

10-315: Introduction to Machine Learning

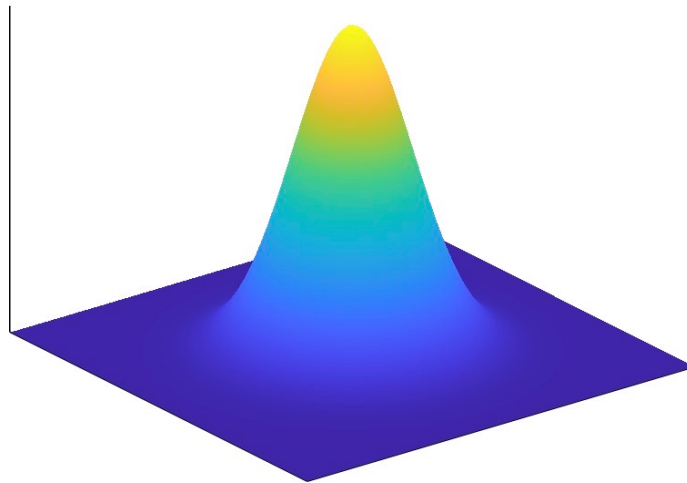
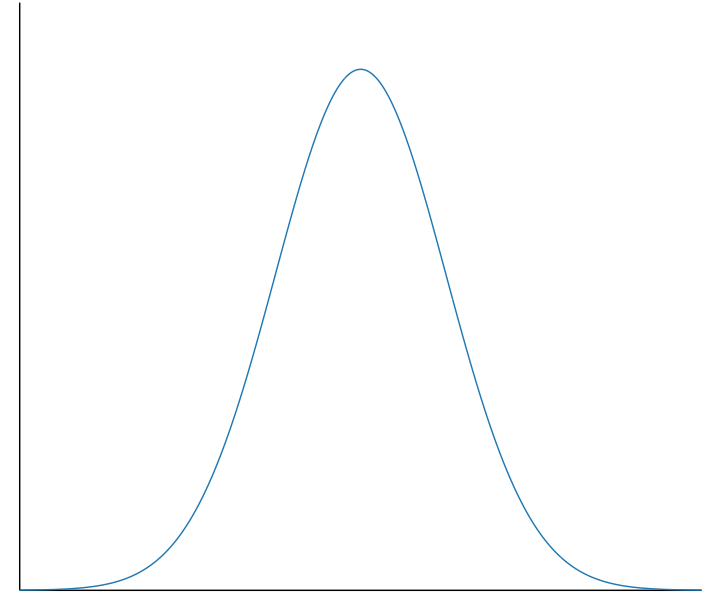
Lecture 23 – Gaussian Processes

Henry Chai

4/18/22

Gaussians

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



- Multivariate Gaussians:

$$\vec{x} = [x_1, \dots, x_p]^T$$
$$\sim \mathcal{N}(\vec{x}; \vec{\mu} = \vec{0}_p, \Sigma = I_p)$$

Some fun facts about Gaussians

- Closure under linear transformations:

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

$$\text{then } C\vec{x} + b \sim \mathcal{N}(C\vec{\mu} + b, C\Sigma C^T)$$

- Closure under addition

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

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

- Closure under conditioning:

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

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

Some old
friends

Gaussian process =
Bayesian linear regression + Kernels

Some old
friends

Gaussian process =

Bayesian linear regression + Kernels

Recall: MAP for Linear Regression

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

- If we assume a linear model with additive Gaussian noise

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \rightarrow \vec{y} \sim N(A\vec{\beta}, \sigma^2 I_n)$$

and a Gaussian prior on the weights...

$$\vec{\beta} \sim N\left(\vec{0}_{p+1}, \frac{\sigma^2}{\lambda} I_{p+1}\right) \rightarrow p(\vec{\beta}) \propto \exp\left(-\frac{1}{2\sigma^2} (\lambda \vec{\beta}^T \vec{\beta})\right)$$

- ... then, the MAP of $\vec{\beta}$ is the ridge regression solution!

$$\begin{aligned} \vec{\beta}_{MAP} &= \underset{\vec{\beta}}{\operatorname{argmin}} (A\vec{\beta} - \vec{y})^T (A\vec{\beta} - \vec{y}) + \lambda \vec{\beta}^T \vec{\beta} \\ &= (A^T A + \lambda I_{p+1})^{-1} A^T \vec{y} \end{aligned}$$

Bayesian Linear Regression

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

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

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p+1}, \Sigma)$$

- then,

$$\vec{y} \sim N(A\vec{0}_{p+1} + \vec{0}_n, A\Sigma A^T + \sigma^2 I_n)$$

Bayesian Linear Regression

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

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

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p+1}, \Sigma)$$

- then,

$$\vec{y} \sim N(\vec{0}_n, A\Sigma A^T + \sigma^2 I_n)$$

Bayesian Linear Regression

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

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

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p+1}, \Sigma)$$

- then,

$$\begin{bmatrix} \vec{\beta} \\ \vec{y} \end{bmatrix} \sim N \left(\begin{bmatrix} \vec{0}_{p+1} \\ \vec{0}_n \end{bmatrix}, \begin{bmatrix} \Sigma & ??? \\ ??? & A\Sigma A^T + \sigma^2 I_n \end{bmatrix} \right)$$

- Covariance between $\vec{y}$ and $\vec{\beta}$:

$$\text{Cov}(\vec{y}, \vec{\beta}) = \text{Cov}(A\vec{\beta} + \vec{\epsilon}, \vec{\beta}) = \text{Cov}(A\vec{\beta}, \vec{\beta}) = A\text{Cov}(\vec{\beta}, \vec{\beta}) = A\Sigma$$

Bayesian Linear Regression

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

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

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p+1}, \Sigma)$$

- then,

$$\begin{bmatrix} \vec{\beta} \\ \vec{y} \end{bmatrix} \sim N \left(\begin{bmatrix} \vec{0}_{p+1} \\ \vec{0}_n \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma A^T \\ A\Sigma & A\Sigma A^T + \sigma^2 I_n \end{bmatrix} \right)$$

- Covariance between $\vec{y}$ and $\vec{\beta}$:

$$\text{Cov}(\vec{y}, \vec{\beta}) = \text{Cov}(A\vec{\beta} + \vec{\epsilon}, \vec{\beta}) = \text{Cov}(A\vec{\beta}, \vec{\beta}) = A\text{Cov}(\vec{\beta}, \vec{\beta}) = A\Sigma$$

Bayesian Linear Regression

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

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

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p+1}, \Sigma)$$

- then,

$$\vec{\beta}|\vec{y} \sim N(\vec{\mu}_{POST}, \Sigma_{POST})$$

where

$$\begin{aligned} \vec{\mu}_{POST} &= \Sigma A^T (A \Sigma A^T + \sigma^2 I_n)^{-1} \vec{y}, \\ \Sigma_{POST} &= \Sigma - \Sigma A^T (A \Sigma A^T + \sigma^2 I_n)^{-1} A \Sigma \end{aligned}$$

Bayesian Linear Regression

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

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

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p+1}, \Sigma)$$

- then given a new test data point $\vec{x}^*$, the prediction is

$$y^* | \vec{y} = \vec{x}^{*T} \vec{\beta} | \vec{y} \sim N(\vec{x}^{*T} \vec{\mu}_{POST}, \vec{x}^{*T} \Sigma_{POST} \vec{x}^*)$$

where

$$\begin{aligned} \vec{\mu}_{POST} &= \Sigma A^T (A \Sigma A^T + \sigma^2 I_n)^{-1} \vec{y}, \\ \Sigma_{POST} &= \Sigma - \Sigma A^T (A \Sigma A^T + \sigma^2 I_n)^{-1} A \Sigma \end{aligned}$$

Bayesian Linear Regression

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

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

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p+1}, \Sigma)$$

- then given a new test data point $\vec{x}^*$, the prediction is

$$y^* | \vec{y} = \vec{x}^{*T} \vec{\beta} | \vec{y} \sim N(\vec{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\vec{\mu}_{PRED} = \vec{x}^{*T} \Sigma A^T (A \Sigma A^T + \sigma^2 I_n)^{-1} \vec{y},$$

$$\Sigma_{PRED} = \vec{x}^{*T} \Sigma \vec{x}^* - \vec{x}^{*T} \Sigma A^T (A \Sigma A^T + \sigma^2 I_n)^{-1} A \Sigma \vec{x}^*$$

Some old
friends

Gaussian process =

Bayesian linear regression + Kernels

Some old
friends

Gaussian process =
Bayesian linear regression + Kernels

Bayesian Linear Regression...

$$A = \begin{bmatrix} 1 & \vec{x}_1^T \\ 1 & \vec{x}_2^T \\ \vdots & \vdots \\ 1 & \vec{x}_n^T \end{bmatrix}$$

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

$$\vec{y} = A\vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p+1}, \Sigma)$$

- then given a new test data point $\vec{x}^*$, the prediction is

$$y^* | \vec{y} = \vec{x}^{*T} \vec{\beta} | \vec{y} \sim N(\vec{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\vec{\mu}_{PRED} = \vec{x}^{*T} \Sigma A^T (A \Sigma A^T + \sigma^2 I_n)^{-1} \vec{y},$$

$$\Sigma_{PRED} = \vec{x}^{*T} \Sigma \vec{x}^* - \vec{x}^{*T} \Sigma A^T (A \Sigma A^T + \sigma^2 I_n)^{-1} A \Sigma \vec{x}^*$$

Bayesian Linear Regression can be kernelized!

$$\Phi = \begin{bmatrix} 1 & \phi(\vec{x}_1)^T \\ 1 & \phi(\vec{x}_2)^T \\ \vdots & \vdots \\ 1 & \phi(\vec{x}_n)^T \end{bmatrix}$$

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

$$\vec{y} = \Phi \vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p'+1}, \Sigma)$$

- then given a new test data point $\vec{x}^*$, the prediction is

$$y^* | \vec{y} = \phi(\vec{x}^*)^T \vec{\beta} | \vec{y} \sim N(\vec{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\vec{\mu}_{PRED} = \phi(\vec{x}^*)^T \Sigma \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I_n)^{-1} \vec{y},$$

$$\begin{aligned} \Sigma_{PRED} &= \phi(\vec{x}^*)^T \Sigma \phi(\vec{x}^*) - \phi(\vec{x}^*)^T \Sigma \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I_n)^{-1} \Phi \Sigma \phi(\vec{x}^*) \end{aligned}$$

Bayesian Linear Regression can be kernelized!

$$\Phi = \begin{bmatrix} 1 & \phi(\vec{x}_1)^T \\ 1 & \phi(\vec{x}_2)^T \\ \vdots & \vdots \\ 1 & \phi(\vec{x}_n)^T \end{bmatrix}$$

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

$$\vec{y} = \Phi \vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p'+1}, \Sigma)$$

- then given a new test data point $\vec{x}^*$, the prediction is

$$y^* | \vec{y} = \phi(\vec{x}^*)^T \vec{\beta} | \vec{y} \sim N(\vec{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\vec{\mu}_{PRED} = \phi(\vec{x}^*)^T \Sigma \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I_n)^{-1} \vec{y},$$

$$\Sigma_{PRED}$$

$$= \phi(\vec{x}^*)^T \Sigma \phi(\vec{x}^*) - \phi(\vec{x}^*)^T \Sigma \Phi^T (\Phi \Sigma \Phi^T + \sigma^2 I_n)^{-1} \Phi \Sigma \phi(\vec{x}^*)$$

- Define the kernel function to be

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

Bayesian Linear Regression can be kernelized!

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

$$\vec{y} = \Phi \vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p'+1}, \Sigma)$$

- then given a new test data point $\vec{x}^*$, the prediction is

$$y^* | \vec{y} \sim N(\vec{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\begin{aligned} \vec{\mu}_{PRED} &= K(\vec{x}^*, A)(K(A, A) + \sigma^2 I_n)^{-1} \vec{y}, \\ \Sigma_{PRED} &= K(\vec{x}^*, \vec{x}^*) - K(\vec{x}^*, A)(K(A, A) + \sigma^2 I_n)^{-1} K(A, \vec{x}^*) \end{aligned}$$

- Define the kernel function to be

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

Wait, what happened to the weights?

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

$$\vec{y} = \Phi \vec{\beta} + \vec{\epsilon} \text{ where } \vec{\epsilon} \sim N(\vec{0}_n, \sigma^2 I_n) \text{ and } \vec{\beta} \sim N(\vec{0}_{p'+1}, \Sigma)$$

- then given a new test data point $\vec{x}^*$, the prediction is

$$y^* | \vec{y} \sim N(\vec{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$\begin{aligned} \vec{\mu}_{PRED} &= K(\vec{x}^*, A)(K(A, A) + \sigma^2 I_n)^{-1} \vec{y}, \\ \Sigma_{PRED} &= K(\vec{x}^*, \vec{x}^*) - K(\vec{x}^*, A)(K(A, A) + \sigma^2 I_n)^{-1} K(A, \vec{x}^*) \end{aligned}$$

- Define the kernel function to be

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

Some old
friends

Gaussian process =
Bayesian linear regression + Kernels

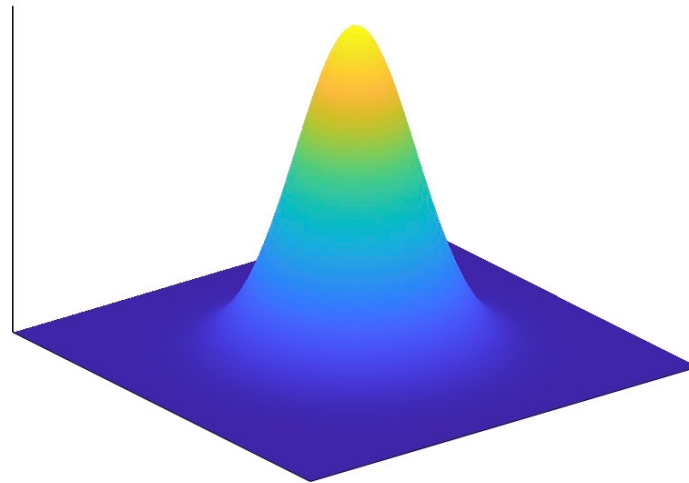
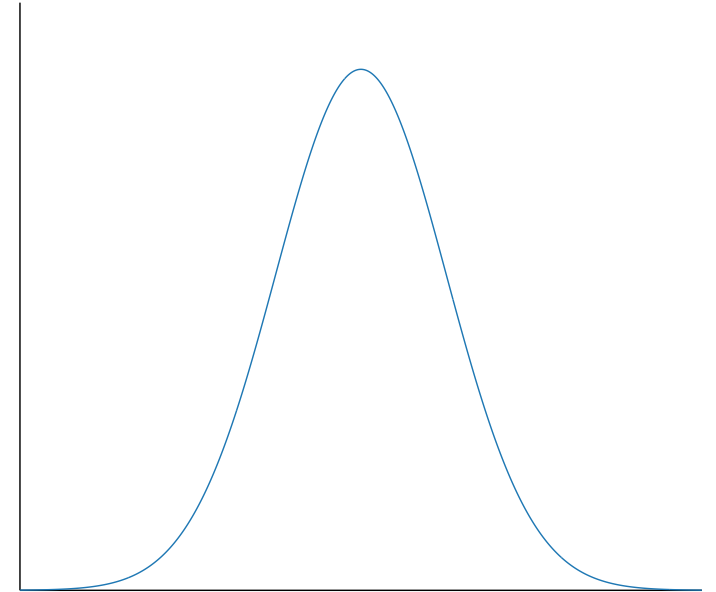
A new perspective

Gaussian process =
The extension of a Gaussian
distribution to functions

Gaussians

- (Univariate) Gaussians:

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



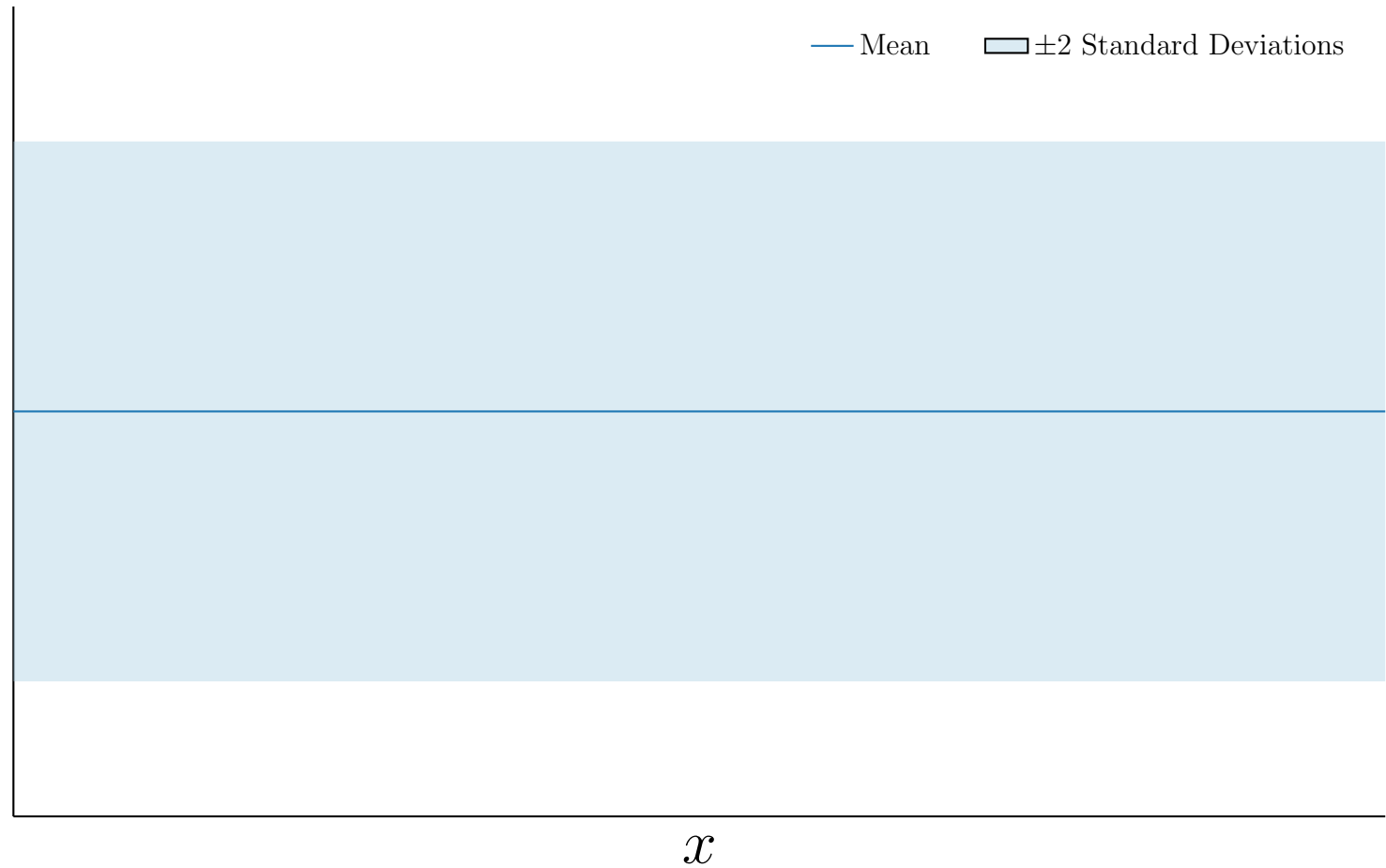
- Multivariate Gaussians:

$$\vec{x} = [x_1, \dots, x_p]^T$$

$$\sim \mathcal{N}(\vec{x}; \vec{\mu} = \vec{0}_p, \Sigma = I_p)$$

Gaussian Process (GP)

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

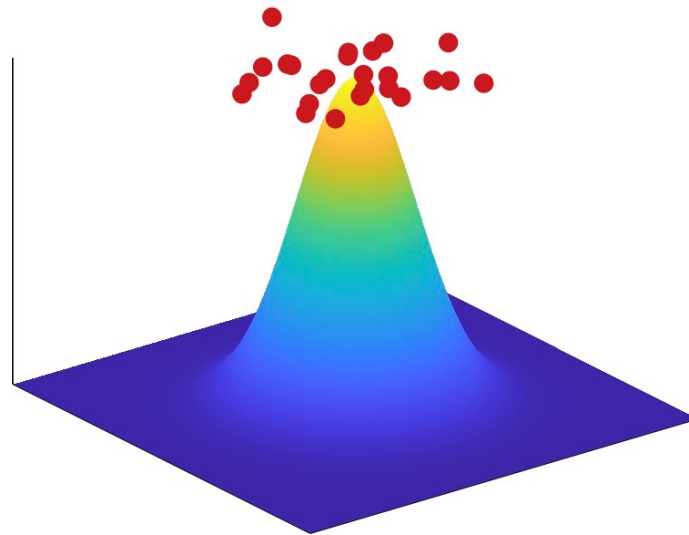
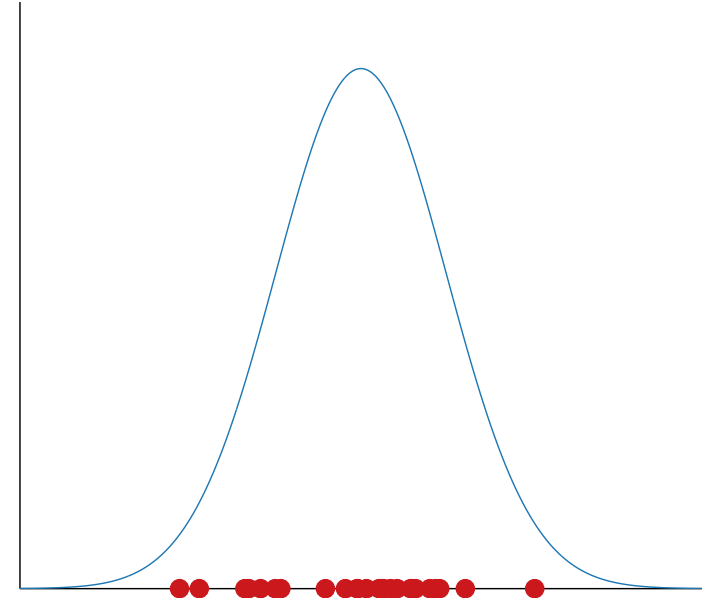


$$f \sim \mathcal{GP}(\mu, \Sigma) \rightarrow f(x) \sim \mathcal{N}(\mu(x), \Sigma(x, x))$$

Gaussians

- (Univariate) Gaussians:

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



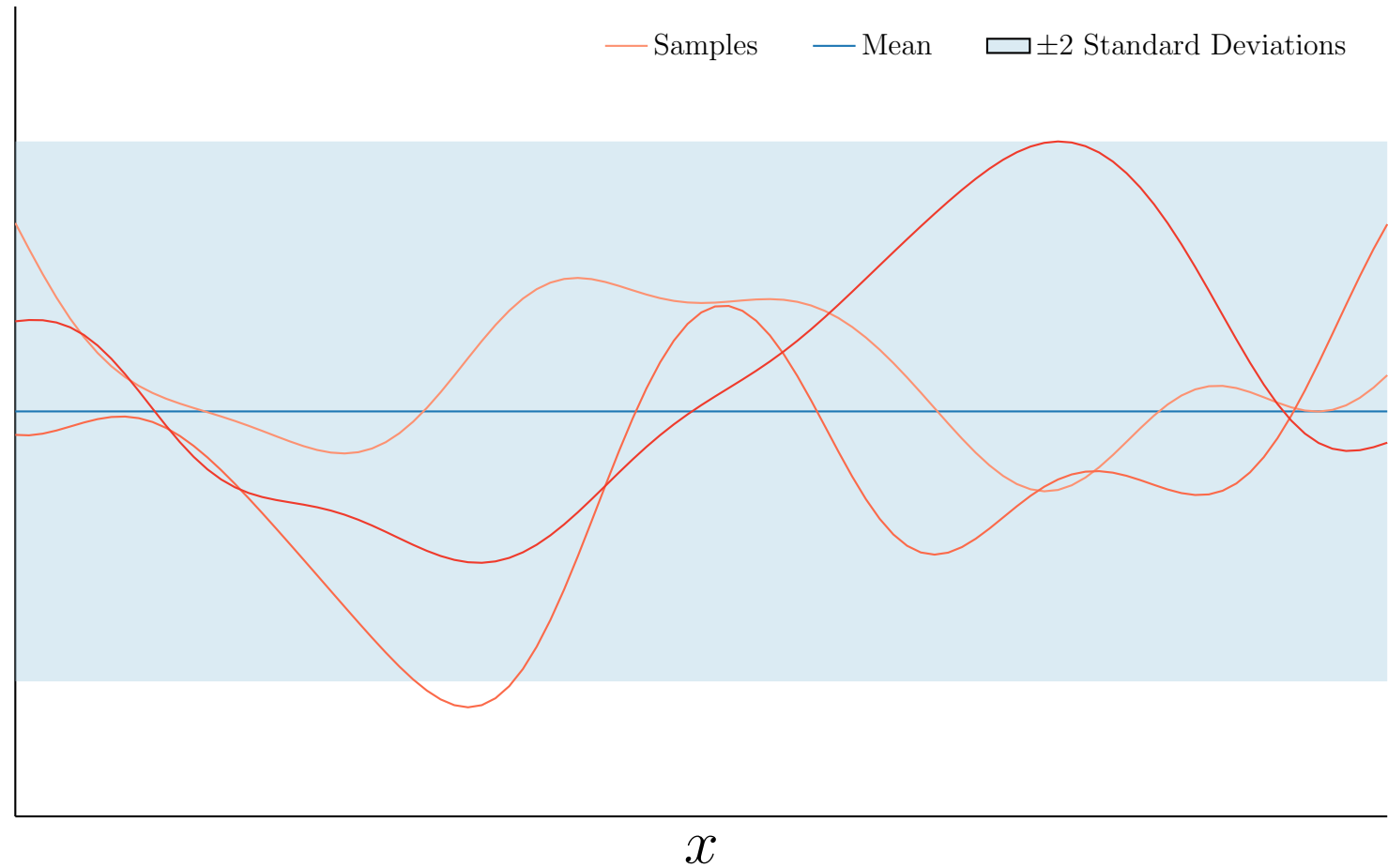
- Multivariate Gaussians:

$$\vec{x} = [x_1, \dots, x_p]^T$$

$$\sim \mathcal{N}(\vec{x}; \vec{\mu} = \vec{0}_p, \Sigma = I_p)$$

Gaussian Process (GP)

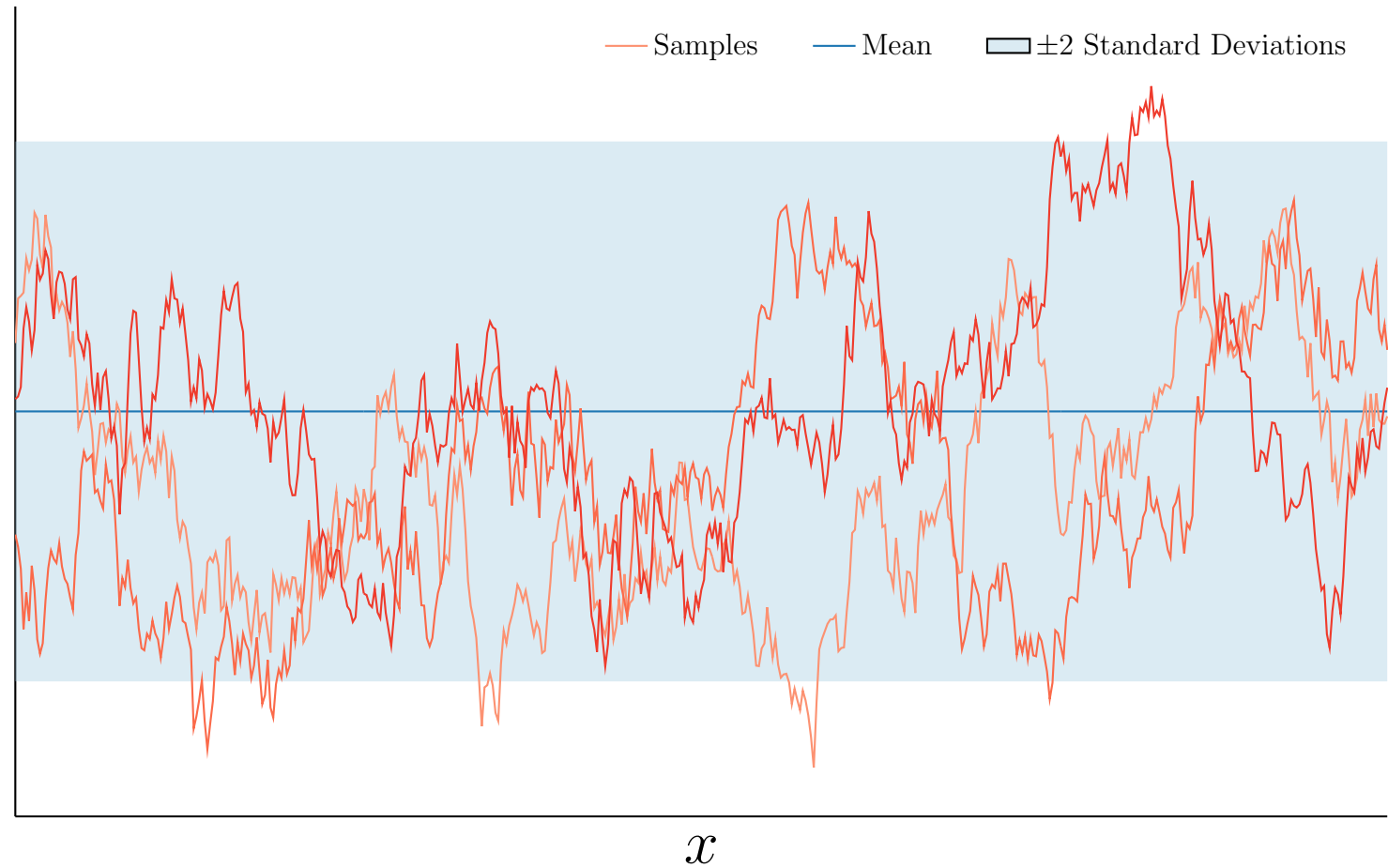
$$f: \mathbb{R}^p \mapsto \mathbb{R} \sim \mathcal{GP}(f; \mu(x) = 0, \Sigma(x, x') = \exp(-(x - x')^2))$$



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

Gaussian Process (GP)

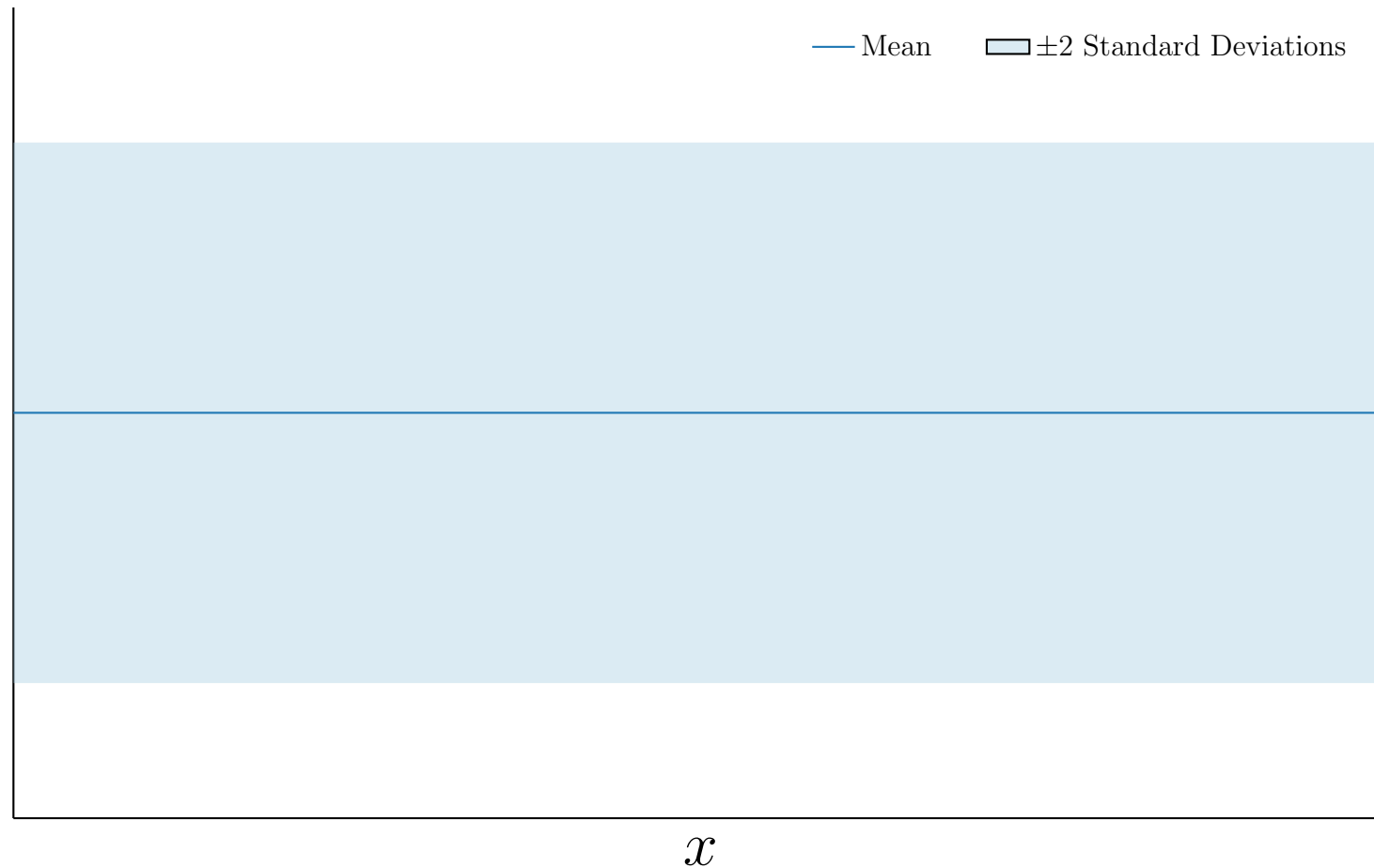
$$f: \mathbb{R}^p \mapsto \mathbb{R} \sim \mathcal{GP}(f; \mu(x) = 0, \Sigma(x, x') = \exp(-|x - x'|))$$



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

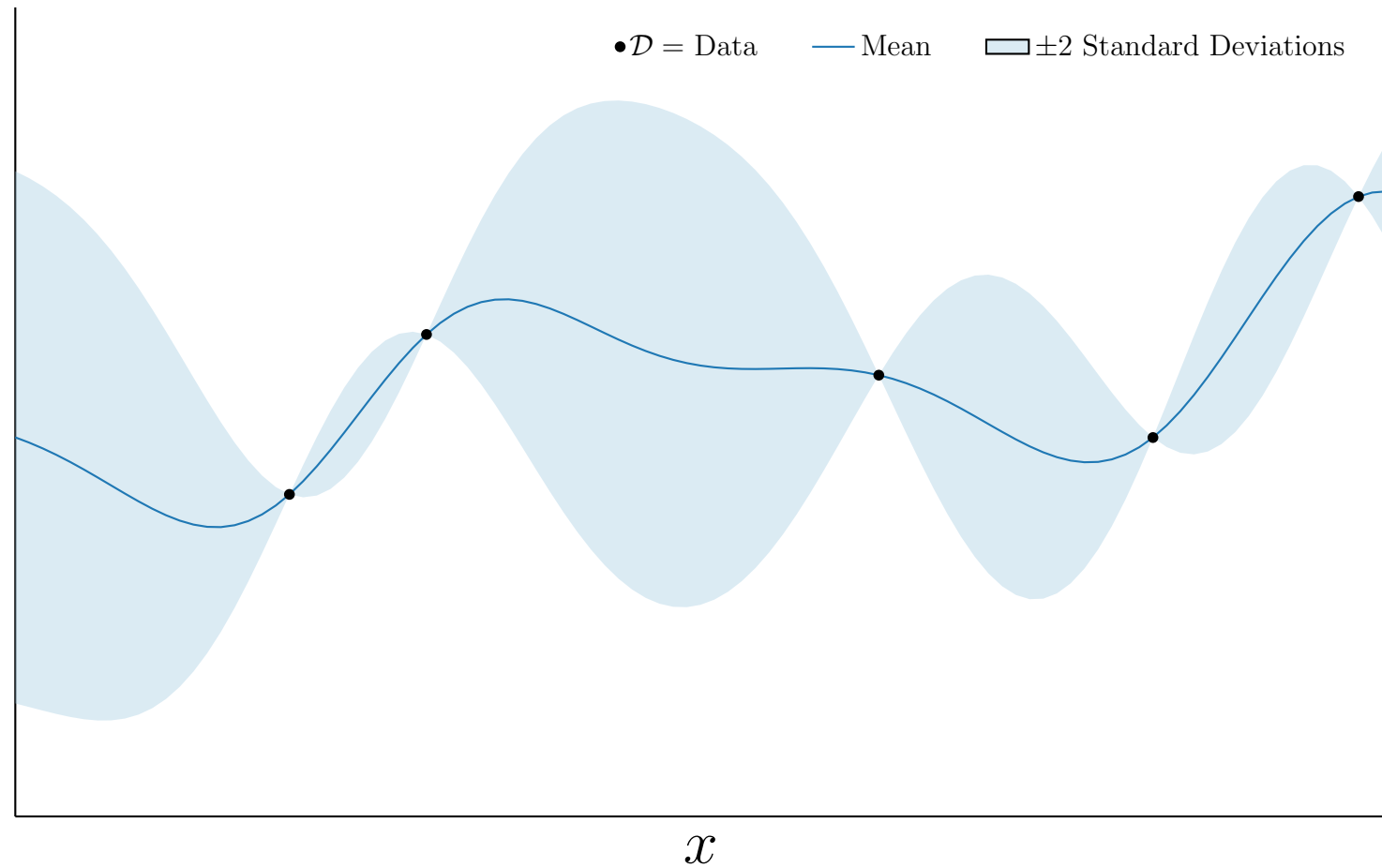
GP Prior

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



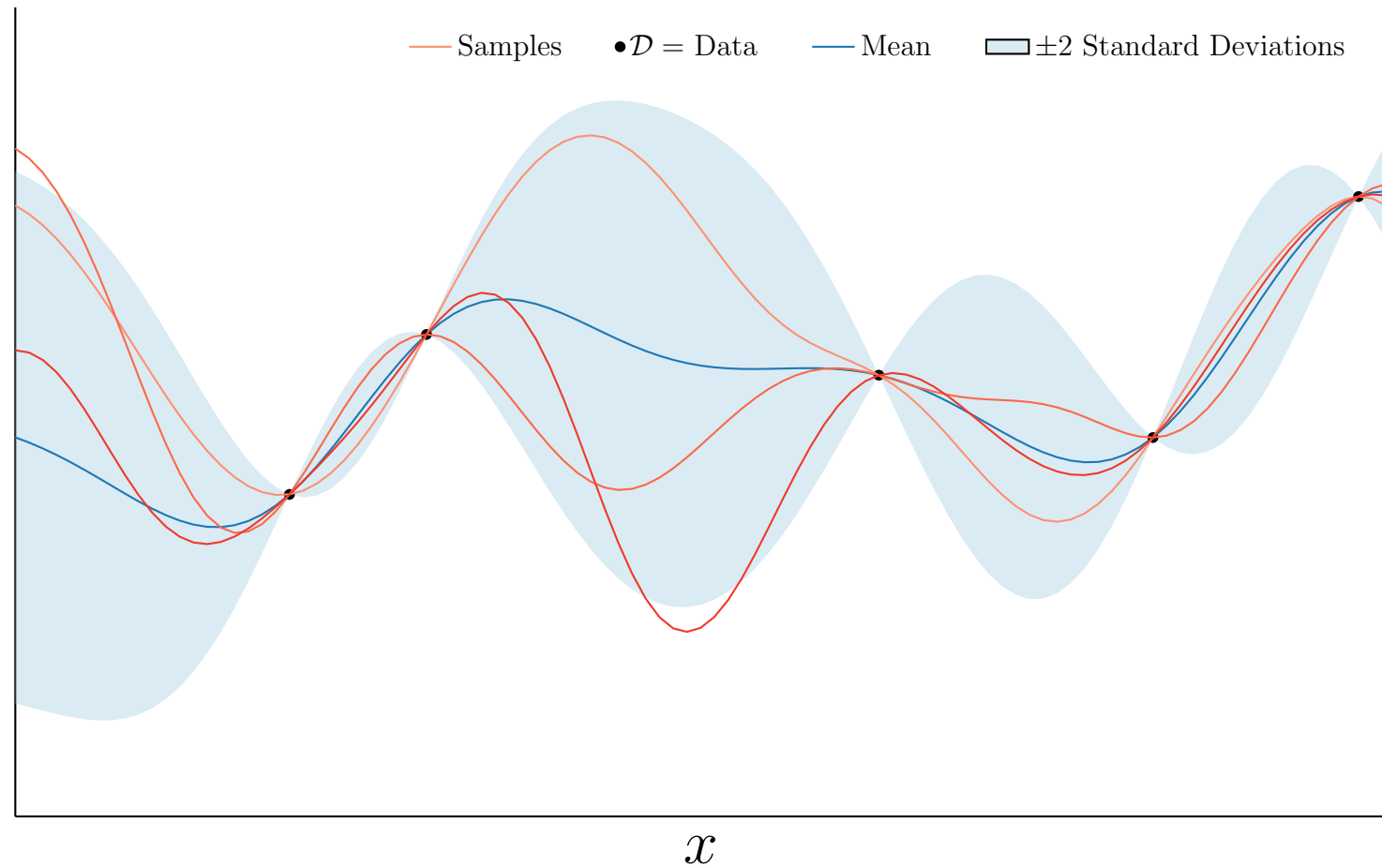
GP Posterior

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

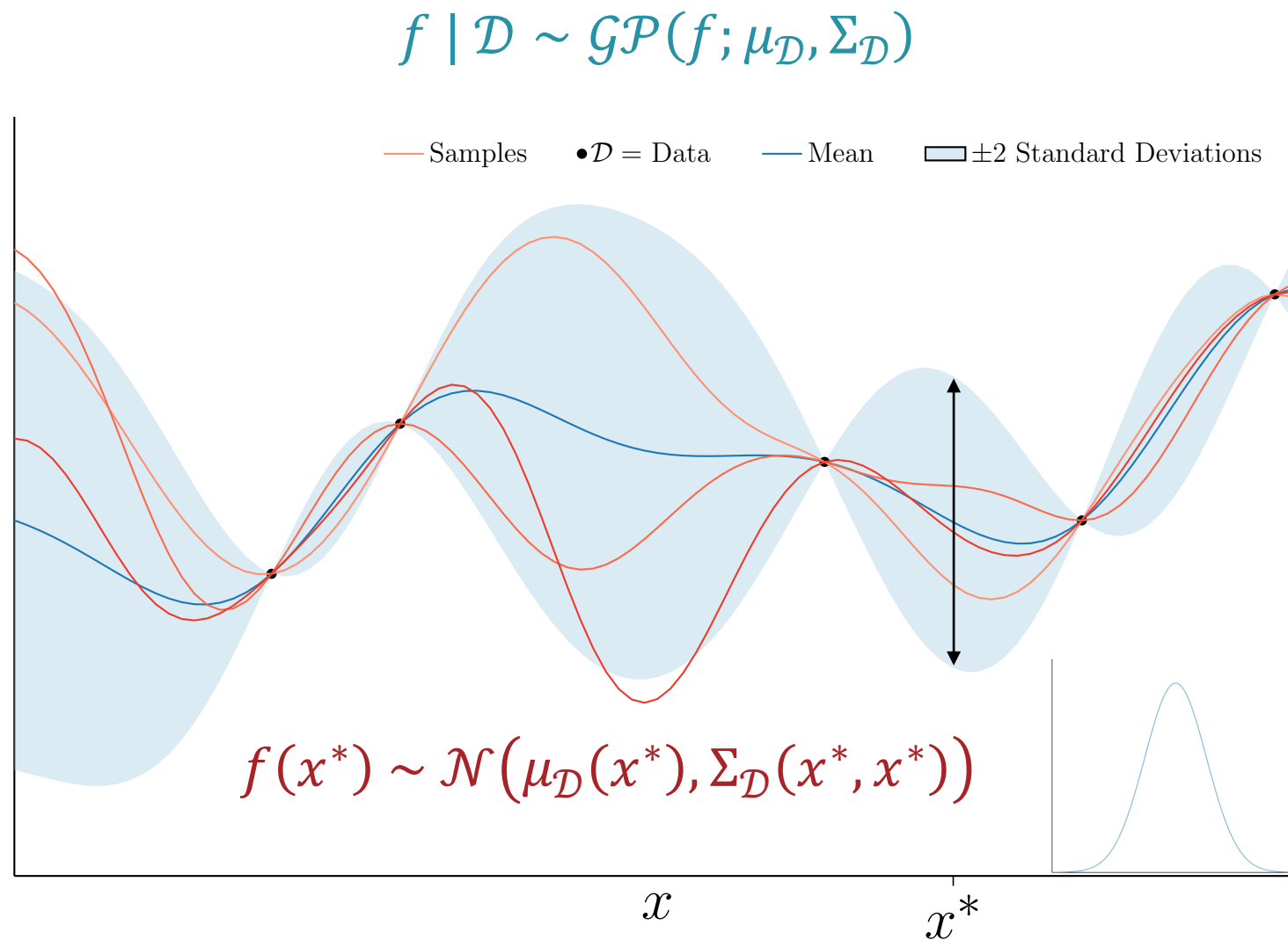


GP Posterior

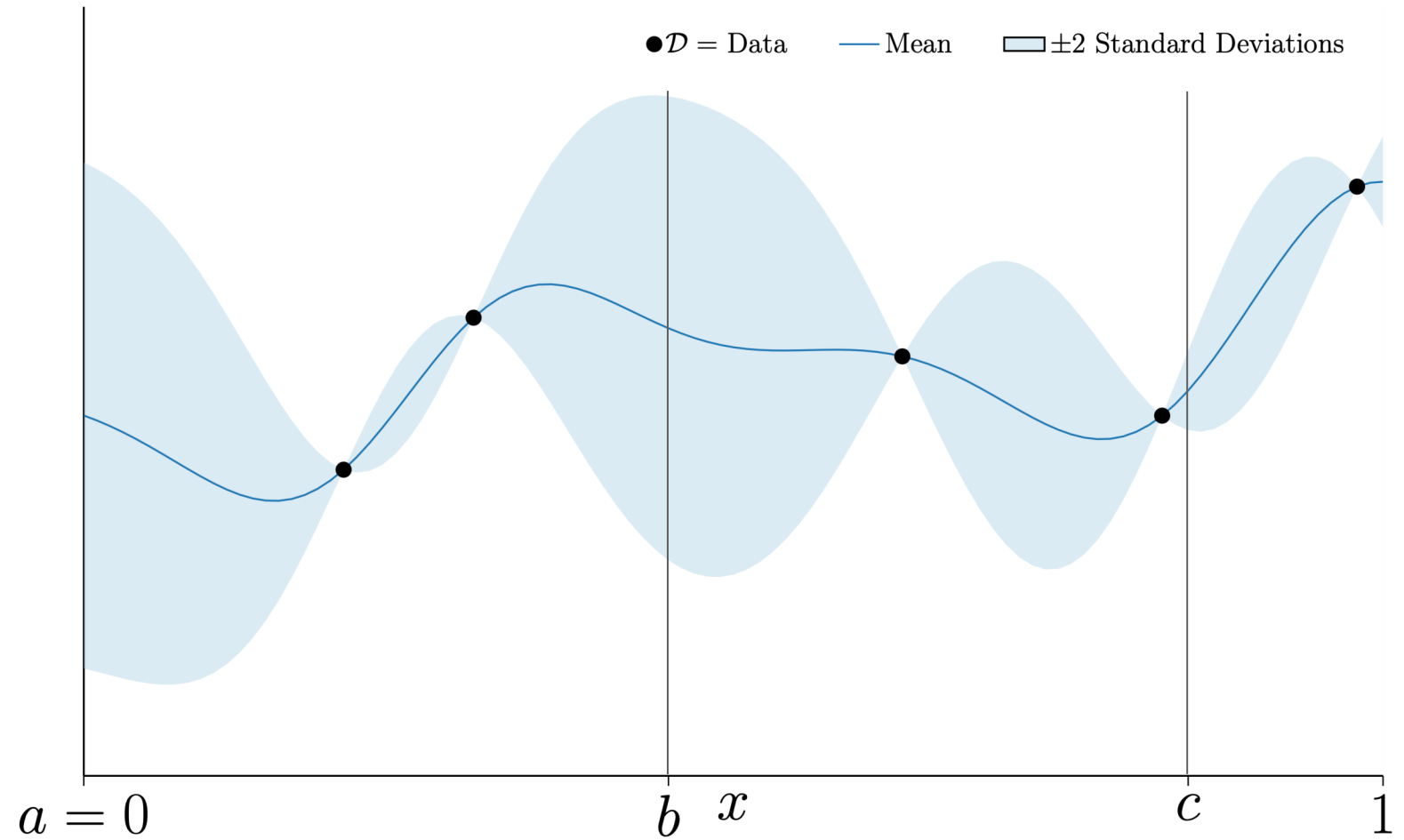
$$f \mid \mathcal{D} \sim \mathcal{GP}(f; \mu_{\mathcal{D}}, \Sigma_{\mathcal{D}})$$



GP Posterior

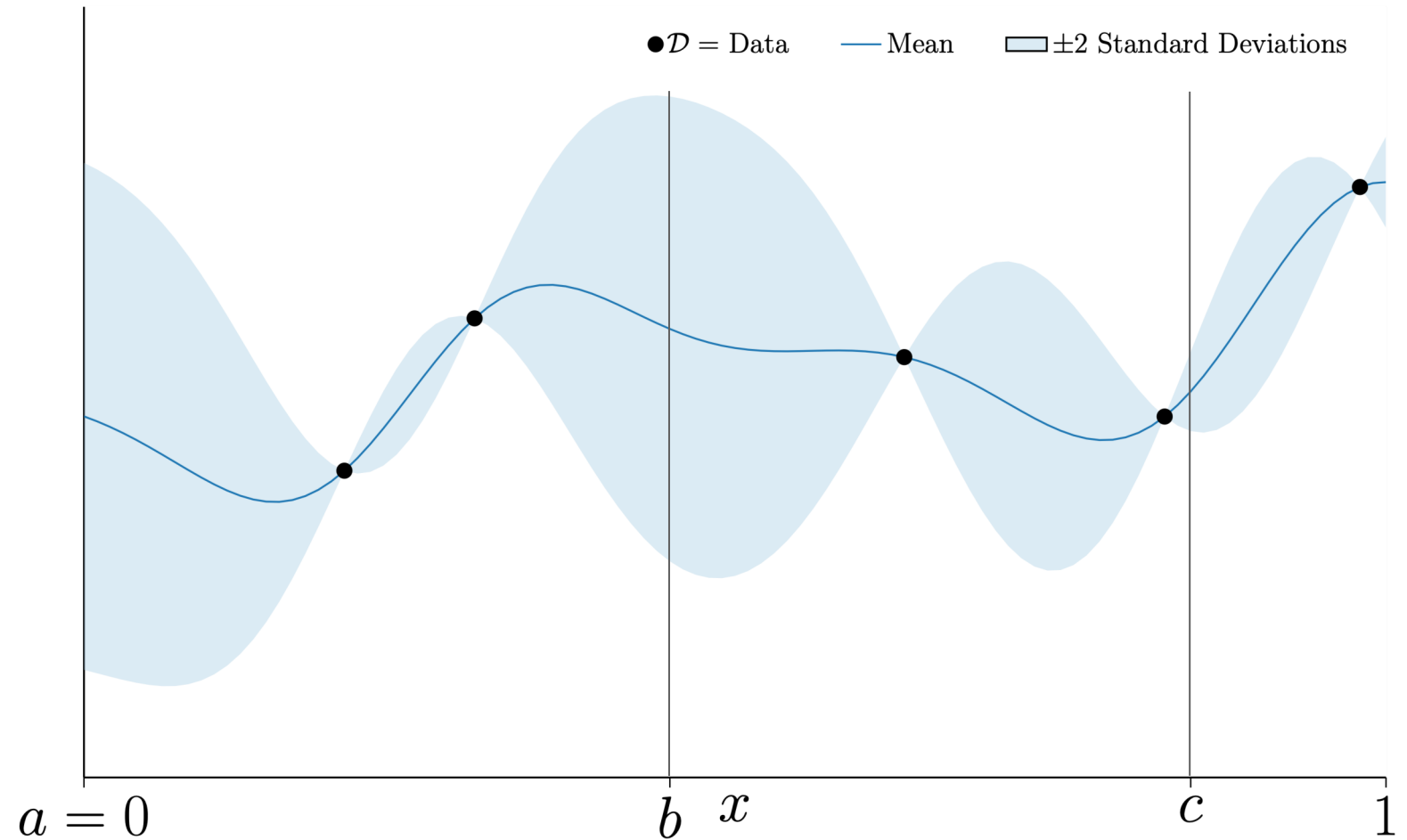


Active Learning



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

Which point would you add and why?

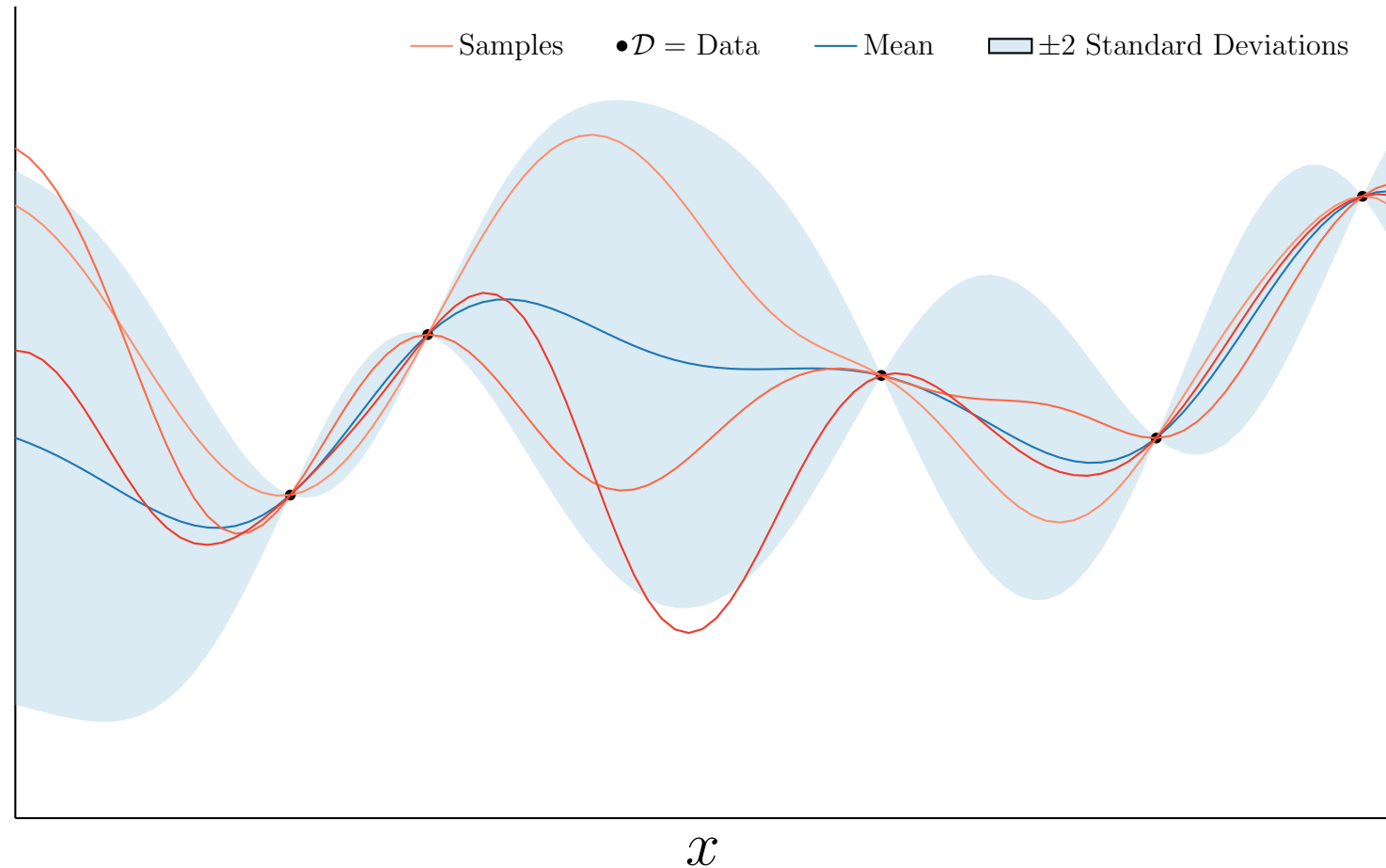


Are GPs:

1. parametric or nonparametric

2. generative or discriminative

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

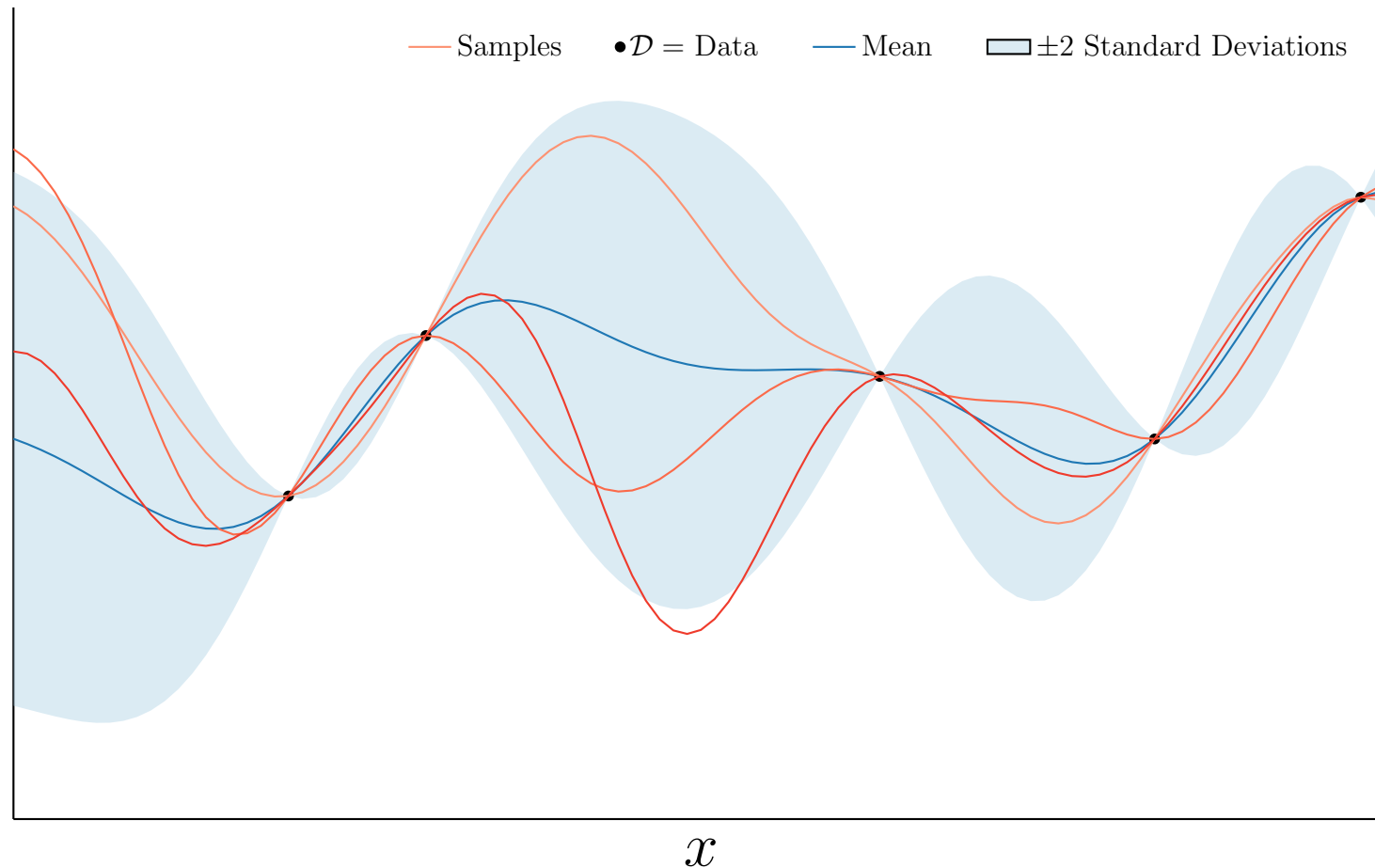


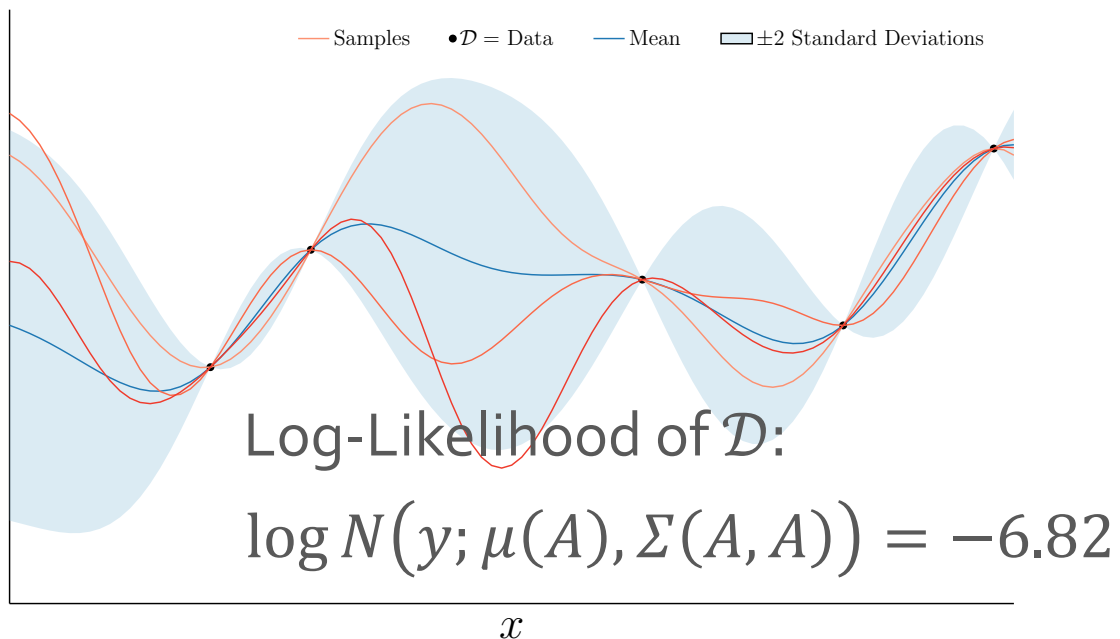
GPs are

1. parametric or nonparametric

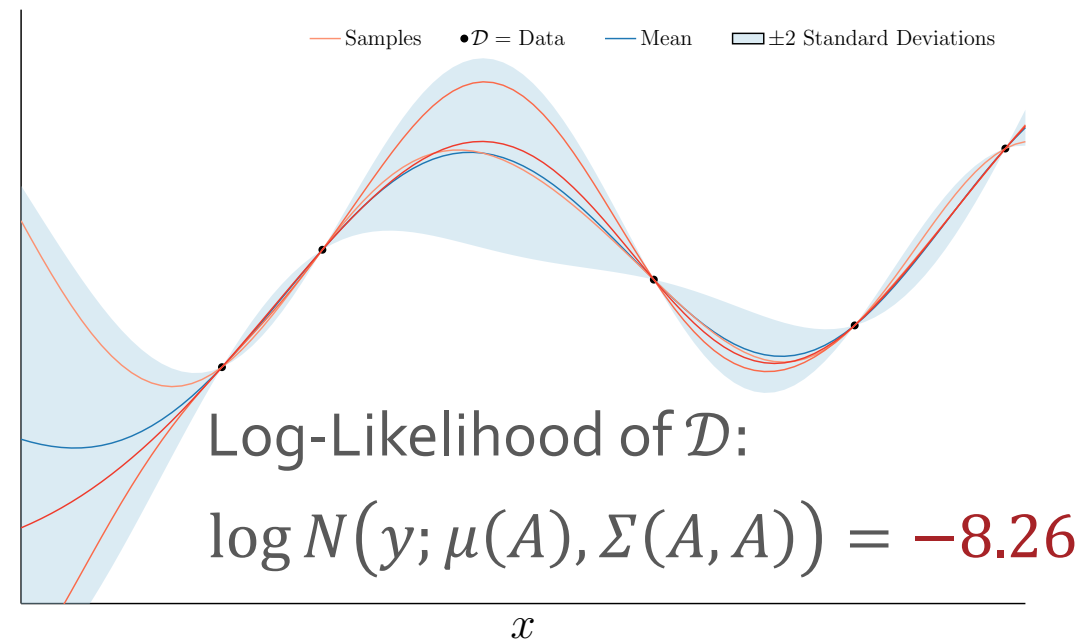
2. generative or discriminative

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





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



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

Kernel Hyperparameters