

10-301/601: Introduction to Machine Learning

Lecture 33 – Gaussian Processes

Henry Chai

6/17/25

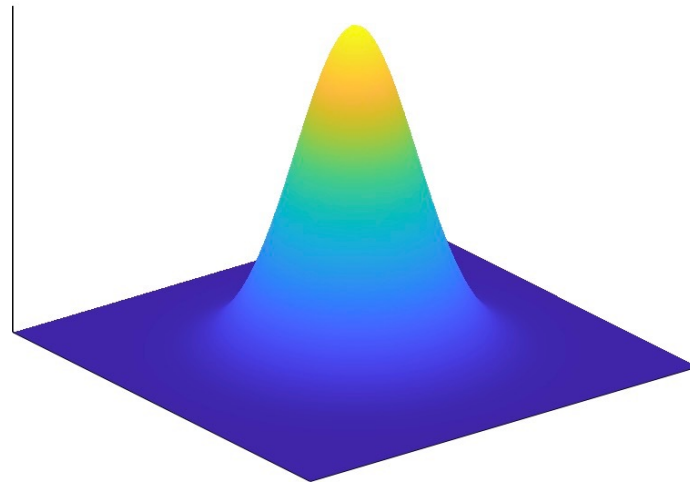
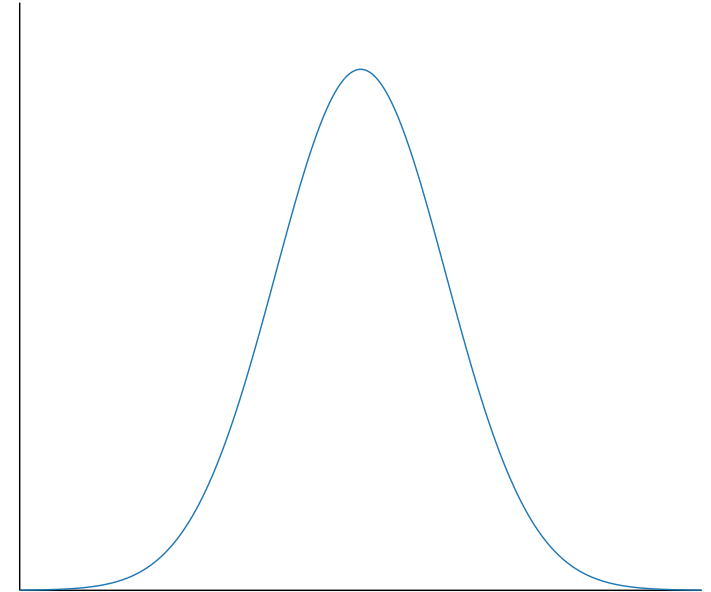
Front Matter

- Announcements
 - HW8 released 6/13, due 6/17 (today!) at 11:59 PM
 - Final on 6/20 (next Friday) at **8:30 AM** in BH A36 (here!)
 - **We will not use the full 3-hour window**
 - All topics from Lectures 17 to 30 are in-scope
 - **The final is *not* cumulative:** pre-midterm content may be referenced but will not be the *primary* focus of any question
 - You are allowed to bring one letter-size sheet of notes; you may put *whatever* you want on *both sides*

Gaussians

(Univariate) Gaussians:

$$x \sim \mathcal{N}(x; \mu = 0, \sigma^2 = 1)$$



- Multivariate Gaussians:

$$\mathbf{x} = [x_1, \dots, x_D]^T$$

$$\sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu} = \mathbf{0}_D, \Sigma = I_D)$$

Some fun facts about Gaussians

- Closure under linear transformations:

$$\text{If } \mathbf{x} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma),$$

$$\text{then } A\mathbf{x} + b \sim \mathcal{N}(A\boldsymbol{\mu} + b, A\Sigma A^T)$$

- Closure under addition

$$\text{If } \mathbf{x} \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) \text{ and } \mathbf{y} \sim \mathcal{N}(\mathbf{y}; \mathbf{m}, S),$$

$$\text{then } \mathbf{x} + \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{m}, \Sigma + S)$$

- Closure under conditioning:

$$\text{If } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$

$$\text{then } x_1 | x_2 = c \sim \mathcal{N}(x_1; \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(c - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

Outline

Gaussian process =
Bayesian linear regression + Kernels

Outline

Gaussian process =

Bayesian linear regression + Kernels

Recall: Maximum a Posteriori (MAP) Estimation

- Insight: sometimes we have *prior* information we want to incorporate into parameter estimation
- Idea: use Bayes rule to reason about the *posterior* distribution over the parameters

- MLE finds $\hat{\theta} = \operatorname{argmax}_{\theta} p(\mathcal{D}|\theta)$

- MAP finds $\hat{\theta} = \operatorname{argmax}_{\theta} p(\theta|\mathcal{D})$
 $= \operatorname{argmax}_{\theta} p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$
 $= \operatorname{argmax}_{\theta} p(\mathcal{D}|\theta)p(\theta)$

likelihood

prior

$$= \operatorname{argmax}_{\theta} \underbrace{\log p(\mathcal{D}|\theta) + \log p(\theta)}_{\text{log-posterior}}$$

MAP for Linear Regression

- If we assume a probabilistic linear model with additive Gaussian noise:

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \rightarrow \mathbf{y} \sim N(X\mathbf{w}, \sigma^2 I_N)$$

and independent, identical Gaussian priors on the weights...

$$\mathbf{w} \sim N\left(\mathbf{w}_{D+1}, \frac{\sigma^2}{\lambda} I_{D+1}\right) \rightarrow p(\mathbf{w}) \propto \exp\left(-\frac{1}{2\sigma^2} (\lambda \mathbf{w}^T \mathbf{w})\right)$$

- ... then, the MAP of $\mathbf{w}$ is the ridge regression (L2-regularized) solution!

$$\begin{aligned} \mathbf{w}_{MAP} &= \underset{\mathbf{w}}{\operatorname{argmin}} (X\mathbf{w} - \mathbf{y})^T (X\mathbf{w} - \mathbf{y}) + \lambda \mathbf{w}^T \mathbf{w} \\ &= (X^T X + \lambda I_{D+1})^{-1} X^T \mathbf{y} \end{aligned}$$

Bayesian Linear Regression

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \mathbf{w} \sim N(\mathbf{0}_{D+1}, \Sigma)$$

then,

$$\mathbf{y} \sim N(X\mathbf{0}_{D+1} + \mathbf{0}_N = \mathbf{0}_N, X\Sigma X^T + \sigma^2 I_N)$$

Bayesian Linear Regression

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \mathbf{w} \sim N(\mathbf{0}_{D+1}, \Sigma)$$

then,

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{y} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0}_{D+1} \\ \mathbf{0}_N \end{bmatrix}, \begin{bmatrix} \Sigma & ??? \\ ??? & X\Sigma X^T + \sigma^2 I_N \end{bmatrix} \right)$$

- Covariance between $\mathbf{y}$ and $\mathbf{w}$:

$$\text{Cov}(\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon}, \mathbf{w}) = \text{Cov}(X\mathbf{w}, \mathbf{w}) = X\text{Cov}(\mathbf{w}, \mathbf{w}) = X\Sigma$$

Bayesian Linear Regression

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \mathbf{w} \sim N(\mathbf{0}_{D+1}, \Sigma)$$

then,

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{y} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0}_{D+1} \\ \mathbf{0}_N \end{bmatrix}, \begin{bmatrix} \Sigma & X\Sigma \\ X\Sigma & X\Sigma X^T + \sigma^2 I_N \end{bmatrix} \right)$$

Bayesian Linear Regression

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \mathbf{w} \sim N(\mathbf{0}_{D+1}, \Sigma)$$

then,

$$\mathbf{w} \mid \mathbf{y} \sim N(\boldsymbol{\mu}_{POST}, \Sigma_{POST})$$

where

$$\begin{aligned}\boldsymbol{\mu}_{POST} &= \Sigma X^T (X \Sigma X^T + \sigma^2 I_N)^{-1} \mathbf{y}, \\ \Sigma_{POST} &= \Sigma - \Sigma X^T (X \Sigma X^T + \sigma^2 I_N)^{-1} X \Sigma\end{aligned}$$

Bayesian Linear Regression

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \mathbf{w} \sim N(\mathbf{0}_{D+1}, \Sigma)$$

then given a new test data point $\mathbf{x}'$, the prediction is

$$y' \mid \mathbf{y} = \mathbf{x}'^T \mathbf{w} \mid \mathbf{y} \sim N(\mathbf{x}'^T \boldsymbol{\mu}_{POST}, \mathbf{x}'^T \Sigma_{POST} \mathbf{x}')$$

where

$$\begin{aligned}\boldsymbol{\mu}_{POST} &= \Sigma X^T (X \Sigma X^T + \sigma^2 I_N)^{-1} \mathbf{y}, \\ \Sigma_{POST} &= \Sigma - \Sigma X^T (X \Sigma X^T + \sigma^2 I_N)^{-1} X \Sigma\end{aligned}$$

Bayesian Linear Regression

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \mathbf{w} \sim N(\mathbf{0}_{D+1}, \Sigma)$$

then given a new test data point $\mathbf{x}'$, the prediction is

$$y' \mid \mathbf{y} = \mathbf{x}'^T \mathbf{w} \mid \mathbf{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\boldsymbol{\mu}_{PRED} = \mathbf{x}'^T \Sigma X^T (X \Sigma X^T + \sigma^2 I_N)^{-1} \mathbf{y},$$

$$\Sigma_{PRED} = \mathbf{x}'^T \Sigma \mathbf{x}' - \mathbf{x}'^T \Sigma X^T (X \Sigma X^T + \sigma^2 I_N)^{-1} X \Sigma \mathbf{x}'$$

Bayesian Non-linear Regression...

$$\Phi = \begin{bmatrix} 1 & \phi(\mathbf{x}^{(1)})^T \\ 1 & \phi(\mathbf{x}^{(2)})^T \\ \vdots & \vdots \\ 1 & \phi(\mathbf{x}^{(N)})^T \end{bmatrix}$$

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = \Phi \boldsymbol{\omega} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \boldsymbol{\omega} \sim N(\mathbf{0}_{D'+1}, \Sigma)$$

then given a new test data point $\mathbf{x}'$, the prediction is

$$y' | \mathbf{y} = \phi(\mathbf{x}')^T \boldsymbol{\omega} | \mathbf{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\boldsymbol{\mu}_{PRED} = \phi(\mathbf{x}')^T \Sigma \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I_N)^{-1} \mathbf{y},$$

$$\begin{aligned} \Sigma_{PRED} &= \phi(\mathbf{x}')^T \Sigma \phi(\mathbf{x}') - \phi(\mathbf{x}')^T \Sigma \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I_N)^{-1} \Phi \Sigma \phi(\mathbf{x}') \end{aligned}$$

Bayesian Non-linear Regression can be “kernelized”

$$\Phi = \begin{bmatrix} 1 & \phi(\mathbf{x}^{(1)})^T \\ 1 & \phi(\mathbf{x}^{(2)})^T \\ \vdots & \vdots \\ 1 & \phi(\mathbf{x}^{(N)})^T \end{bmatrix}$$

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = \Phi \boldsymbol{\omega} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \boldsymbol{\omega} \sim N(\mathbf{0}_{D'+1}, \Sigma)$$

then given a new test data point $\mathbf{x}'$, the prediction is

$$\mathbf{y}' \mid \mathbf{y} = \phi(\mathbf{x}')^T \boldsymbol{\omega} \mid \mathbf{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\boldsymbol{\mu}_{PRED} = \phi(\mathbf{x}')^T \Sigma \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I_N)^{-1} \mathbf{y},$$

$$\Sigma_{PRED}$$

$$= \phi(\mathbf{x}')^T \Sigma \phi(\mathbf{x}') - \phi(\mathbf{x}')^T \Sigma \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I_N)^{-1} \Phi \Sigma \phi(\mathbf{x}')$$

- Define a **kernel function** as

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \Sigma \phi(\mathbf{x}')$$

Bayesian Linear Regression can be kernelized!

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = \Phi \boldsymbol{\omega} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \boldsymbol{\omega} \sim N(\mathbf{0}_{D'+1}, \Sigma)$$

then given a new test data point $\mathbf{x}'$, the prediction is

$$y' \mid \mathbf{y} = \phi(\mathbf{x}')^T \boldsymbol{\omega} \mid \mathbf{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\boldsymbol{\mu}_{PRED} = K(\mathbf{x}', X)(K(X, X) + \sigma^2 I_N)^{-1} \mathbf{y},$$

$$\Sigma_{PRED} = K(\mathbf{x}', \mathbf{x}') - K(\mathbf{x}', X)(K(X, X) + \sigma^2 I_N)^{-1} K(X, \mathbf{x})$$

- Define the kernel function to be

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \Sigma \phi(\mathbf{x}')$$

Wait, what happened to the weights?

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = \Phi \boldsymbol{\omega} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \boldsymbol{\omega} \sim N(\mathbf{0}_{D'+1}, \Sigma)$$

then given a new test data point $\mathbf{x}'$, the prediction is

$$y' \mid \mathbf{y} = \phi(\mathbf{x}')^T \boldsymbol{\omega} \mid \mathbf{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\boldsymbol{\mu}_{PRED} = K(\mathbf{x}', X)(K(X, X) + \sigma^2 I_N)^{-1} \mathbf{y},$$

$$\Sigma_{PRED} = K(\mathbf{x}', \mathbf{x}') - K(\mathbf{x}', X)(K(X, X) + \sigma^2 I_N)^{-1} K(X, \mathbf{x})$$

- Define the kernel function to be

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \Sigma \phi(\mathbf{x}')$$

Outline

Gaussian process =
Bayesian linear regression + Kernels

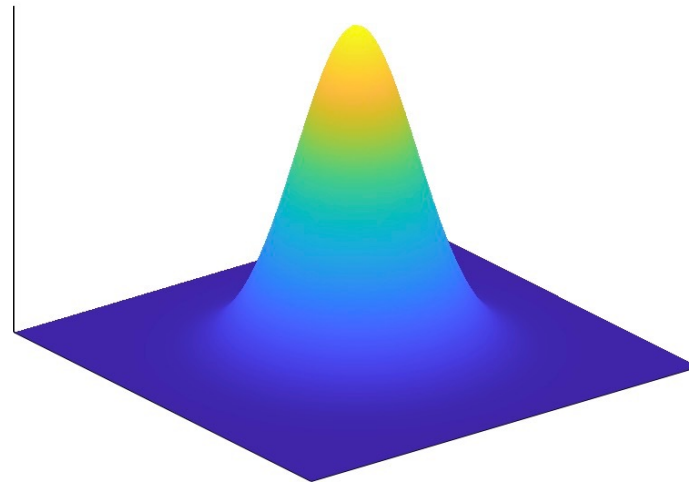
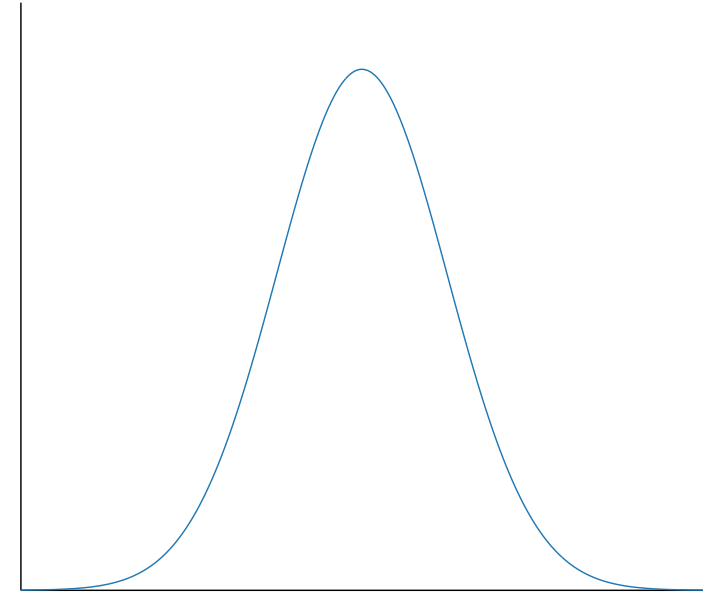
A new perspective

Gaussian process =
The extension of a Gaussian
distribution to functions

Gaussians

- (Univariate) Gaussians:

$$x \sim \mathcal{N}(x; \mu = 0, \sigma^2 = 1)$$

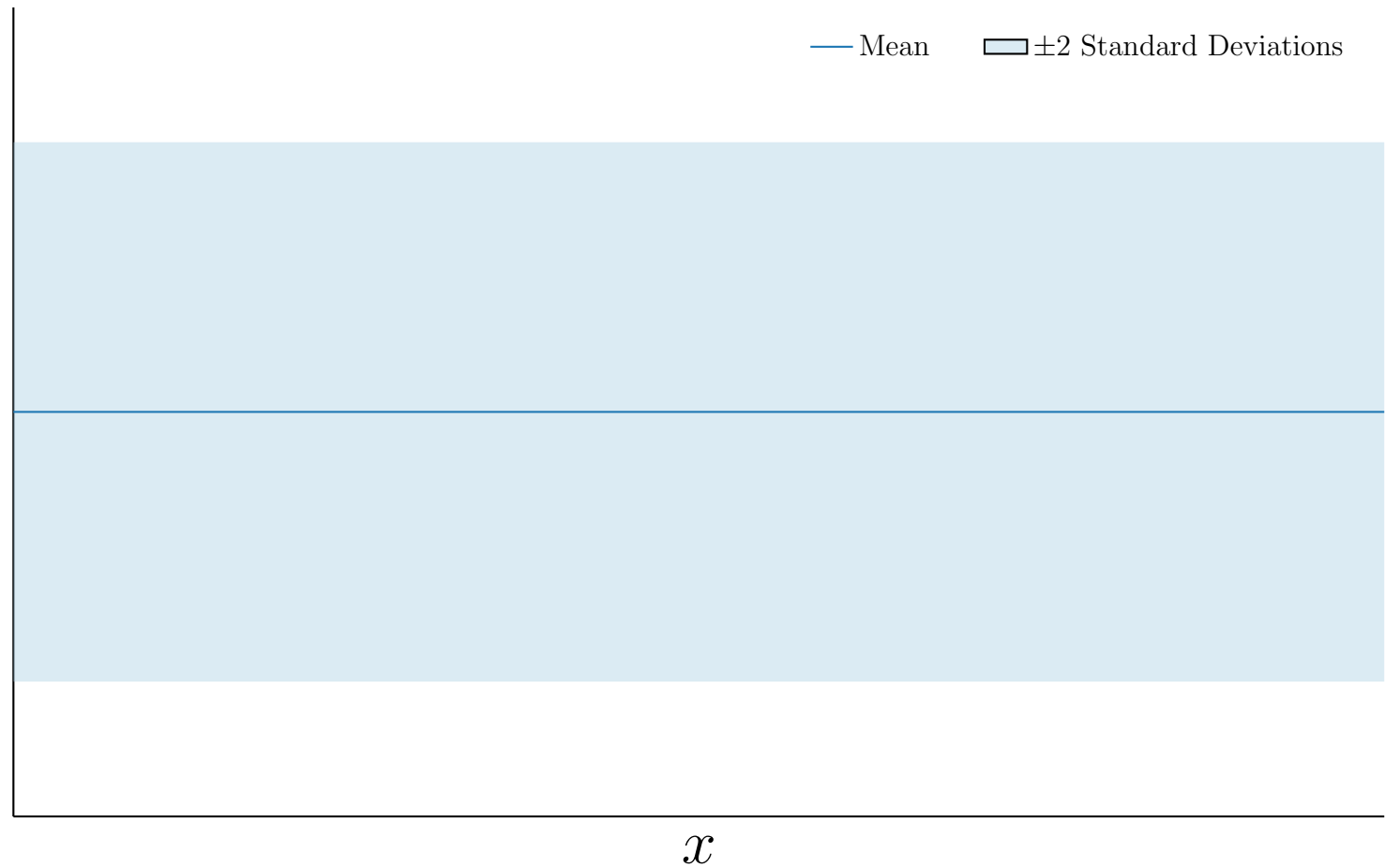


- Multivariate Gaussians:

$$\begin{aligned} \mathbf{x} &= [x_1, \dots, x_D]^T \\ &\sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu} = \mathbf{0}_D, \Sigma = I_D) \end{aligned}$$

Gaussian Process (GP)

$$f: \mathbb{R}^p \mapsto \mathbb{R} \sim \mathcal{GP}(f; \mu(x), \Sigma(x, x'))$$

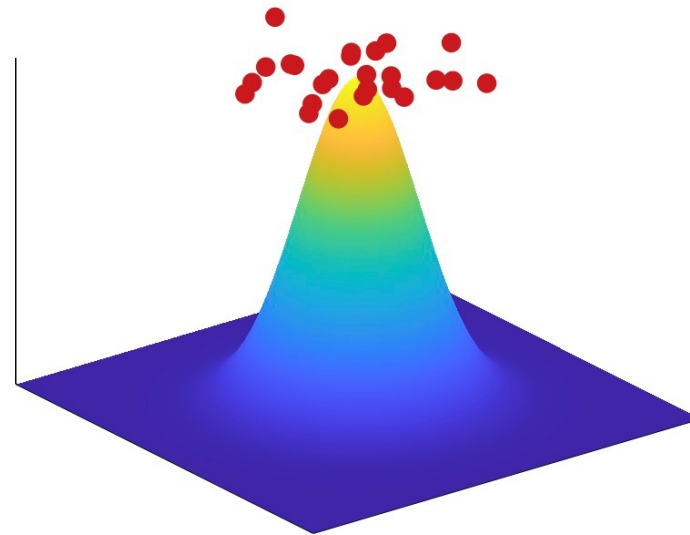
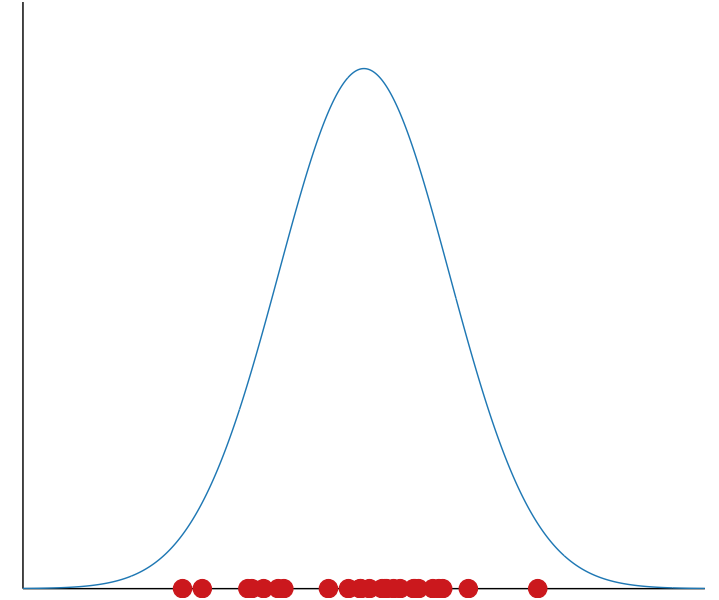


$$f \sim \mathcal{GP}(\mu, \Sigma) \rightarrow f(x) \sim \mathcal{N}(\mu(x), \Sigma(x, x))$$

Gaussians

- (Univariate) Gaussians:

$$x \sim \mathcal{N}(x; \mu = 0, \sigma^2 = 1)$$



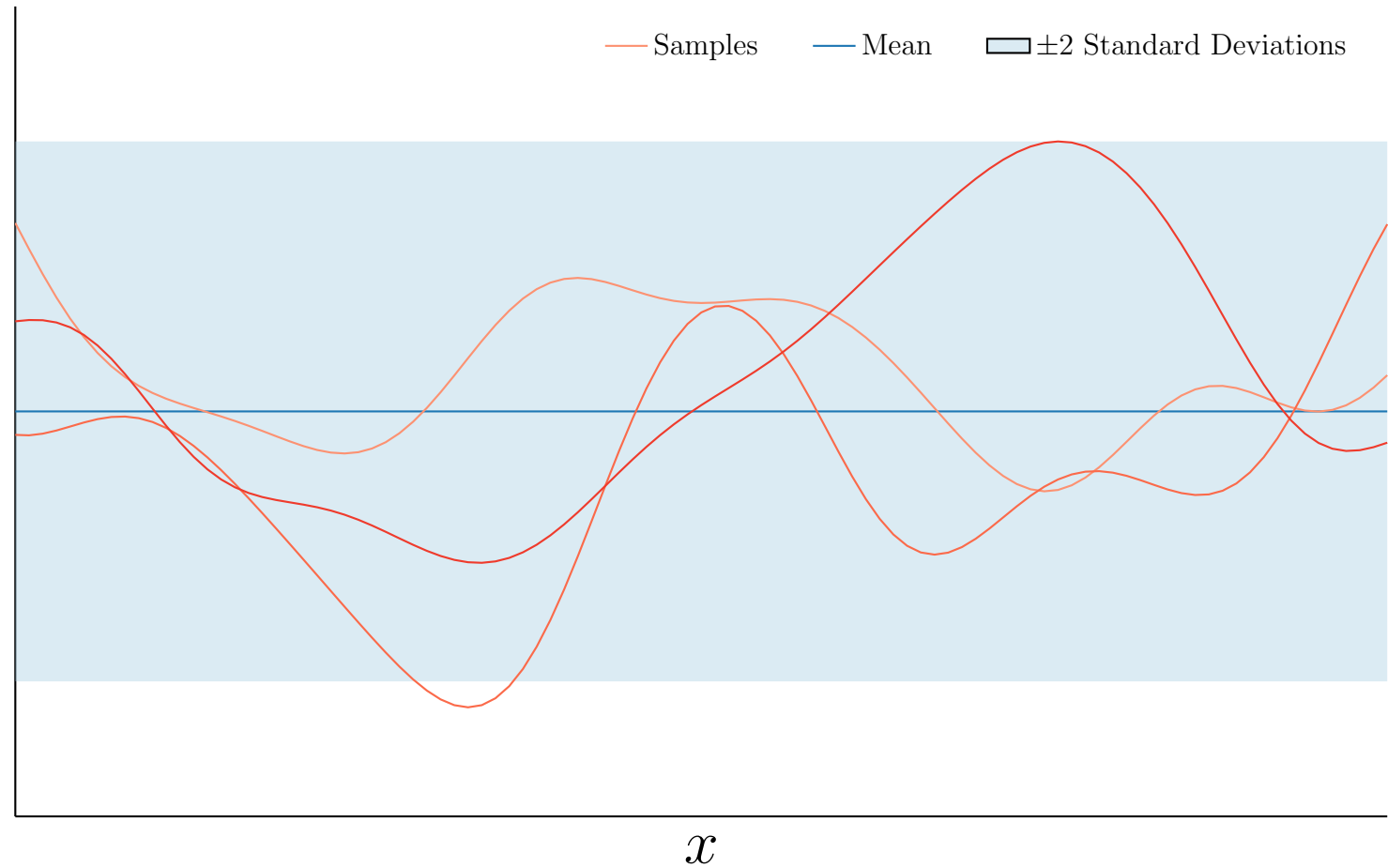
- Multivariate Gaussians:

$$\mathbf{x} = [x_1, \dots, x_D]^T$$

$$\sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu} = \mathbf{0}_D, \Sigma = I_D)$$

Gaussian Process (GP)

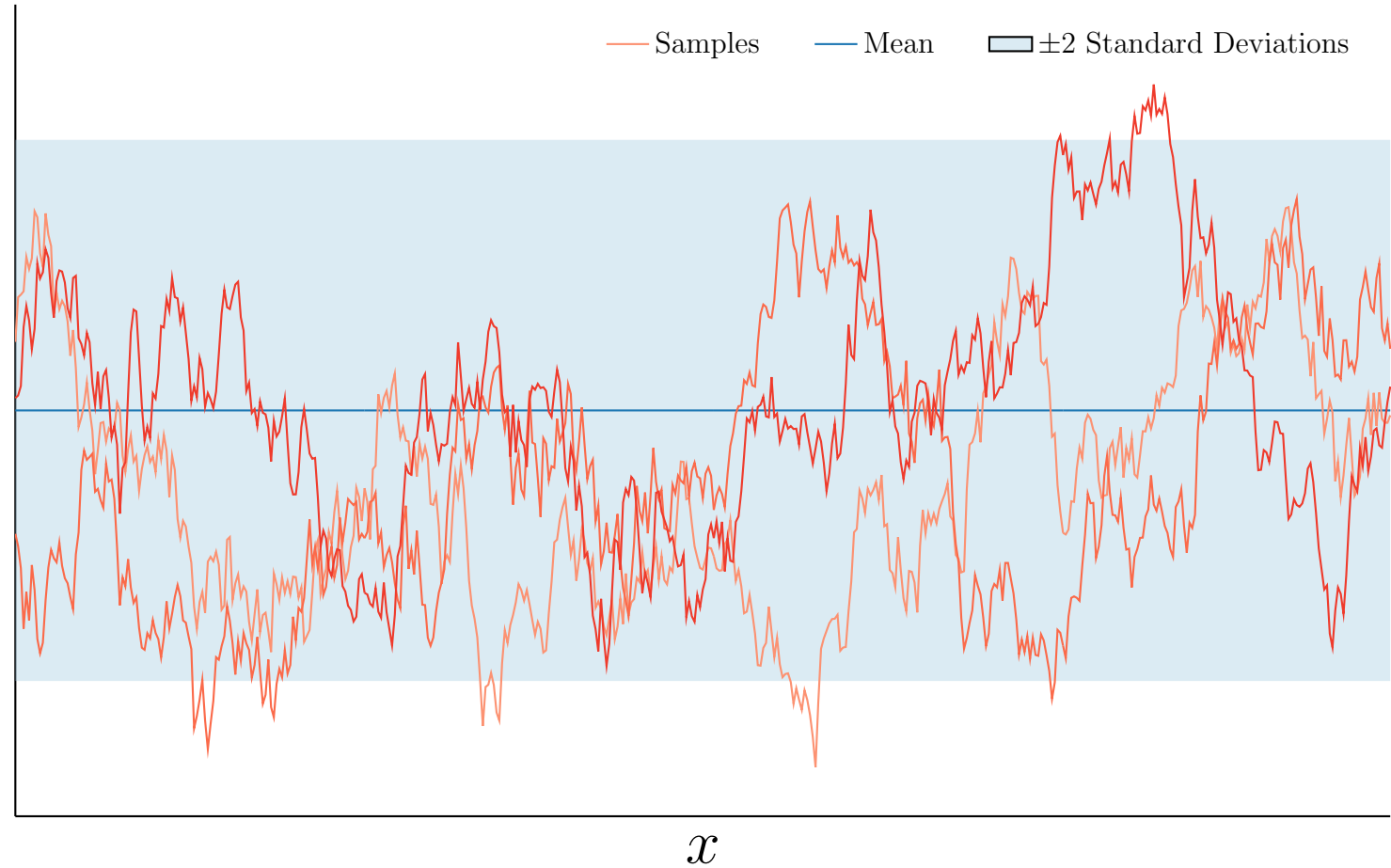
$$f: \mathbb{R}^p \mapsto \mathbb{R} \sim \mathcal{GP}(f; \mu(x) = 0, \Sigma(x, x') = \exp(-(x - x')^2))$$



$$f \sim \mathcal{GP}(\mu, \Sigma) \rightarrow f(x) \sim \mathcal{N}(\mu(x), \Sigma(x, x))$$

Gaussian Process (GP)

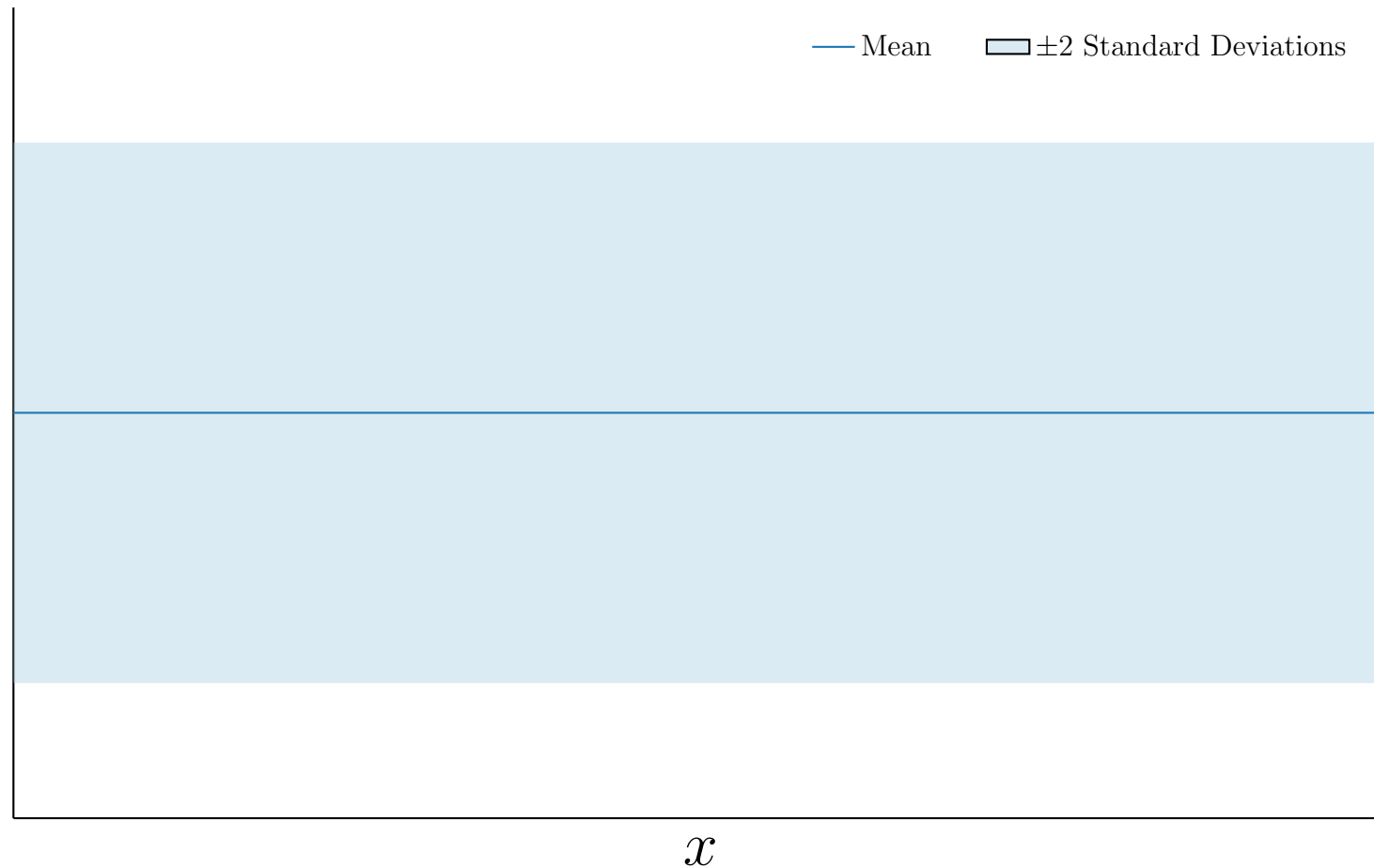
$$f: \mathbb{R}^p \mapsto \mathbb{R} \sim \mathcal{GP}(f; \mu(x) = 0, \Sigma(x, x') = \exp(-|x - x'|))$$



$$f \sim \mathcal{GP}(\mu, \Sigma) \rightarrow f(x) \sim \mathcal{N}(\mu(x), \Sigma(x, x))$$

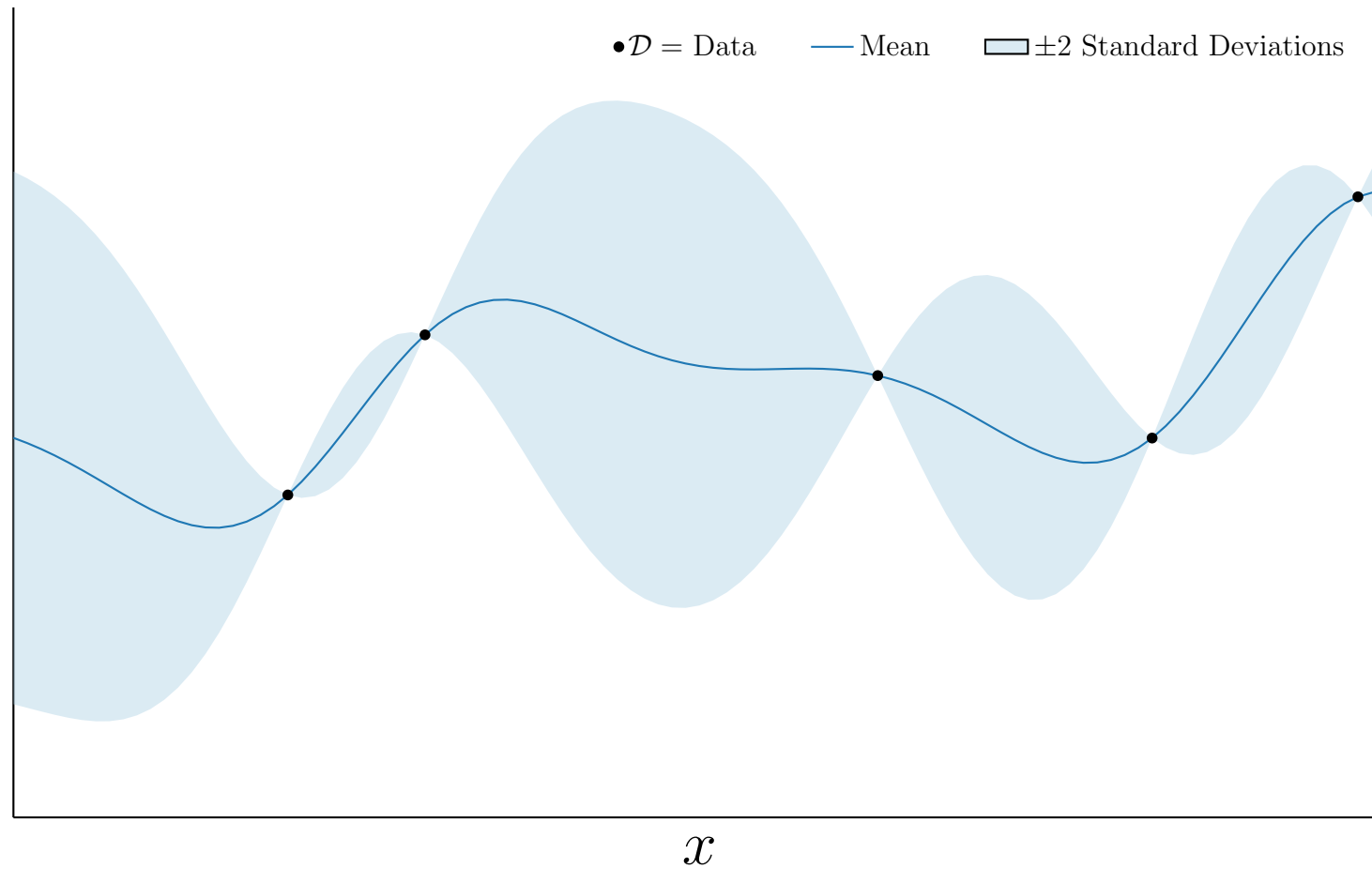
GP Prior

$$f: \mathbb{R}^p \mapsto \mathbb{R} \sim \mathcal{GP}(f; \mu(x) = 0, \Sigma(x, x') = \exp(-(x - x')^2))$$



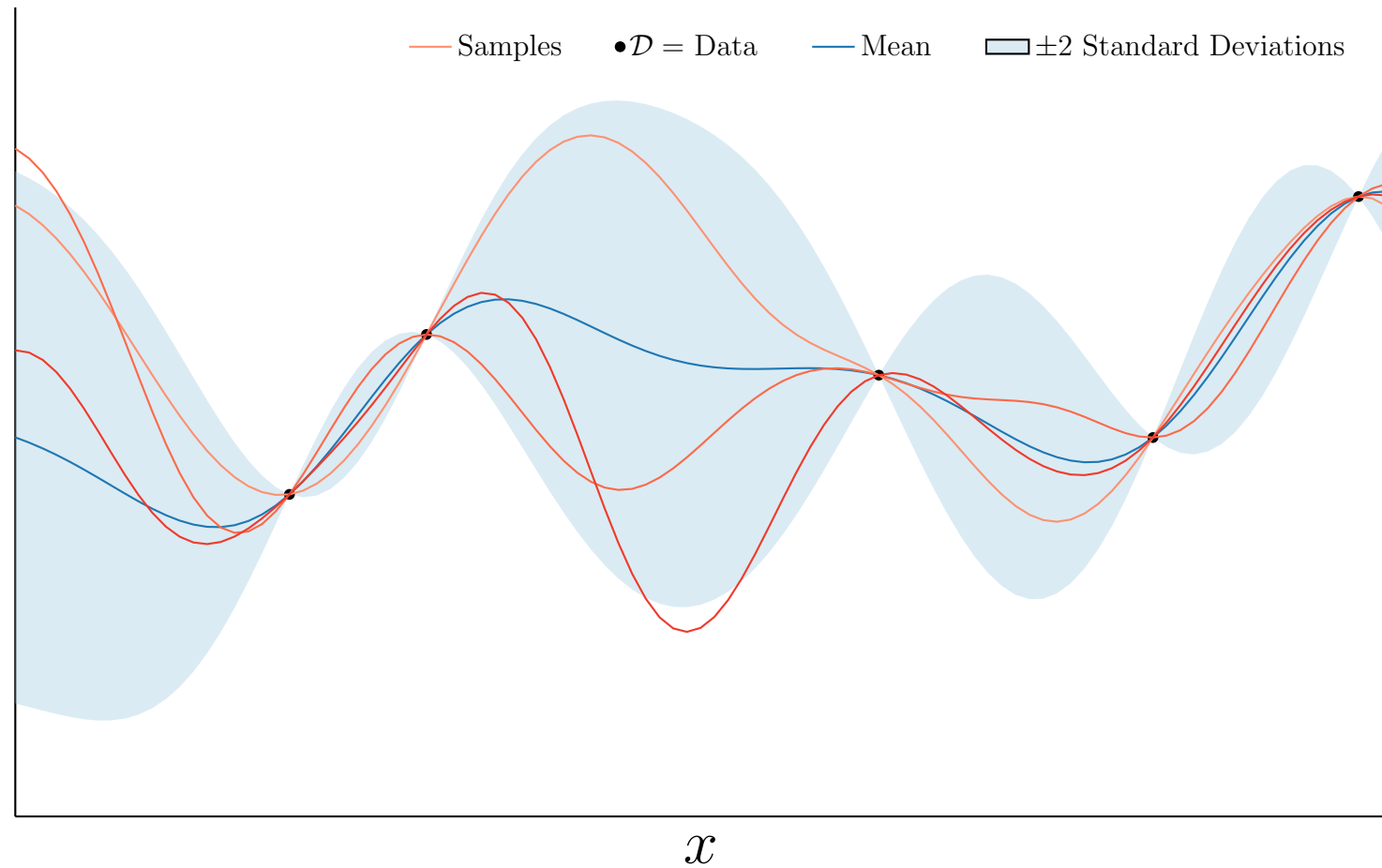
GP Posterior

$$f \mid \mathcal{D} \sim \mathcal{GP}(f; \mu_{\mathcal{D}}, \Sigma_{\mathcal{D}})$$

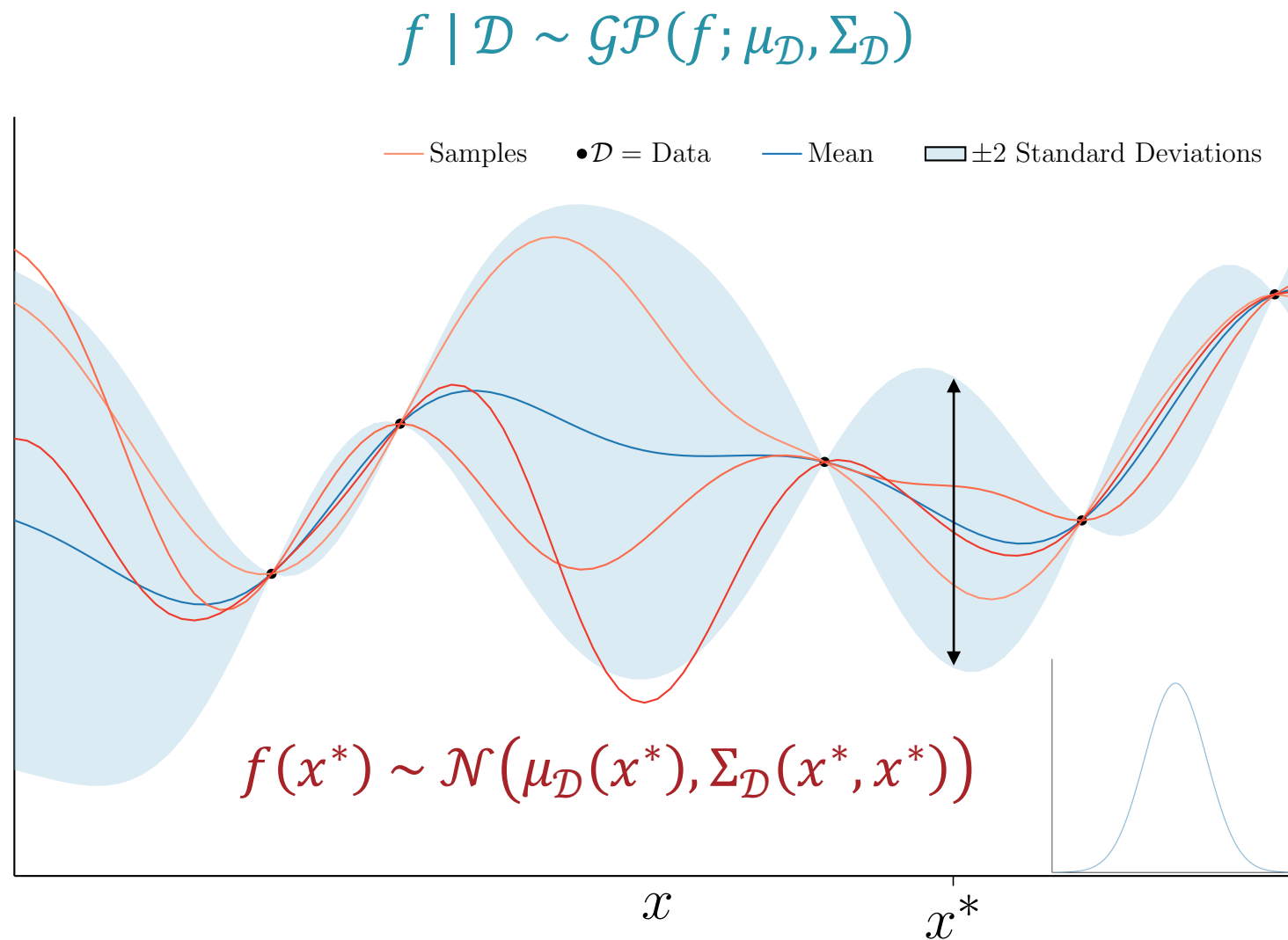


GP Posterior

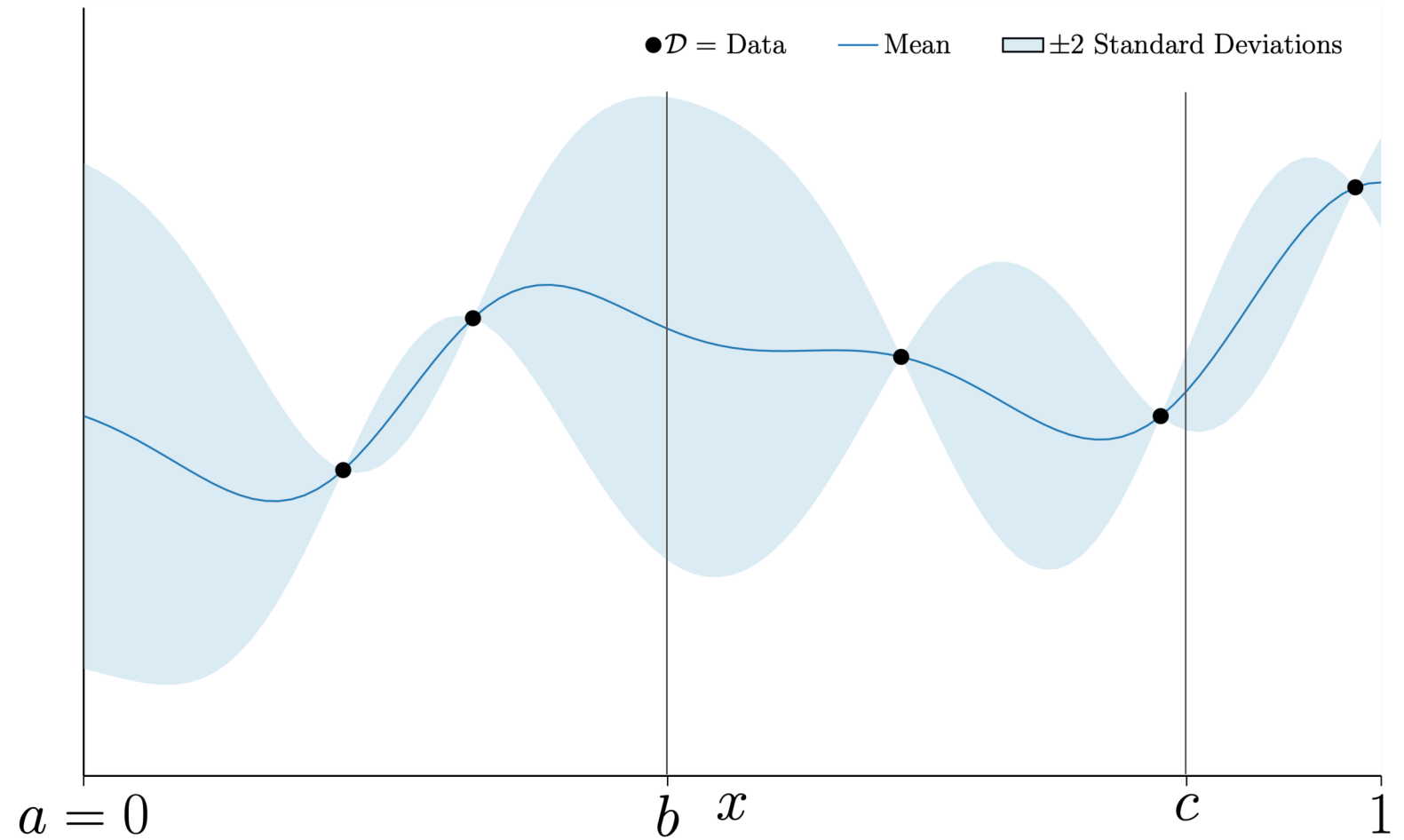
$$f \mid \mathcal{D} \sim \mathcal{GP}(f; \mu_{\mathcal{D}}, \Sigma_{\mathcal{D}})$$



GP Posterior

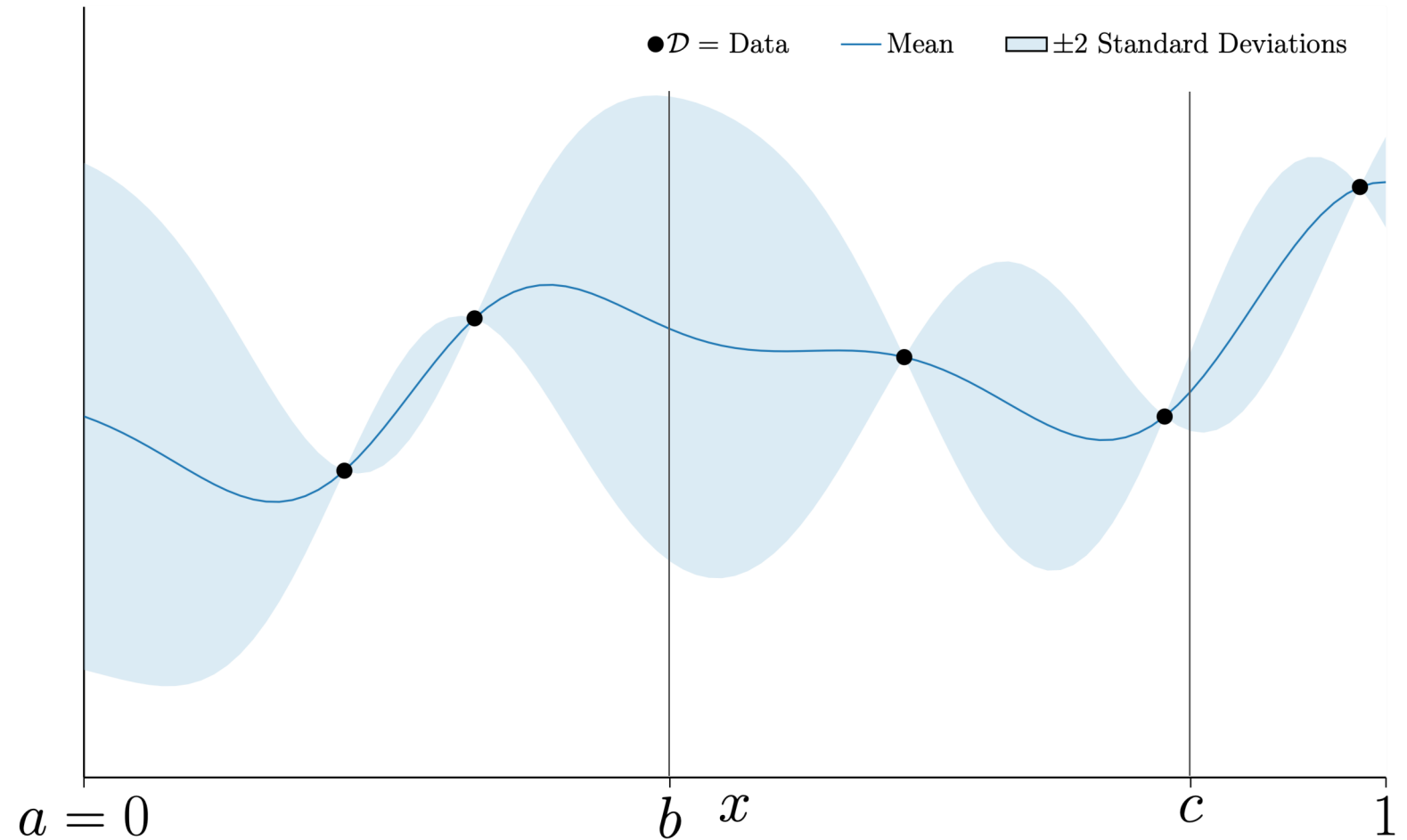


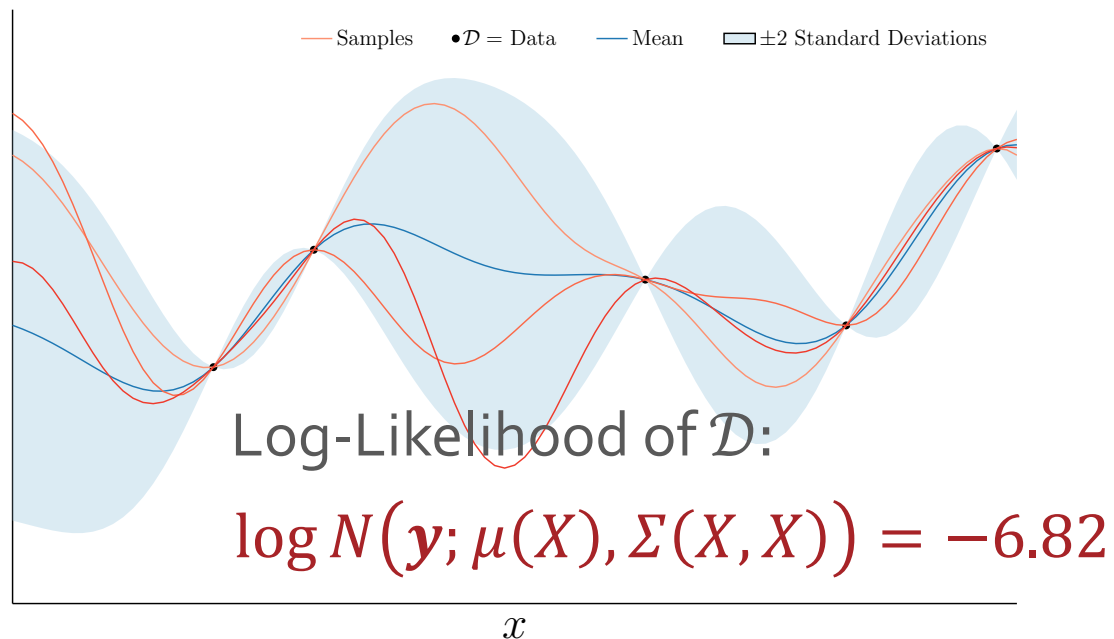
Active Learning



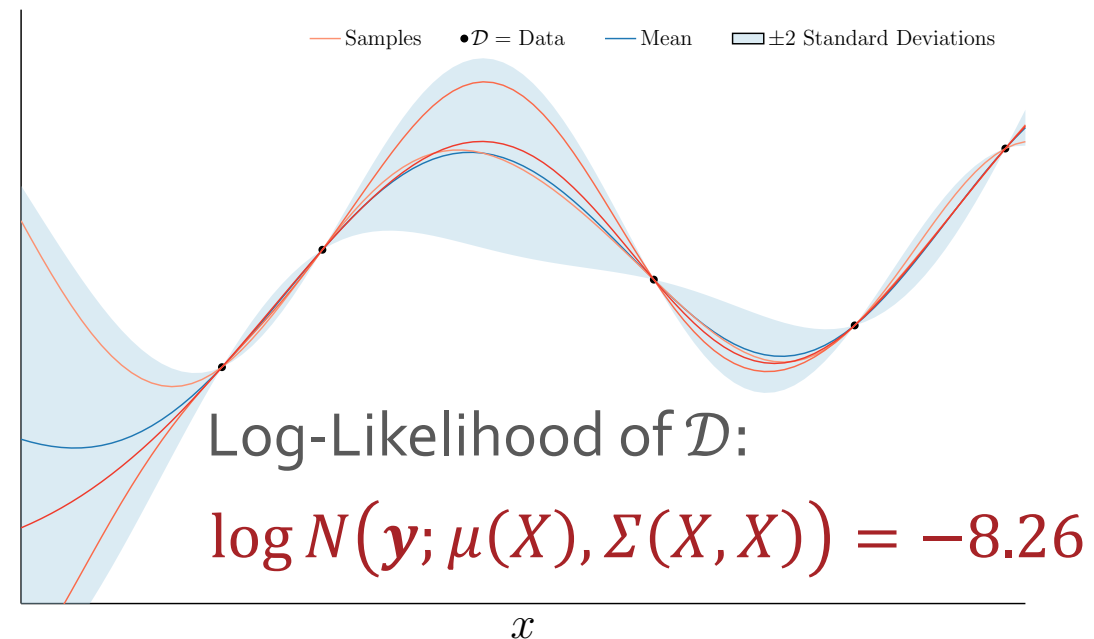
Suppose you can add one data point to your training dataset.

Which value of x would you add and why?





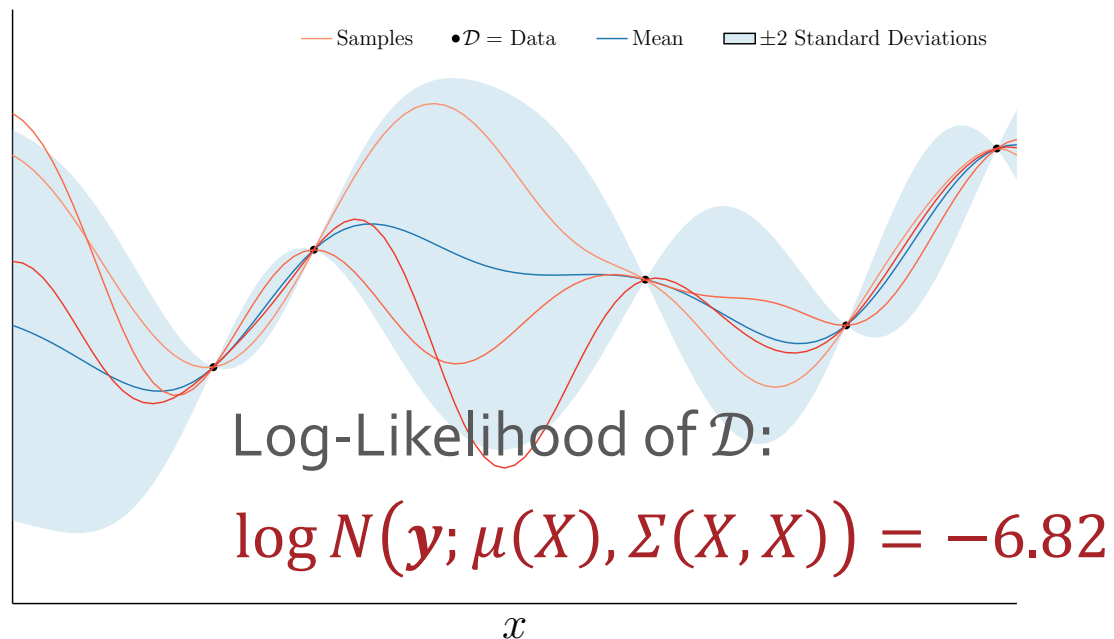
$$f \sim \mathcal{GP} \left(f; 0, (1^2) \exp \left(-\frac{(x - x')^2}{1^2} \right) \right)$$



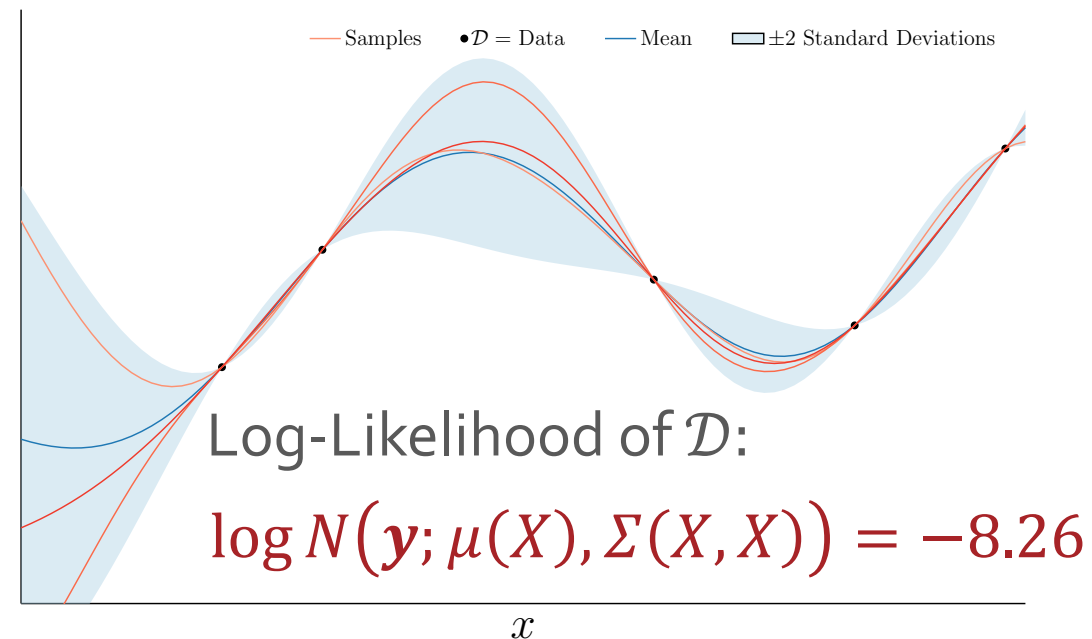
$$f \sim \mathcal{GP} \left(f; 0, (2^2) \exp \left(-\frac{(x - x')^2}{2^2} \right) \right)$$

Kernel Hyperparameters

- Can be set via MLE
- As long as μ and Σ are differentiable, the log-likelihood is differentiable with respect to the kernel hyperparameters



$$f \sim \mathcal{GP}\left(f; 0, (1^2) \exp\left(-\frac{(x - x')^2}{1^2}\right)\right)$$



$$f \sim \mathcal{GP}\left(f; 0, (2^2) \exp\left(-\frac{(x - x')^2}{2^2}\right)\right)$$

Wait doesn't this always get zero training error???

Noise

- Assume a linear model with additive Gaussian noise and a zero-mean Gaussian prior on the weights:

$$\mathbf{y} = \Phi \boldsymbol{\omega} + \boldsymbol{\epsilon} \text{ where } \boldsymbol{\epsilon} \sim N(\mathbf{0}_N, \sigma^2 I_N) \text{ and } \boldsymbol{\omega} \sim N(\mathbf{0}_{D'+1}, \Sigma)$$

then given a new test data point $\mathbf{x}'$, the prediction is

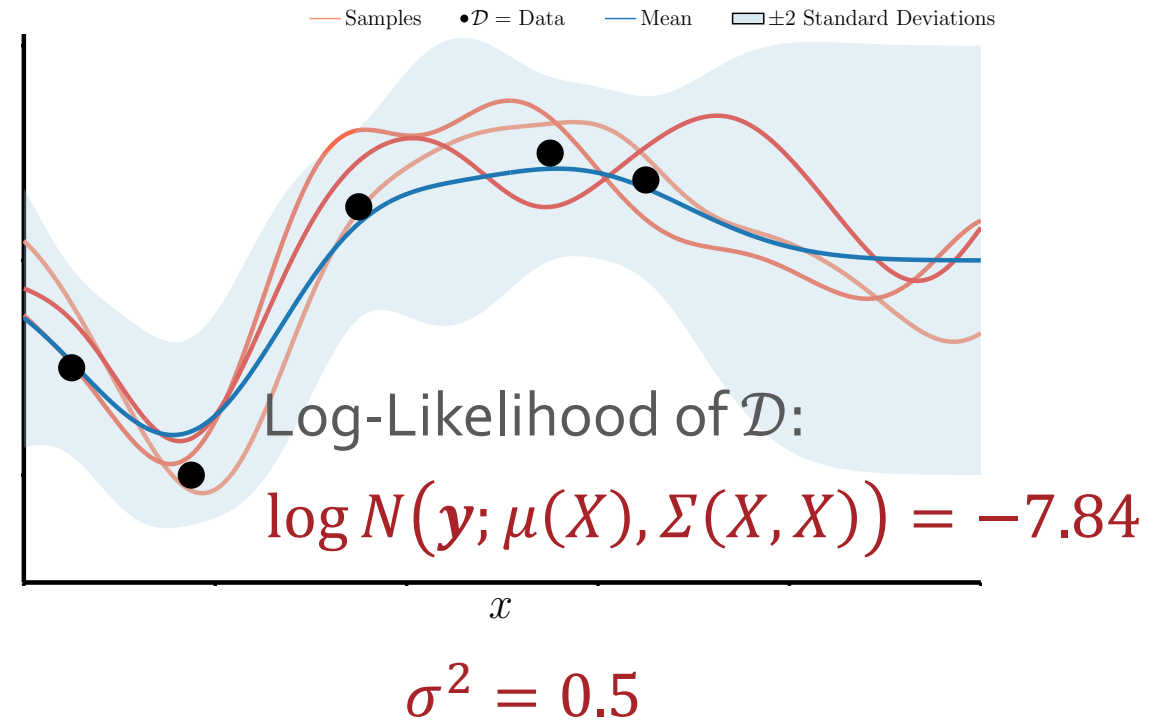
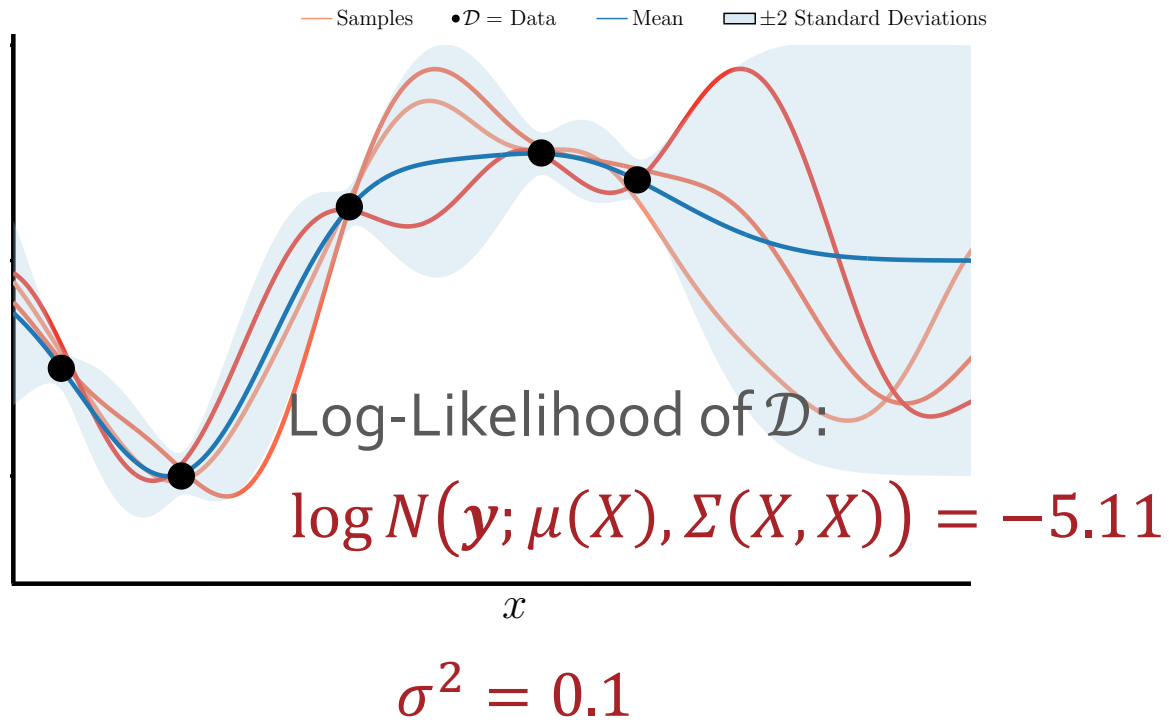
$$y' | \mathbf{y} = \phi(\mathbf{x}')^T \boldsymbol{\omega} | \mathbf{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\boldsymbol{\mu}_{PRED} = K(\mathbf{x}', X)(K(X, X) + \sigma^2 I_N)^{-1} \mathbf{y},$$

$$\Sigma_{PRED} = K(\mathbf{x}', \mathbf{x}') - K(\mathbf{x}', X)(K(X, X) + \sigma^2 I_N)^{-1} K(X, \mathbf{x})$$

- σ^2 is another hyperparameter we can tune
 - $\sigma^2 = 0$ is a noiseless fit: the mean will always pass through the observations exactly; $\sigma^2 > 0$ allows for deviations



Noise