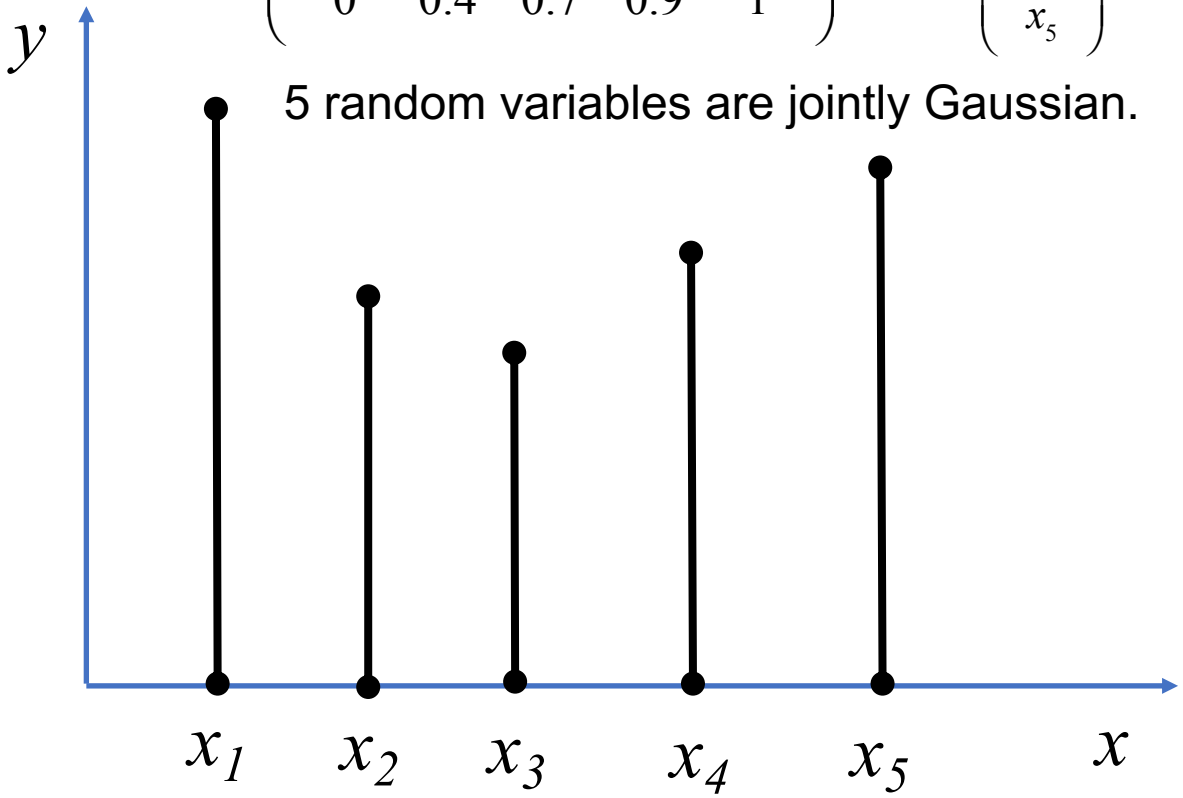


Observation of x_1 does not influence x_2.

Gaussian Processes in a Nutshell

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.7 & 0.4 & 0 \\ 0.9 & 1 & 0.9 & 0.7 & 0.4 \\ 0.7 & 0.9 & 1 & 0.9 & 0.7 \\ 0.4 & 0.7 & 0.9 & 1 & 0.9 \\ 0 & 0.4 & 0.7 & 0.9 & 1 \end{pmatrix} \qquad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \sim \mathbb{N}(\mu, \Sigma)$$

5 random variables are jointly Gaussian.



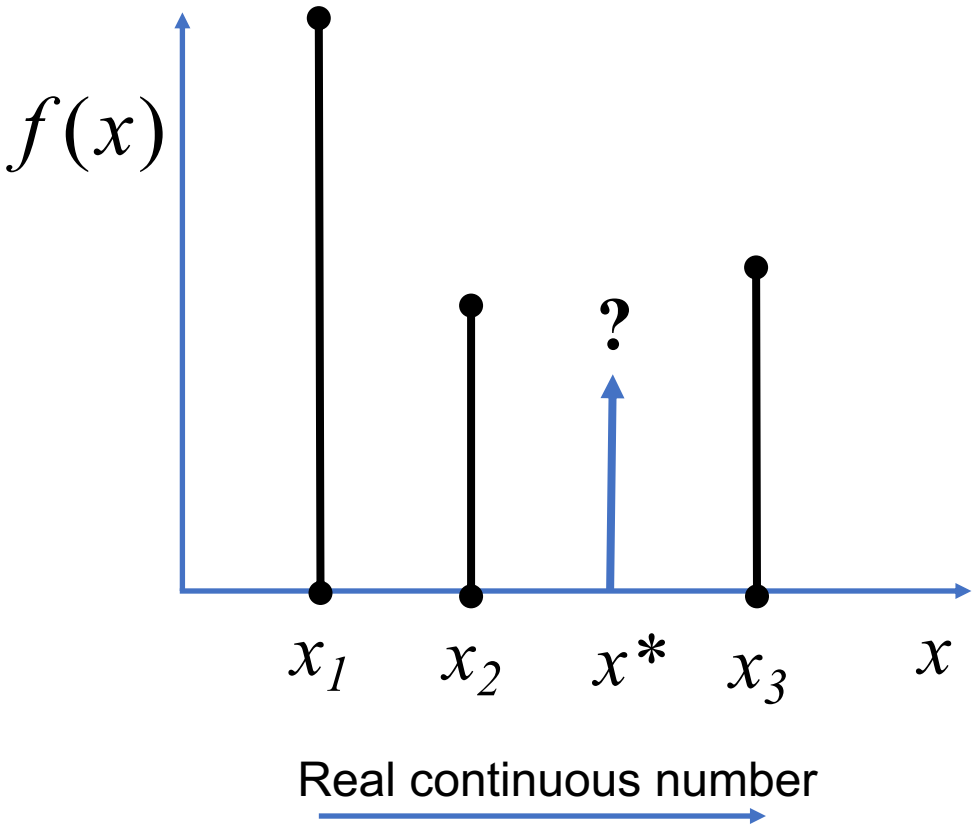
- What if we consider an uncountable infinite number of random variables?
- Instead of using index 1,2,…., we use a variable x, which takes a continuous real number.

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \end{pmatrix} \Rightarrow f(x)$$

- The vector becomes a function in a Hilbert space.

Definition of Gaussian Process

A Gaussian process is a type of random process: a collection of random variables, any finite number of which have a joint Gaussian distribution.



- Consider a finite number of random variables taken from a Gaussian process, and place them in a vector $\boldsymbol{f}$.

$$\boldsymbol{f} \sim \mathbb{N}(\mu, \Sigma)$$

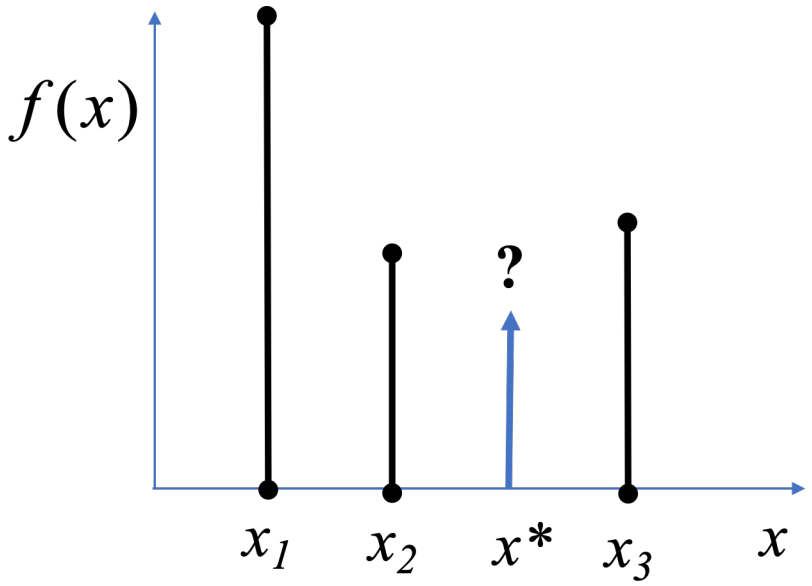
- Also consider a set of test points placed in another vector $\boldsymbol{f}^*$, which we want to predict. Collectively, these random variables are jointly Gaussian.

$$\begin{pmatrix} \boldsymbol{f} \\ \boldsymbol{f}^* \end{pmatrix} \sim \mathbb{N}\left(\begin{pmatrix} \mu \\ \mu^* \end{pmatrix}, \begin{pmatrix} \Sigma & C \\ C^T & \Sigma^* \end{pmatrix} \right)$$

Gaussian Processes in a Nutshell

- Jointly Gaussian distribution: Before conditioning f^* on f, the joint distribution has the following mean and covariance, called "Prior".

$$\begin{pmatrix} f \\ f^* \end{pmatrix} \sim \mathbb{N}\left(\begin{pmatrix} \mu \\ \mu^* \end{pmatrix}, \begin{pmatrix} \Sigma & C \\ C^T & \Sigma^* \end{pmatrix} \right)$$



- Suppose that f are observed. Based on the observation, find f^*, namely, obtain the conditional distribution of f^*.

$$p(f^* \mid f) \quad : \text{ Posterior distribution}$$

- This can be obtained by using the following Gaussian Identity.

$$f^* \mid f \sim \mathbb{N}\left(\mu^* + C^T\Sigma^{-1}(f-\mu), \Sigma^* - C^T\Sigma^{-1}C\right)$$

- Usually, means, μ and μ^*, are set to zero.

$$\boxed{f^* \mid f \sim \mathbb{N}\left(C^T\Sigma^{-1}f, \Sigma^* - C^T\Sigma^{-1}C\right)}$$

- The posterior mean shifts to

$$m(f^* \mid f) = C^T\Sigma^{-1}f$$

- The posterior variance $\leq$ The prior variance

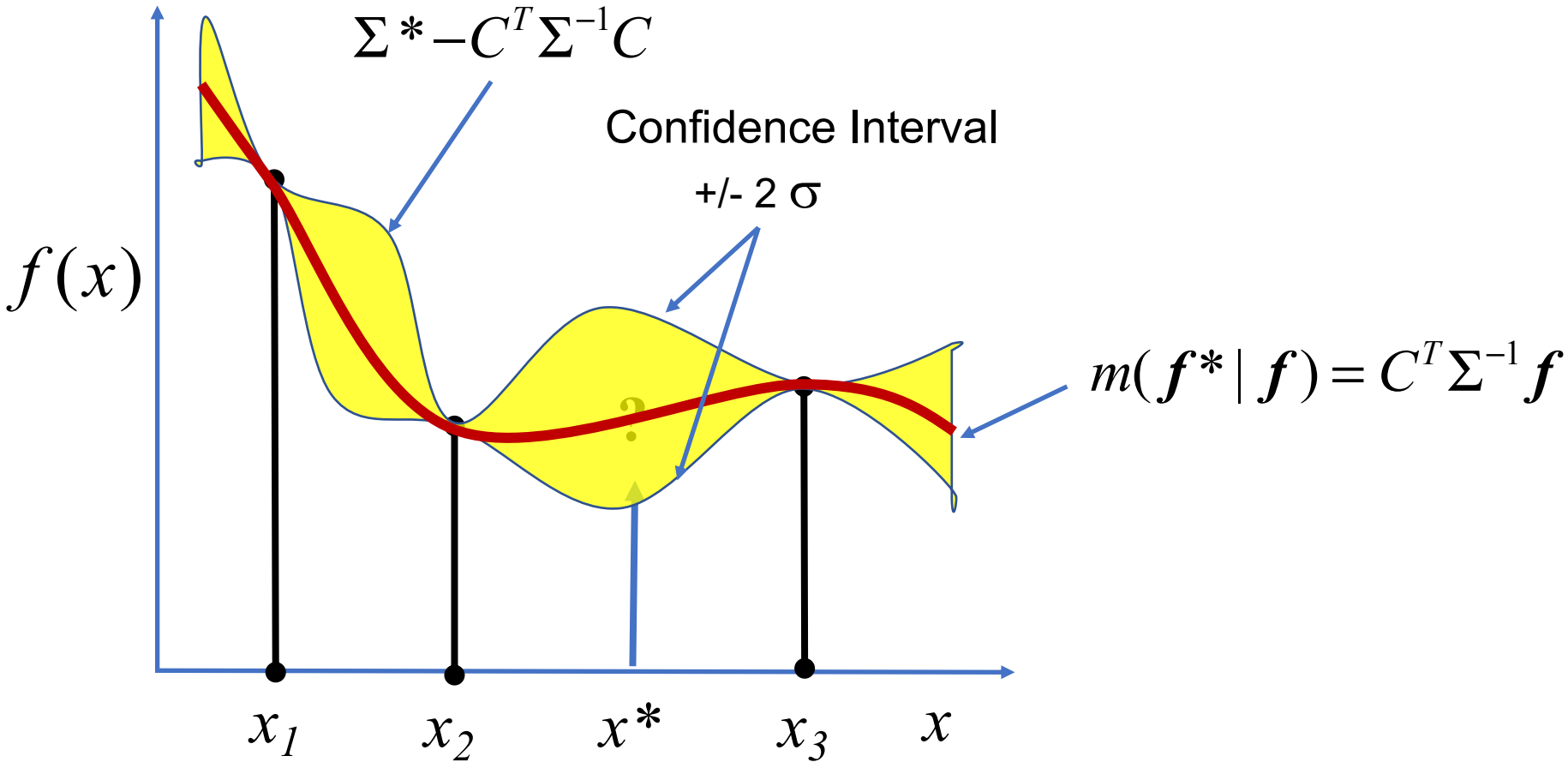
$$\Sigma^* - C^T\Sigma^{-1}C$$

Prior variance

Reduction in uncertainty due to observations

Gaussian Processes in a Nutshell

PS #6 Problem 3-c)



Outline: Roadmap

- Problem formulation: Noise must be considered in observation. We formulate the following noise corrupted observation model.

$$y = f(x) + \varepsilon$$

$$\varepsilon \sim \mathbb{N}(0, \sigma_n^2)$$

where ε is additive noise: independent, identically distributed (iid) Gaussian.

- Data matrices:

$$X = (\ x_1 \quad x_2 \quad \cdots \quad x_N\ ),\ Y = \begin{pmatrix} y_1 & y_2 & \cdots & y_N \end{pmatrix}^T$$

Combining these, $D = \{X, Y\}$

- Note that both ε and f are random. The latter f is a random process indexed with x.

1. To make it easier, consider a function with linear parametric representation:

$$f(x) = x^T w$$

where $x \in \Re^n$,

$f : \Re^n \to \Re$, a scalar function,

$w \in \Re^n$ weight vector

and w is a random variable.

2. We will expand the linear regression using ***kernels***.
3. Finally, we will obtain a general case, which is non-parametric.
   - Gaussian Process: $f(x)$ in a Hilbert Space.

23.1 Linear Parametric Model

- Giving a specific structure to $f(x)$ will be easier in considering Gaussian Process.

$$y = f(x) + \varepsilon$$

$$f(x) = x^T w \qquad w : \text{Random variable}$$

$$\varepsilon \sim \mathbb{N}(0, \sigma_n^2)$$

$$X = (\ x_1 \quad x_2 \quad \cdots \quad x_N\ ),$$

$$Y = \begin{pmatrix} y_1 & y_2 & \cdots & y_N \end{pmatrix}^T$$

- Consider the probability of observing Y, given a weight vector w and input X. The only random variable is ε in this case. Since the observation noise ε is independent and identically distributed, we can write:

$$p(Y \mid X, w) = \prod_{i=1}^{N} p(y_i \mid x_i, w)$$

$\varepsilon = y - f(x)$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y_i - x_i^T w)^2}{2\sigma_n^2} \right)$$

$$= \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left(-\sum_{i=1}^{N} \frac{(y_i - x_i^T w)^2}{2\sigma_n^2} \right)$$

- Therefore, this conditional probability distribution is Gaussian with mean $X^T w$ and variance $\sigma_n^2 I$

$$p(Y \mid X, w) \propto \exp\left(-\frac{|Y - X^T w|^2}{2\sigma_n^2} \right) \sim \mathbb{N}(X^T w, \sigma_n^2 I) \qquad \text{(A)}$$

I is the identity matrix of order N.

The Posterior Distribution of Weight w

- Our objective is to obtain the distribution of f, which depends on weight vector w: $f(x) = x^T w$
- Let us obtain the distribution of w from data D by using Bayesian Inference (based on Bayes' Rule).
- Assuming that the prior of weight distribution is zero-mean Gaussian.

$$w \sim \mathbb{N}(0, \Sigma_p) \quad \text{where} \quad \Sigma_p = E[ww^T]$$

$$p(w) = \frac{1}{(2\pi \det \Sigma_p)^{n/2}} \exp\left(-\frac{1}{2} w^T \Sigma_p^{-1} w\right) \quad \text{(B)}$$

- Recall Bayes' Rule

$$p(x \mid y) = \frac{p(y \mid x) p(x)}{p(y)}$$

(posterior) ~ (likelihood) x (prior)

Replacing x by w and y by Y,

$$p(w \mid Y, X) \sim p(Y \mid X, w) p(w)$$

(A) (B)

From (A) and (B),

$$p(w \mid X, Y) \sim \exp\left(-\frac{|Y - X^T w|^2}{2\sigma_n^2}\right) \exp\left(-\frac{1}{2} w^T \Sigma_p^{-1} w\right)$$

$$= \exp\left(-\frac{1}{2\sigma_n^2}(Y^T Y - 2Y^T X^T w + w^T XX^T w) - \frac{1}{2} w^T \Sigma_p^{-1} w\right)$$

- We can reduce the above expression to the following quadratic form.

$$p(w \mid X, Y) \sim \exp\left(-\frac{1}{2}(w - \bar{w})^T A (w - \bar{w})\right)$$

where $A = XX^T / \sigma_n^2 + \Sigma_p^{-1}$

- Comparing the right-hand sides of the above expressions,

$$\frac{1}{\sigma_n^2} w^T XY = w^T A \bar{w}, \forall w \in \Re^n$$

$$\therefore \bar{w} = \frac{1}{\sigma_n^2} A^{-1} XY$$

- Note that, since Y is given, exp(constant x $Y^T Y$) is a constant that does not affect the above equations.

Output prediction in response to test point input x^*

- In summary, given data $D = \{X, Y\}$ with noisy observation, the posterior distribution of parameter w is a Gaussian distribution with the following mean and covariance:

$$p(w \mid X, Y) \sim \exp\left(-\frac{1}{2}(w - \bar{w})^T A (w - \bar{w}) \right)$$

$$\bar{w} = \frac{1}{\sigma_n^2} A^{-1} XY$$

$$A = \frac{1}{\sigma_n^2} XX^T + \Sigma_p^{-1}$$

- Now a test point x^* is given, what is the expected output y^*?

$$y = x^T w + \varepsilon$$

Note that both w and ε are random variables. Output y^* can be reached through many combinations of w and ε values. Recall the Chapman-Kolmogorov equation.

- Namely, $y^* = x^{*T} w \leftarrow w \leftarrow (X, Y)$

$$p(y^* \mid x^*, X, Y) = \int p(y^* \mid x^*, w) p(w \mid X, Y) dw$$

$$\exp\left(-\frac{1}{2\sigma_n^2}(y - x^{*T} w)^2 - \frac{1}{2}(w - \bar{w})^T A^{-1} (w - \bar{w}) \right)$$

- Computing this, we can find that the predictive distribution of y^* is Gaussian, again.

$$p(y^* \mid x^*, X, Y) \sim \mathbb{N}(x^{*T} \bar{w}, x^{*T} A^{-1} x^*)$$

- The mean of y^* is the posterior mean of weight. Multiplied by the test input x^*:

$$\bar{y}^* = x^{*T} \bar{w}$$

- The predictive variance is the quadratic form of the posterior covariance evaluated at x^*.

$$\sigma^* = x^{*T} A^{-1} x^*$$

23.2 Kernel Trick

- So far, we used a linear parametric model, but the input x can be projected to a high dimensional feature space.
- The Kernel Trick is applicable to Gaussian Process.

$$f(x) = x^T w \quad \rightarrow \quad f(x) = \varphi(x)^T w$$

Now $w \in \Re^m, \quad m > n$

Example: $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \Re^n \quad \rightarrow \quad \varphi(x) = \begin{pmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ \vdots \end{pmatrix} \in \Re^m$

- Define $\Phi(X) = \left(\varphi(x_1), \cdots, \varphi(x_N)\right) \in \Re^{m \times N}$

- The test point distribution is now given by

$$p(y^* \mid x^*, X, Y) \sim \mathbb{N}\left(\frac{1}{\sigma_n^2} \varphi(x^*)^T A^{-1} \Phi(X) Y, \quad \varphi(x^*)^T A^{-1} \varphi(x^*) \right)$$

where $$A = \frac{1}{\sigma_n^2} \Phi(X) \Phi(X)^T + \Sigma_p^{-1} \in \Re^{m \times m}$$

Inverse Computation

- One drawback of the above nonlinear regression based on Gaussian Process is that it entails the inversion of matrix A, which can be of high dimension.

$$A = \frac{1}{\sigma_n^2}\Phi(X)\Phi(X)^T + \Sigma_p^{-1} \in \Re^{m\times m}$$

- Inversion of m-by-m matrix, however, can be reduced to the inversion of N-by-N matrix by using the following matrix:

$$K = \Phi(x)^T \Sigma_p \Phi(x) \in \Re^{N\times N}$$

  - Consider

$$\frac{1}{\sigma_n^2}\Phi(K + \sigma_n^2 I) = \frac{1}{\sigma_n^2}\Phi(\Phi^T\Sigma_p\Phi + \sigma_n^2 I) = \frac{1}{\sigma_n^2}\Phi\Phi^T \cdot \Sigma_p\Phi + \Phi$$

$$= (A - \Sigma_p^{-1}) \cdot \Sigma_p\Phi + \Phi = A \cdot \Sigma_p\Phi - \Phi + \Phi = A\Sigma_p\Phi$$

Inverse Computation (Continued)

$$\frac{1}{\sigma_n^2}\Phi(K+\sigma_n^2 I) = A\Sigma_p\Phi$$

The posterior distribution (after observations are taken.)

$$p(y^*\,|\,x^*,X,Y) \sim \mathbb{N}\left(\frac{1}{\sigma_n^2}\varphi(x^*)^T A^{-1}\Phi(X)Y,\ \ \varphi(x^*)^T A^{-1}\varphi(x^*)\right)$$

- Pre-multiplying A^{-1} and post-multiplying $(K+\sigma_n^2 I)^{-1}$ yields

$$\frac{1}{\sigma_n^2} A^{-1}\Phi = \Sigma_p\Phi(K+\sigma_n^2 I)^{-1}$$

- Note that the matrix inversion of the left-hand side is m-by-m, while that of the right-hand side is N-by-N. (m > N)
- The mean of the posterior distribution can be computed

$$\frac{1}{\sigma_n^2}\varphi(x^*)^T A^{-1}\Phi(X)Y = \varphi(x^*)^T\Sigma_p\Phi(K+\sigma_n^2 I)^{-1}$$

- The covariance of the posterior can be reduced by using the Matrix Inversion Lemma:

$$\varphi(x^*)^T A^{-1}\varphi(x^*) = \varphi(x^*)^T\Sigma_p\varphi(x^*) - \varphi(x^*)^T\Sigma_p\Phi(K+\sigma_n^2 I)^{-1}\Phi^T\Sigma_p\varphi(x^*)$$

Prove this on your own.

Covariance Kernels

- The posterior distribution of the output in response to input x^*

$$p(y^*|x^*,X,Y) \sim \mathbb{N}\left(\varphi^{*T}\Sigma_p\Phi(K+\sigma_n^2 I)^{-1},\quad \varphi^{*T}\Sigma_p\varphi^* - \varphi^{*T}\Sigma_p\Phi(K+\sigma_n^2 I)^{-1}\Phi^T\Sigma_p\varphi^*\right)$$

where $\varphi^* = \varphi(x^*), \Phi = \Phi(X) = (\varphi(x_1),\cdots,\varphi(x_N))$

- In the above expression, φ^* and $\varphi(x_i)$ show up in quadratic form

$\varphi(x)^T\Sigma_p\varphi(x')$ where x and x' are wither x^* or x_i.

- Note that matrix $\Sigma_p = E[ww^T]$ is positive-definite and symmetric. Therefore, a kernel exists!

$\varphi(x)^T\Sigma_p\varphi(x') \rightarrow <\psi(x),\psi(x')>$ Inner Product

- Matrix Σ_p is decomposed to

$$\Sigma_p = VDV^T \qquad \text{where} \quad V = \left(v_1\cdots v_m\right),\quad D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix}, \lambda_i > 0$$

Covariance Kernels

- Define

$$\psi(x) = D^{1/2} V^T \varphi(x), \qquad \text{where } D^{1/2} = \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_m} \end{pmatrix}, \qquad \text{a square root matrix}$$

- The quadratic form can be written as an inner product of $\psi(x)$ and $\psi(x')$.

$$\varphi(x)^T \Sigma_p \varphi(x') = \underbrace{\varphi(x)^T V D^{1/2}}_{\psi(x)^T} \cdot \underbrace{D^{1/2} V^T \varphi(x')}_{\psi(x')} = <\psi(x), \psi(x')>$$

- A kernel function is defined for the covariance matrix:

$$k(x, x') \triangleq <\psi(x), \psi(x')>$$

- This type of kernel is called a covariance kernel.
- If an algorithm is given solely in terms of inner product, the kernel is more important than the corresponding feature space, which may be of infinite dimension.

23.3 Gaussian Processes in Hilbert Space (Function Space)

- The posterior distribution of output has been obtained for the parametric model:

$$f(x) = \varphi(x)^T w$$

- However, the mean and variance of the posterior distribution can be written with respect to covariance kernels $k(x, x')$, where the expanded features $\varphi(x)$ are not explicitly involved.
- This allows us to represent and analyze Gaussian processes without use of parametric models. Instead, exploiting kernels we can formulate a gaussian process directly in function space where a kernel, that is, an inner product is defined. Namely, Gaussian processes are defined in Hilbert Space.
- A Gaussian process is completely specified by its mean and covariance (kernel) functions.

$$m(x) = E[f(x)]$$

$$k(x,x') = E[(f(x) - m(x))(f(x') - m(x')]$$

- Then we write the Gaussian process as

$$f(x) \sim GP(m(x), k(x,x'))$$

Gaussian Processes in Hilbert Space

- If the function $f(x)$ has a linear-parametric representation, $f(x) = \varphi(x)^T w$,

$$E[f(x)] = \varphi(x)^T E[w] = 0$$

$$E[f(x) f(x')] = \varphi(x)^T E[ww^T]\varphi(x') = \varphi(x)^T \Sigma_p \varphi(x')$$

$$= \psi(x)^T \psi(x') = k(x, x')$$

  - Note that parameter w is mean-zero, and the covariance is a kernel function. Often times, the mean of a Gaussian process is set to zero : $f(x) \sim GP(0, k(x, x'))$.

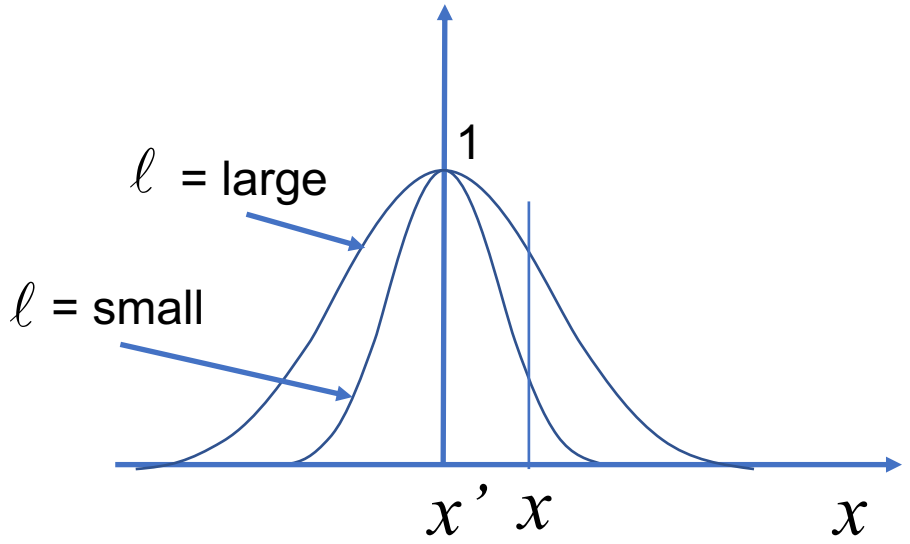
- The above argument implies that mean and covariance, the two quantities that completely characterize a Gaussian process, can be represented without use of feature vectors $\varphi(x)$.
- For any finite number of input variables, $X = (x_1, \cdots, x_N)$, an NxN covariance matrix can be formed with kernel functions.

$$\boldsymbol{f}(X) = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{pmatrix}, \quad \mathrm{cov}[\boldsymbol{f}(X), \boldsymbol{f}(X)] = K(X, X) = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix} \in \Re^{N \times N}$$

Radial-Basis Function as a Covariance Kernel

- Radial-Basis Function is one of the most widely used kernel functions for Gaussian Processes.

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



where parameter l is the characteristic length scale.
In Radial-Basis Function Network, we used a slightly different expression with $\beta = l$ and $\gamma = x'$.
If $x = x'$, the covariance kernel takes 1, the largest value.
As the characteristic length-scale l gets larger, the covariance is large in a wider range.

- Note that this kernel can be decomposed to the inner product of feature vectors, but the vector is infinite in dimension, as we discussed previously in Kernel Trick. We do not care the original feature vector $\varphi(x)$.

Prediction with Noise-Free Observations

- Our interest is to predict an output in response to a set of test inputs, $X^* = (x_1^*, \cdots, x_{N^*}^*)$ with a Gaussian Process tuned with training data, $X = (x_1, \cdots, x_N)$.
- First, let us consider the case where observations are noise free. The covariance matrix of the Gaussian process associated with inputs X and X* can be written as a block symmetric matrix of kernel functions.

$$\begin{pmatrix} \boldsymbol{f}(X) \\ \boldsymbol{f}(X^*) \end{pmatrix} \sim \mathbb{N}\left(0, \boldsymbol{K}\right), \quad \boldsymbol{K} = \begin{pmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{pmatrix} \in \Re^{(N+N^*)\times(N+N^*)}$$

- The prediction of $\boldsymbol{f}(X^*)$, that is the Gaussian distribution with particular mean and variance, can be provided as the following posterior distribution by using the Gaussian Identity.

$$\boldsymbol{f}^* \mid X^*, X, Y \sim \mathbb{N}(\bar{\boldsymbol{f}}^*, \mathrm{cov}(\boldsymbol{f}^*))$$

Mean: $\bar{\boldsymbol{f}}^* = K(X^*,X)K(X,X)^{-1}\boldsymbol{f}$

Covariance: $\mathrm{cov}(\boldsymbol{f}^*) = K(X^*,X^*) - K(X^*,X)K(X,X)^{-1}K(X,X^*)$

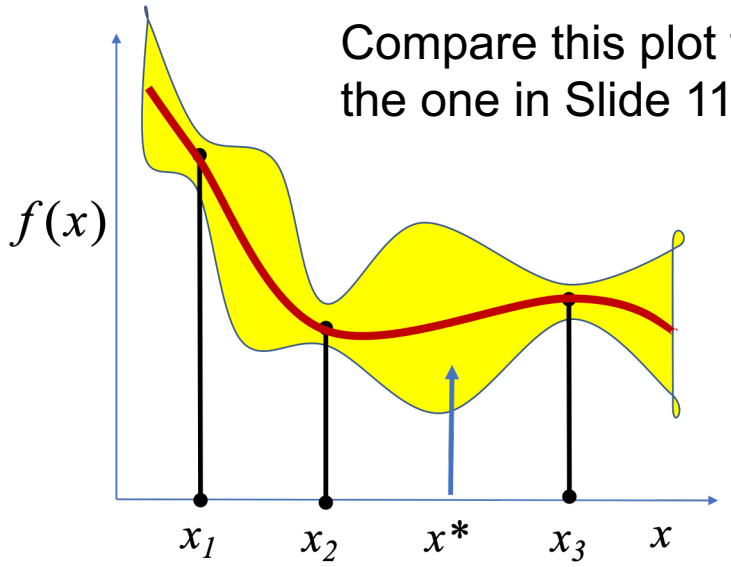
Prediction with Noisy Observations

- In realistic setting, observations are corrupted with noise.
- Assuming that observation noise ε is independent and identically distributed Gaussian with variance σ_n^2 , the prior of noisy observations is given by

$$\mathrm{cov}(\boldsymbol{y}) = K(X,X) + \sigma_n^2 I$$ ← Since the noise is independent, all the cross-terms are zero.

- Replacing $K(X,X)$ by the above covariance, the joint distribution becomes.

$$\begin{pmatrix} \boldsymbol{y}(X) \\ \boldsymbol{f}(X^*) \end{pmatrix} \sim \mathbb{N}\left(0, \begin{pmatrix} K(X,X)+\sigma_n^2 I & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{pmatrix}\right)$$



Compare this plot to the one in Slide 11.

- Using the Gaussian Identity again, we can obtain the prediction of $\boldsymbol{f}(X^*)$,

$$\boldsymbol{f}^* \mid X^*, X, Y \sim \mathbb{N}(\mu^*, C^*)$$

Mean: $$\mu^* = K(X^*,X)(K(X,X)+\sigma_n^2)^{-1} Y$$

Covariance: $$C^* = K(X^*,X^*) - K(X^*,X)(K(X,X)+\sigma_n^2 I)^{-1} K(X,X^*)$$

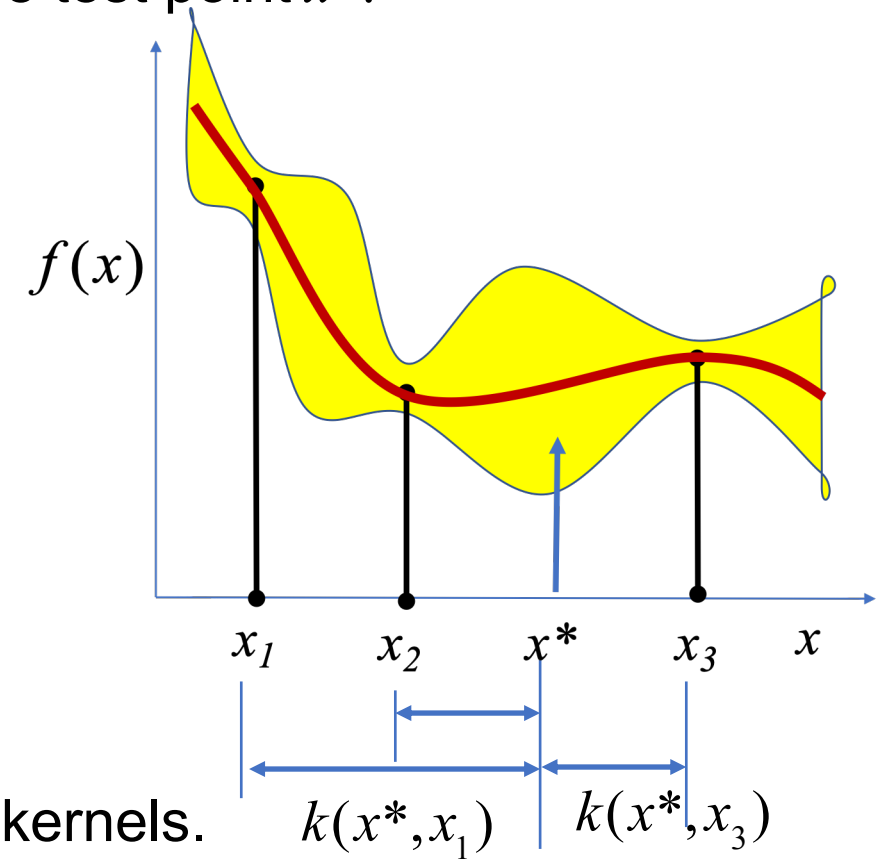
Interpretation of the Predictive Mean

- The predictive distribution of $f(x^*)$ in response to a single test point x^*.

$$f^* \mid x^*, X, Y \sim \mathbb{N}(\mu^*, C^*)$$

Mean: $\mu^* = K(x^*, X)(K(X,X) + \sigma_n^2)^{-1} Y$

where $K(x^*, X) = \left(k(x^*, x_1), k(x^*, x_2), \cdots k(x^*, x_N)\right)$



- Making the following replacement,

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix} \triangleq (K(X,X) + \sigma_n^2)^{-1} Y$$

we can find that the mean is a linear combination of the kernels.

$$\mu^* = \alpha_1 k(x^*, x_1) + \alpha_2 k(x^*, x_2) + \cdots + \alpha_N k(x^*, x_N)$$

- Recall that kernels represent how similar or close two data points are. Depending on the nearness, the contribution from each training data point is evaluated.

Reflection

- Gaussian Process (GP) model is a nonlinear, non-parametric model.
- GP predicts outputs in response to test points as a conditional Gaussian distribution with mean and variance.
- The GP variance is expressed as a kernel matrix.
- Posterior mean $\mu^* = \alpha_1 k(x^*, x_1) + \alpha_2 k(x^*, x_2) + \cdots + \alpha_N k(x^*, x_N)$ is given by a linear combination of kernels representing the nearness (similarity) between a test point and training data points.
- Drawback: It is not effective for large data, since NxN matrix inversion is involved.
- Hyperparameter tuning is required.