10-701: Introduction to Machine Learning Lecture 26 – Gaussian Processes

Henry Chai

4/22/24

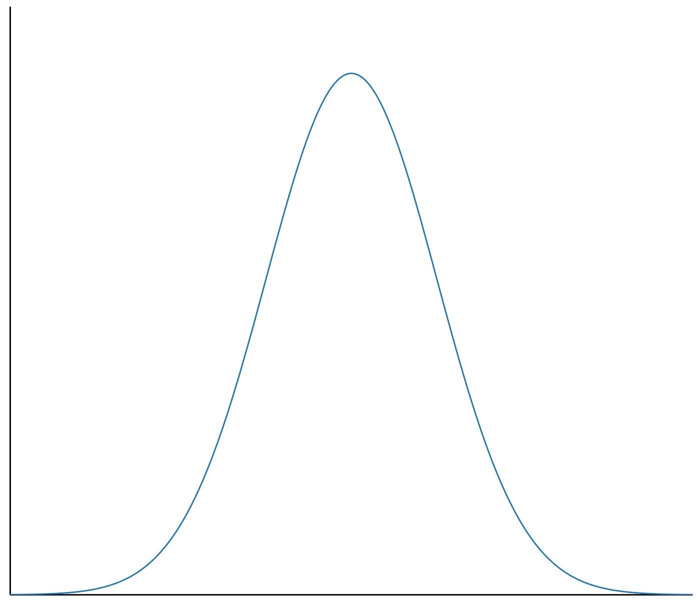
Front Matter

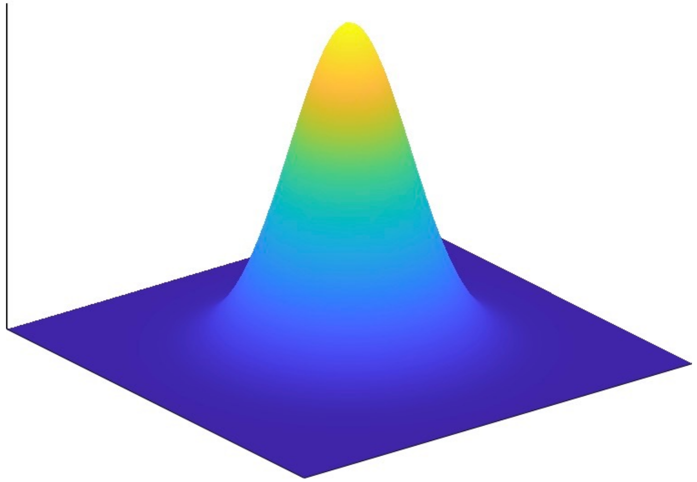
- Announcements:
  - Exam 2 on 5/6 **from 1 PM – 3 PM in TEP 1403**
    - You are allowed to bring one letter-/A4-size sheet of notes; you can put *whatever* you want on *both sides*
    - **Pre-midterm material may be referenced but will not be the primary focus of any question**
  - Project Final Reports due on 4/26 (Friday) at 11:59 PM
    - **No late days can be used on project deliverables**
- Recommended Readings:
  - Murphy, Chapters 15.1-15.2

Gaussians

(Univariate) Gaussians:

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





- Multivariate Gaussians:

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

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

Some fun facts about Gaussians

- Closure under linear transformations:

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

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

- Closure under addition

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

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

- Closure under conditioning:

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

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

Some old friends

Gaussian process =

Bayesian linear regression + Kernels

Some old friends

Gaussian process =

Bayesian linear regression + Kernels

Recall: MAP for Linear Regression

- If we assume a linear model with additive Gaussian noise

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

and independent identical Gaussian priors on the weights…

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

- … then, the MAP of $\boldsymbol{w}$ is the ridge regression solution!

$$\boldsymbol{w}_{MAP} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \, (X\boldsymbol{w} - \boldsymbol{y})^T (X\boldsymbol{w} - \boldsymbol{y}) + \lambda \boldsymbol{w}^T \boldsymbol{w}$$
$$= (X^T X + \lambda I_{D+1})^{-1} X^T \boldsymbol{y}$$

Bayesian Linear Regression

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

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

then,

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

Bayesian Linear Regression

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

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

then,

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

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

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

Bayesian Linear Regression

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

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

then,

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

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

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

Bayesian Linear Regression

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

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

then,

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

where

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

Bayesian Linear Regression

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

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

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

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

where

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

Bayesian Linear Regression

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

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

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

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

where

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

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

Some old friends

Gaussian process =

Bayesian linear regression + Kernels

Some old friends

Gaussian process =

Bayesian linear regression + Kernels

Bayesian Linear Regression...

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

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

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

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

where

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

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

Bayesian Linear Regression can be kernelized!

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

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

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

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

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

where

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

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

Bayesian Linear Regression can be kernelized!

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

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

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

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

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

where

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

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

- Define the kernel function to be

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

Bayesian Linear Regression can be kernelized!

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

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

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

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

where

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

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

- Define the kernel function to be

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

Wait, what happened to the weights?

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

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

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

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

where

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

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

- Define the kernel function to be

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

Some old friends

Gaussian process =

Bayesian linear regression + Kernels

A new perspective

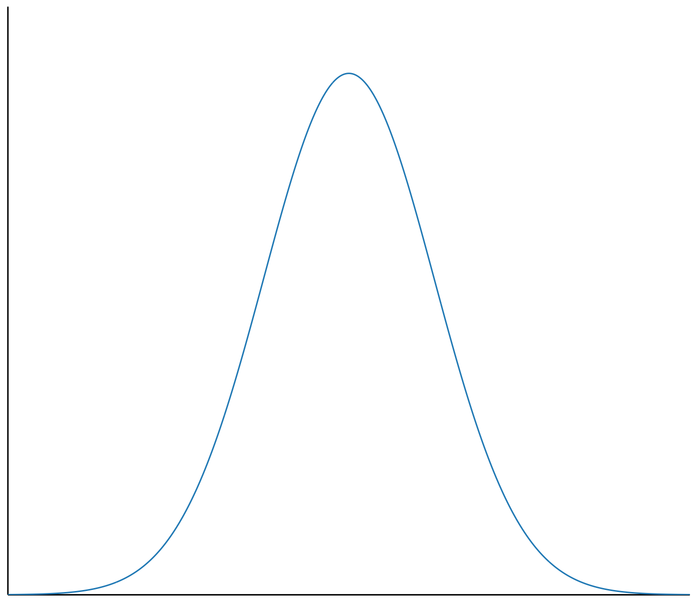
Gaussian process =

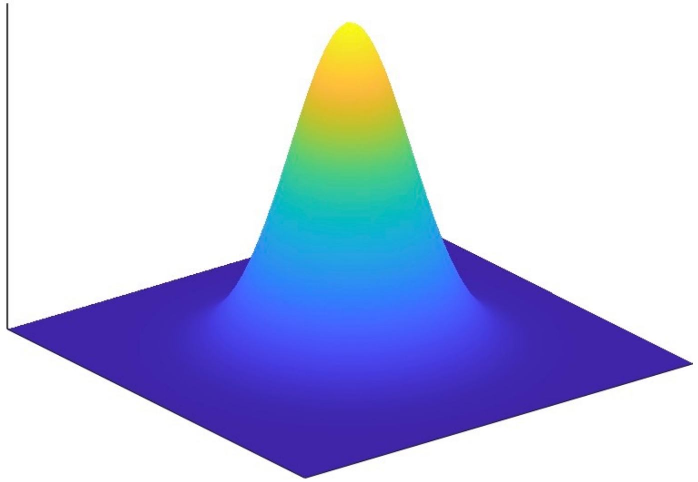
The extension of a Gaussian distribution to functions

Gaussians

- (Univariate) Gaussians:

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





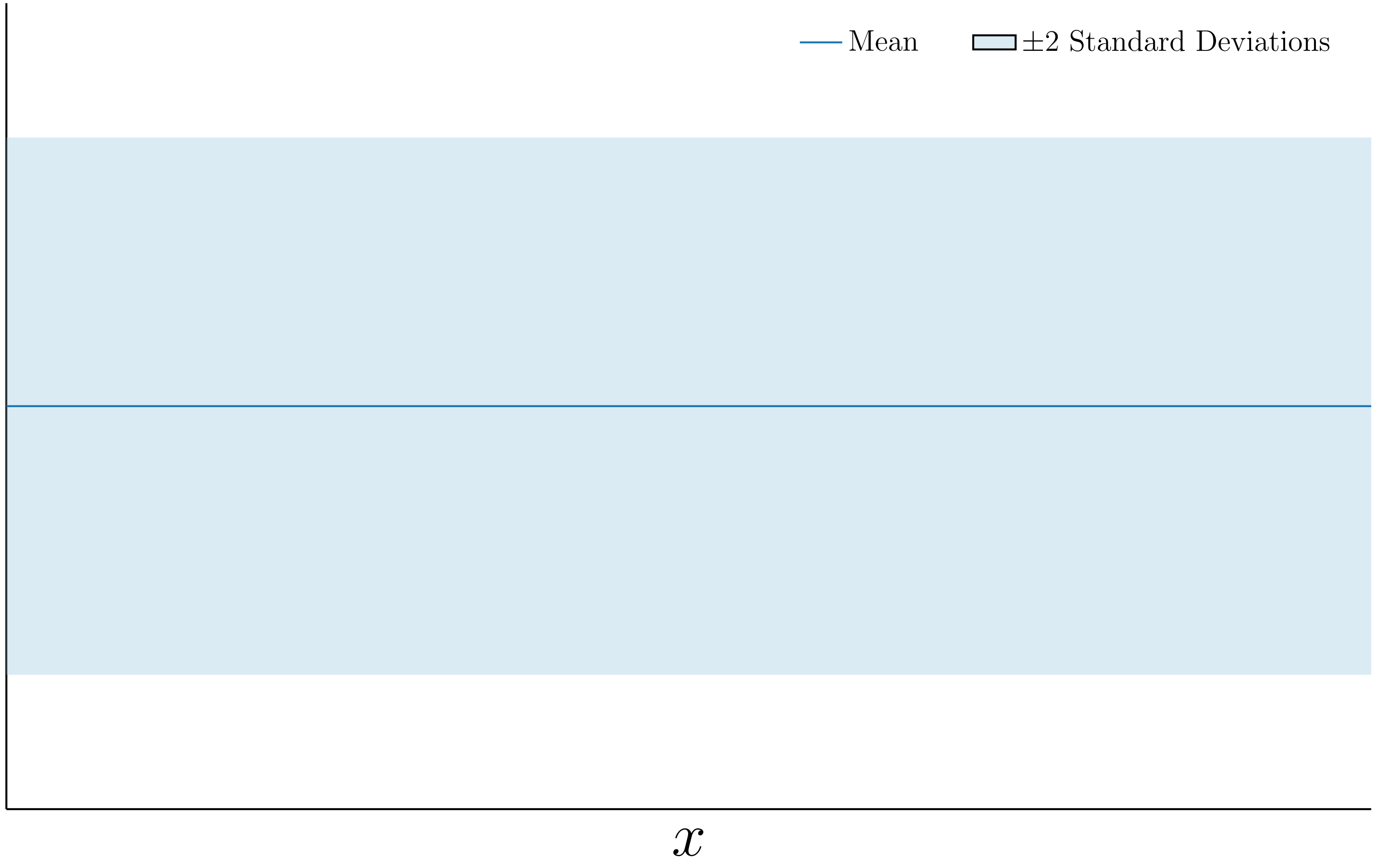
- Multivariate Gaussians:

  $\boldsymbol{x} = [x_1, \ldots, x_D]^T$

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

Gaussian Process (GP)

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

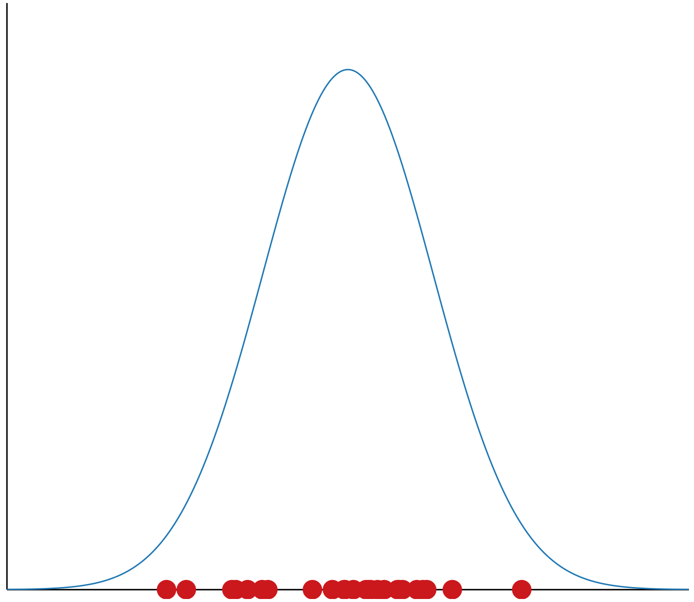


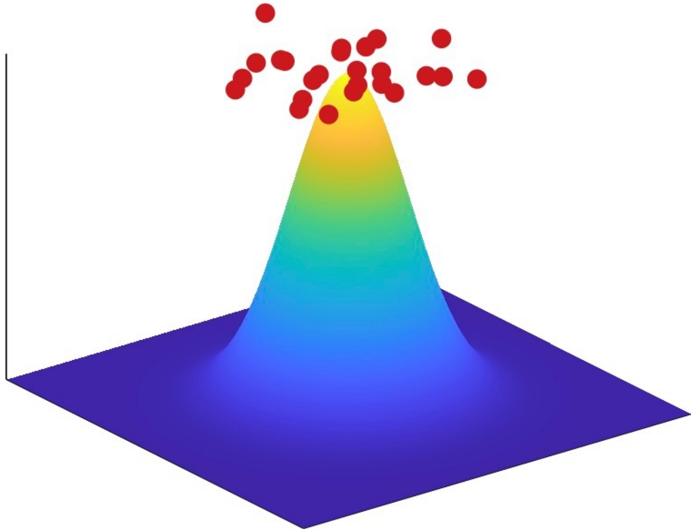
$f \sim \mathcal{GP}(\mu, \Sigma) \rightarrow f(x) \sim \mathcal{N}\big(\mu(x), \Sigma(x, x)\big)$

Gaussians

- (Univariate) Gaussians:

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



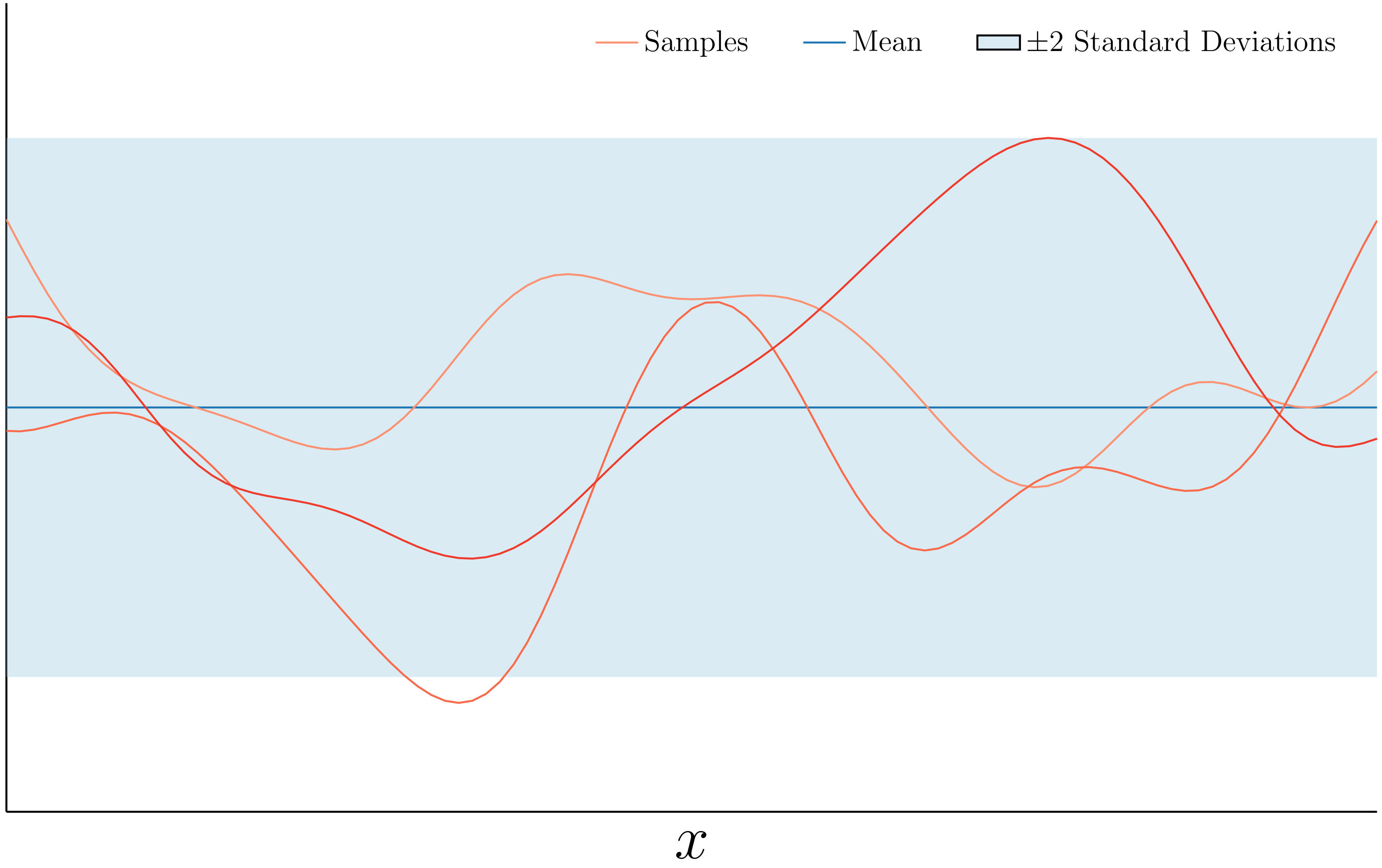


- Multivariate Gaussians:

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

Gaussian Process (GP)

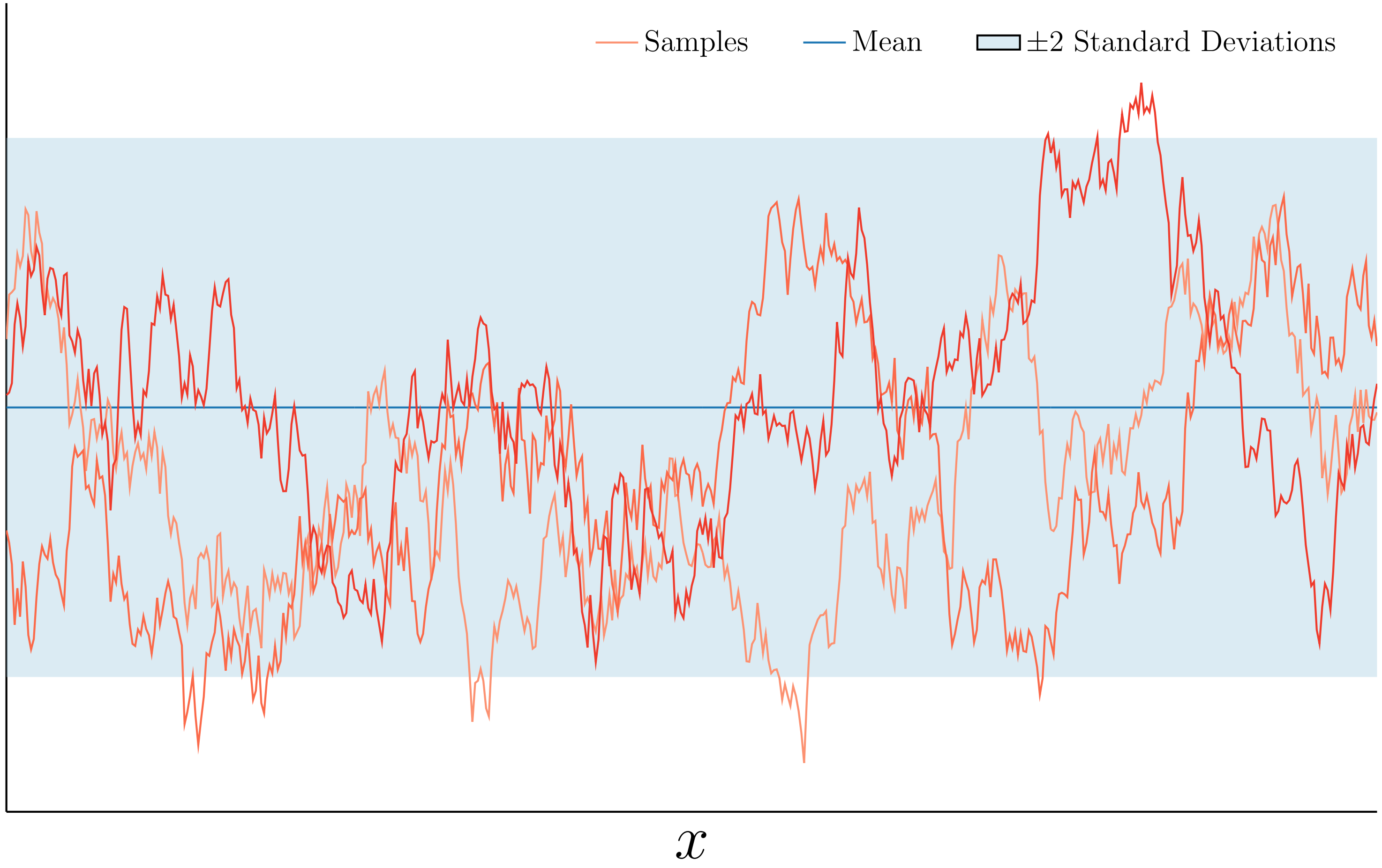
$$f: \mathbb{R}^p \mapsto \mathbb{R} \sim \mathcal{GP}(f; \mu(x) = 0, \Sigma(x, x') = \exp(-(x - x')^2))$$



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

Gaussian Process (GP)

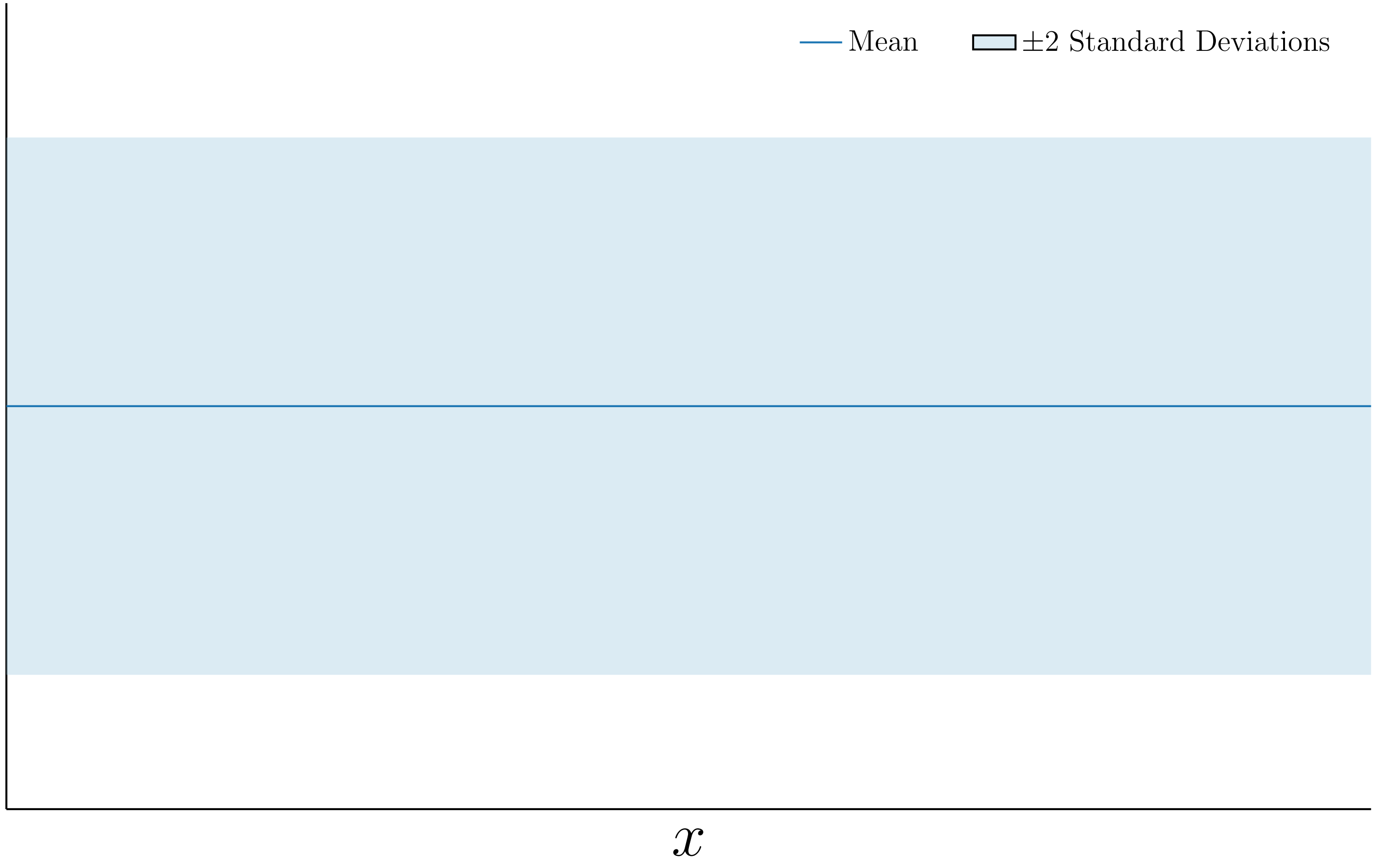
$$f: \mathbb{R}^p \mapsto \mathbb{R} \sim \mathcal{GP}(f; \mu(x) = 0, \Sigma(x, x') = \exp(-|x - x'|))$$



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

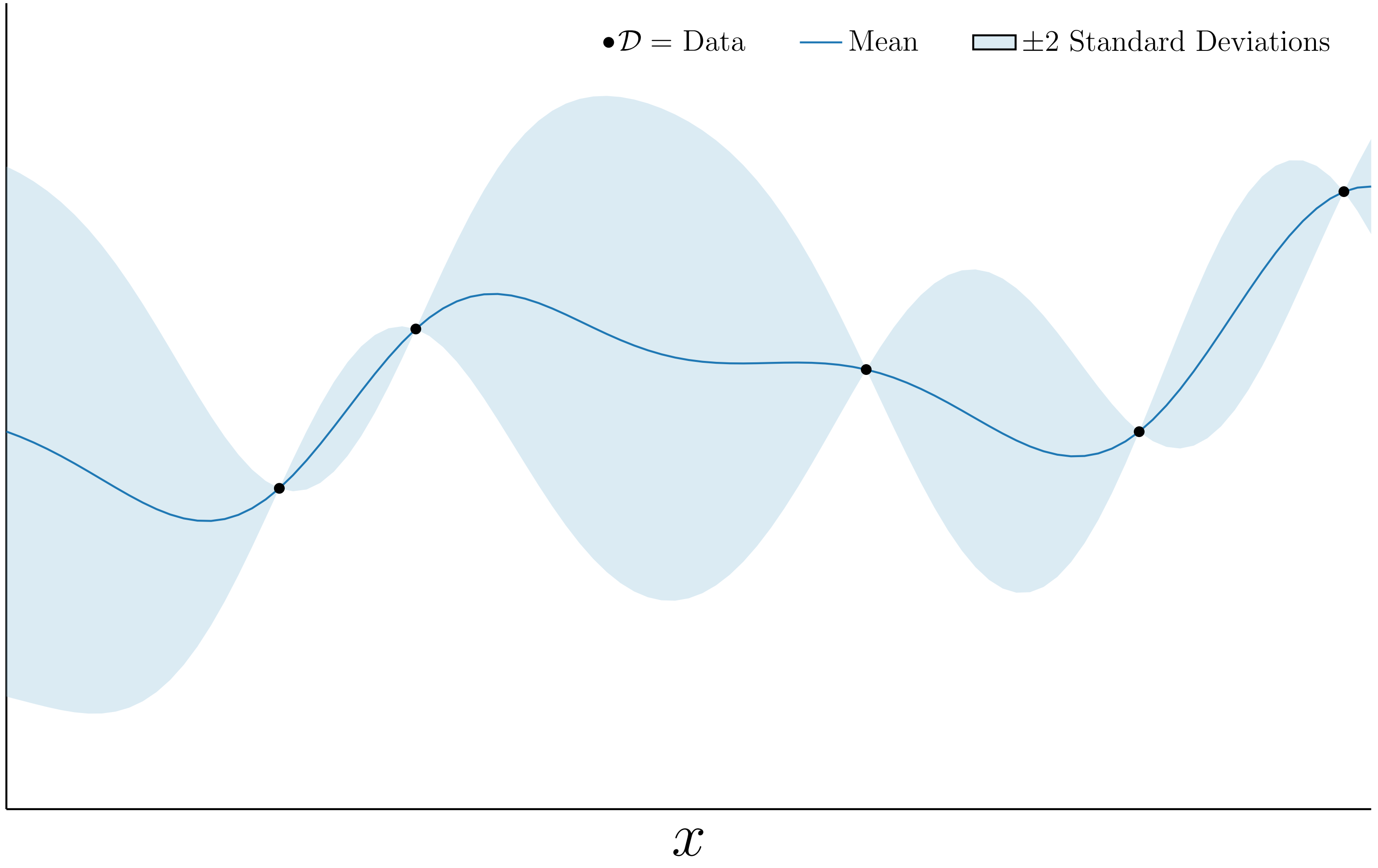
GP Prior

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



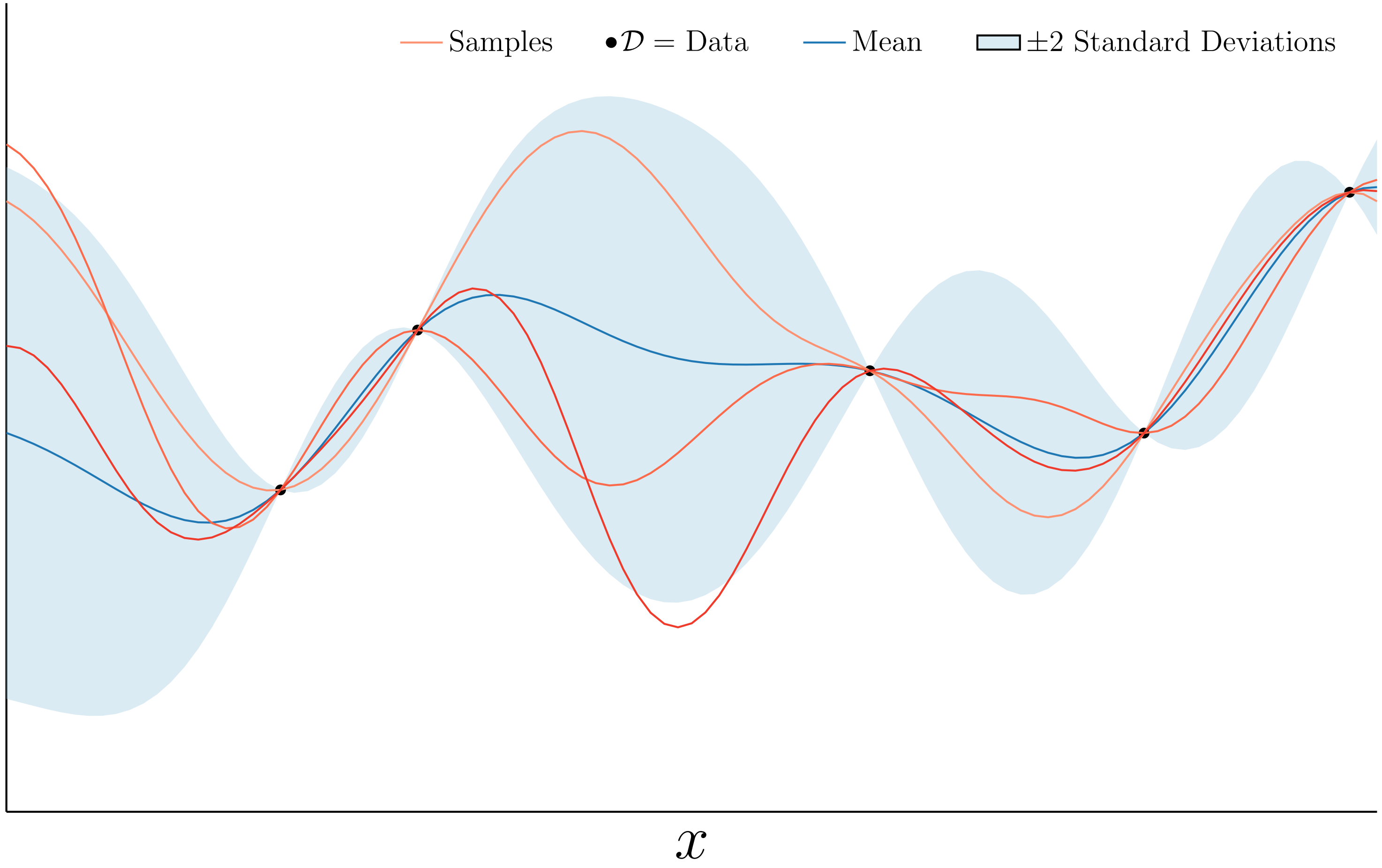
GP Posterior

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



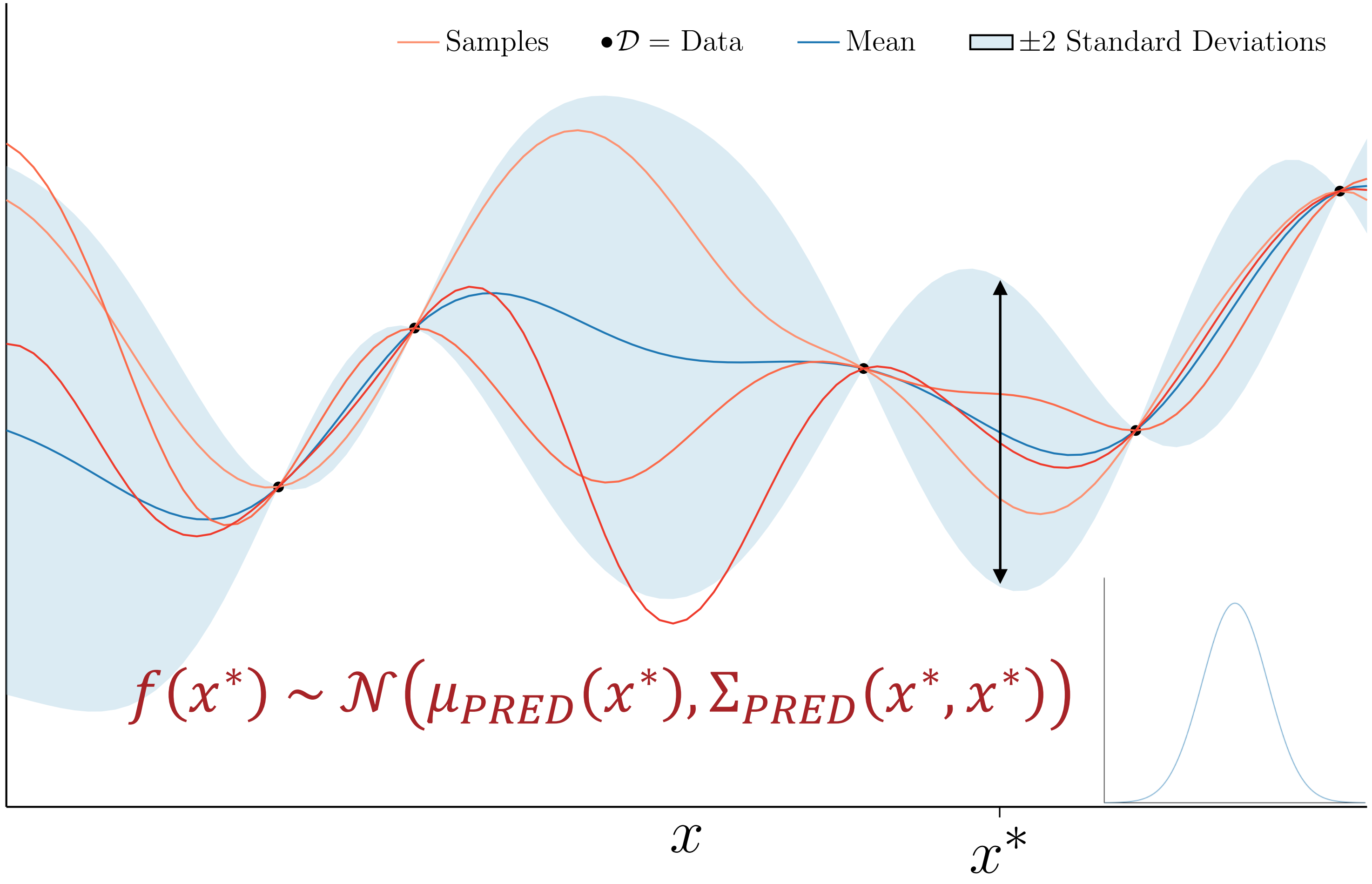
GP Posterior

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



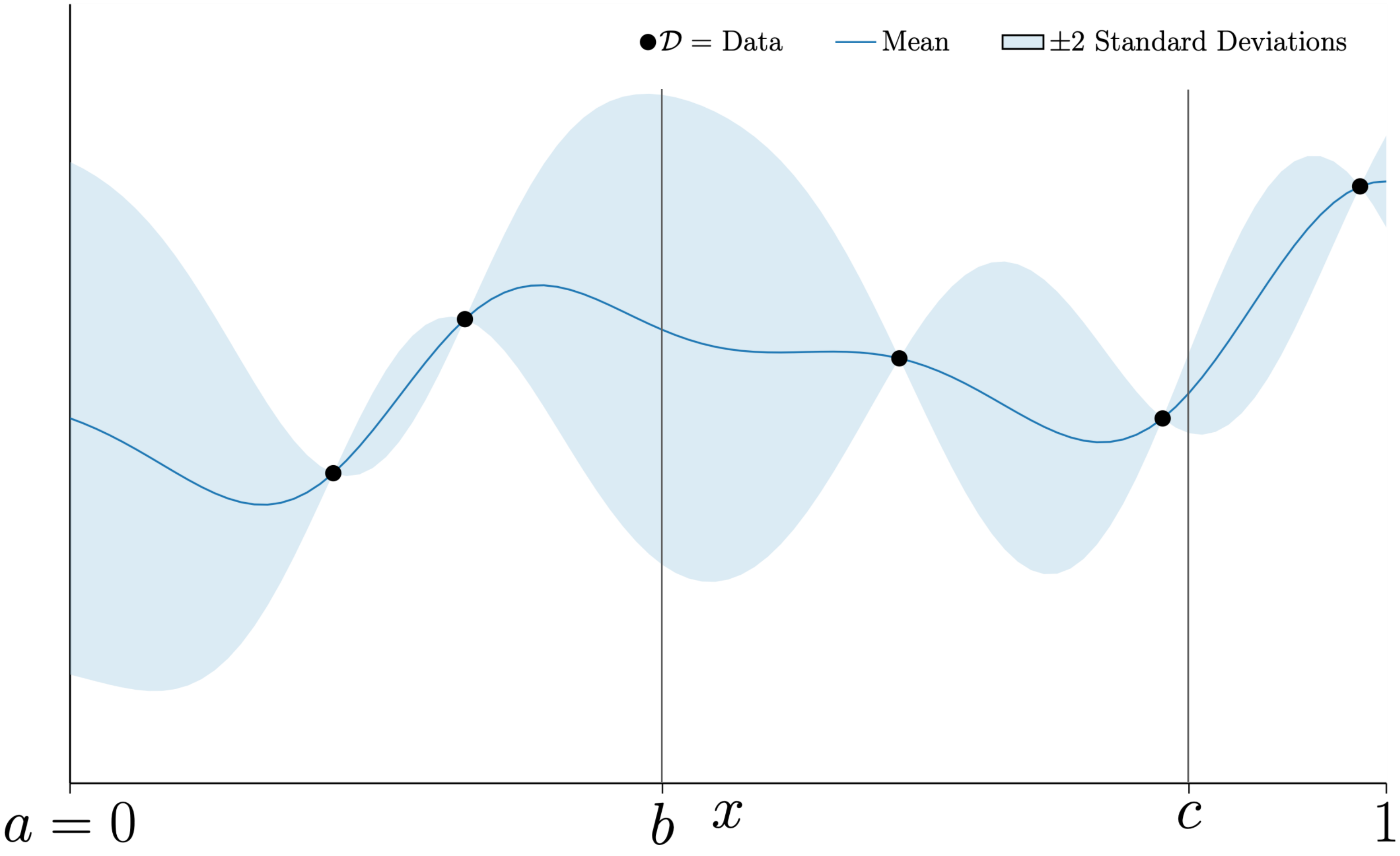
GP Posterior

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



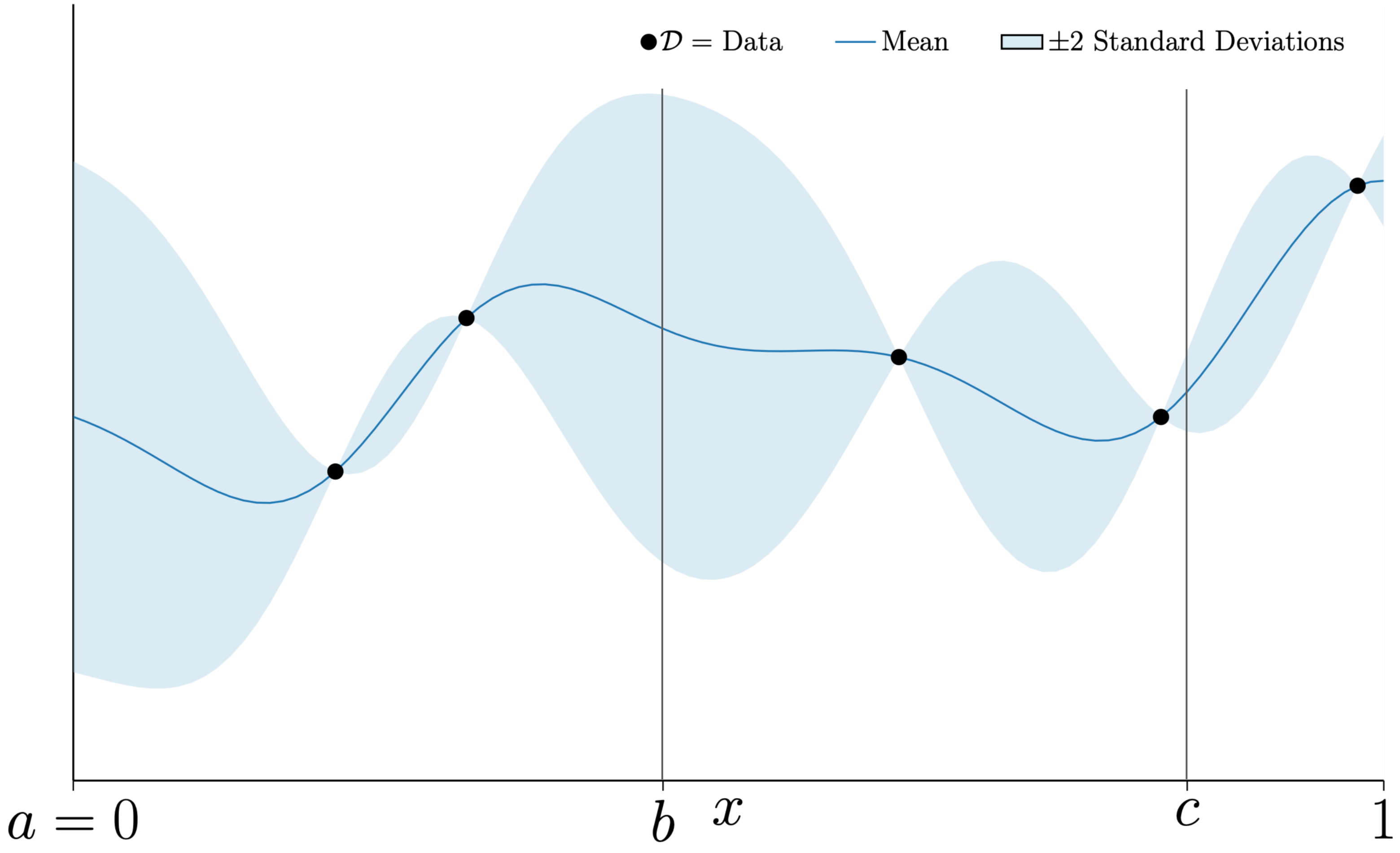
$$f(x^*) \sim \mathcal{N}\left(\mu_{PRED}(x^*), \Sigma_{PRED}(x^*, x^*)\right)$$

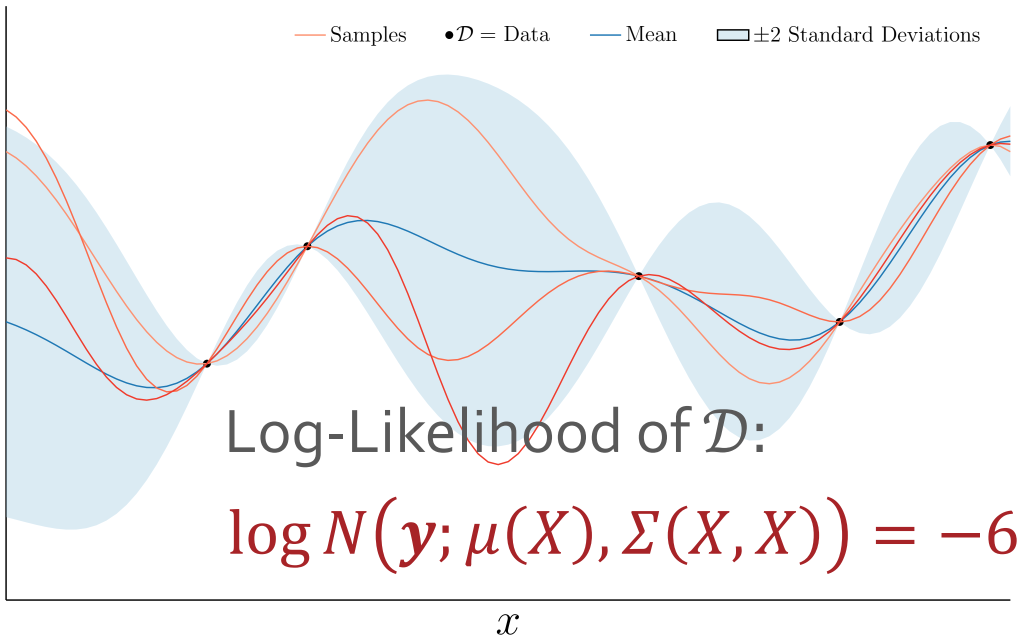
Active Learning



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

Which value of x would you add and why?

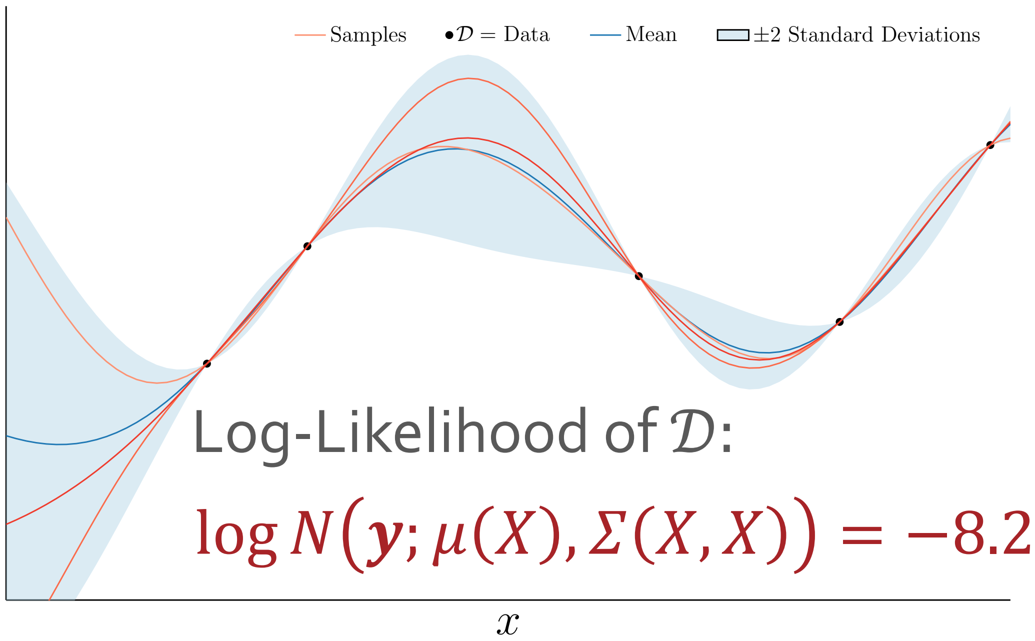




Log-Likelihood of $\mathcal{D}$:

$$\log N\big(\boldsymbol{y}; \mu(X), \Sigma(X, X)\big) = -6.82$$

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



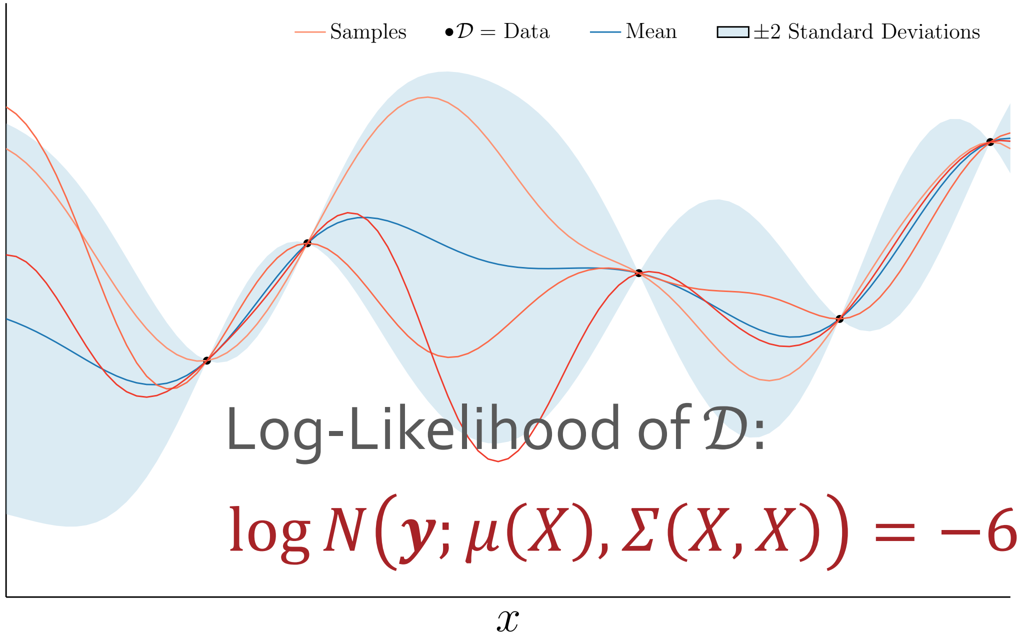
Log-Likelihood of $\mathcal{D}$:

$$\log N\big(\boldsymbol{y}; \mu(X), \Sigma(X, X)\big) = -8.26$$

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

Kernel Hyperparameters

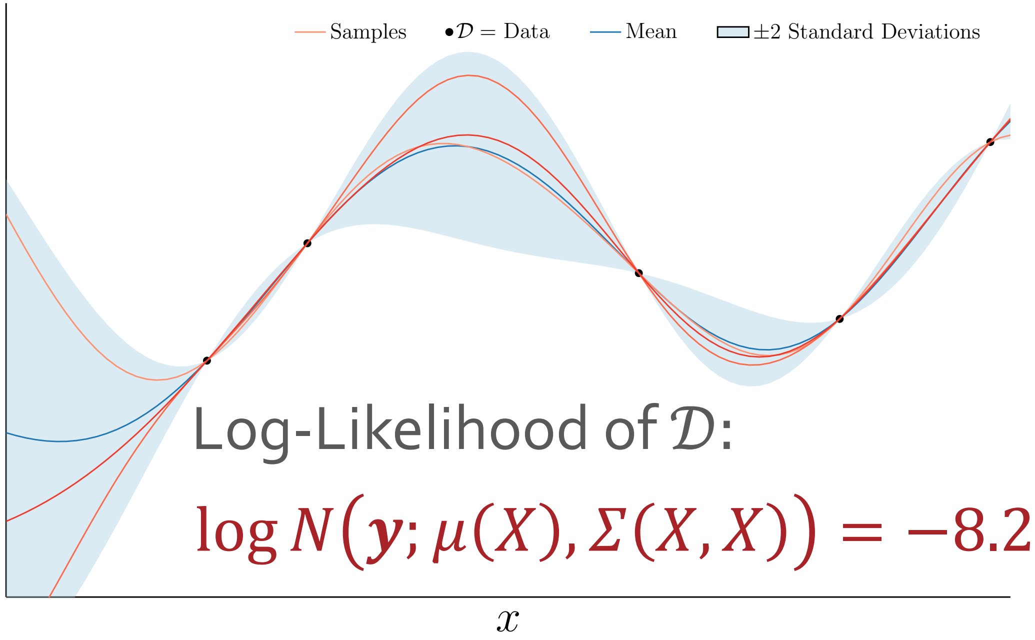
- Can be set via MLE
- As long as μ and Σ are differentiable, the log-likelihood is differentiable with respect to the kernel hyperparameters



Log-Likelihood of $\mathcal{D}$:

$\log N\left(\boldsymbol{y}; \mu(X), \Sigma(X, X)\right) = -6.82$

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



Log-Likelihood of $\mathcal{D}$:

$\log N\left(\boldsymbol{y}; \mu(X), \Sigma(X, X)\right) = -8.26$

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

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:

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

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

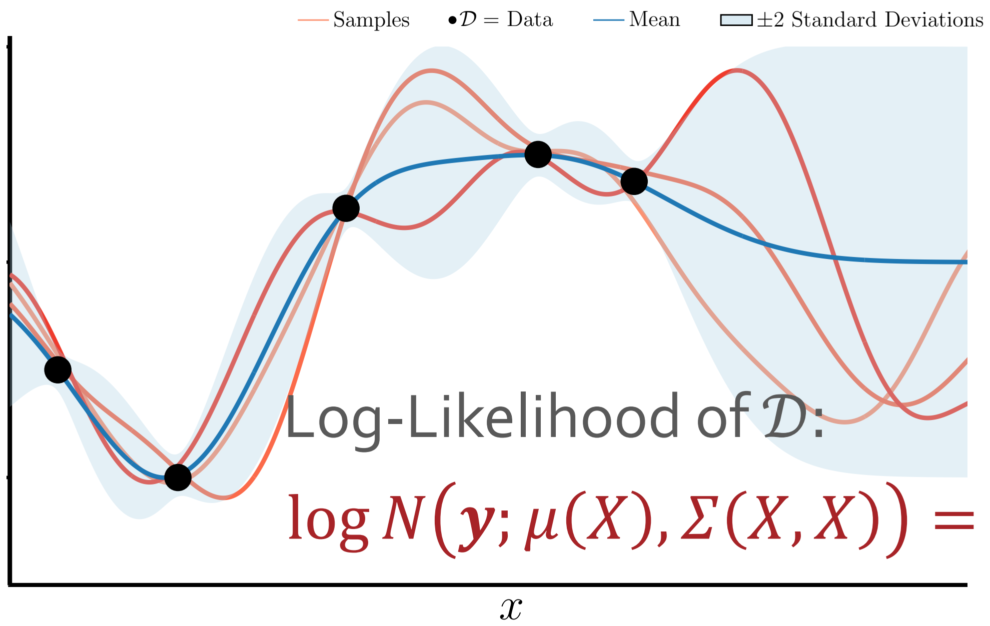
$$y' \mid \boldsymbol{y} = \phi(\boldsymbol{x}')^T\boldsymbol{\omega} \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

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

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

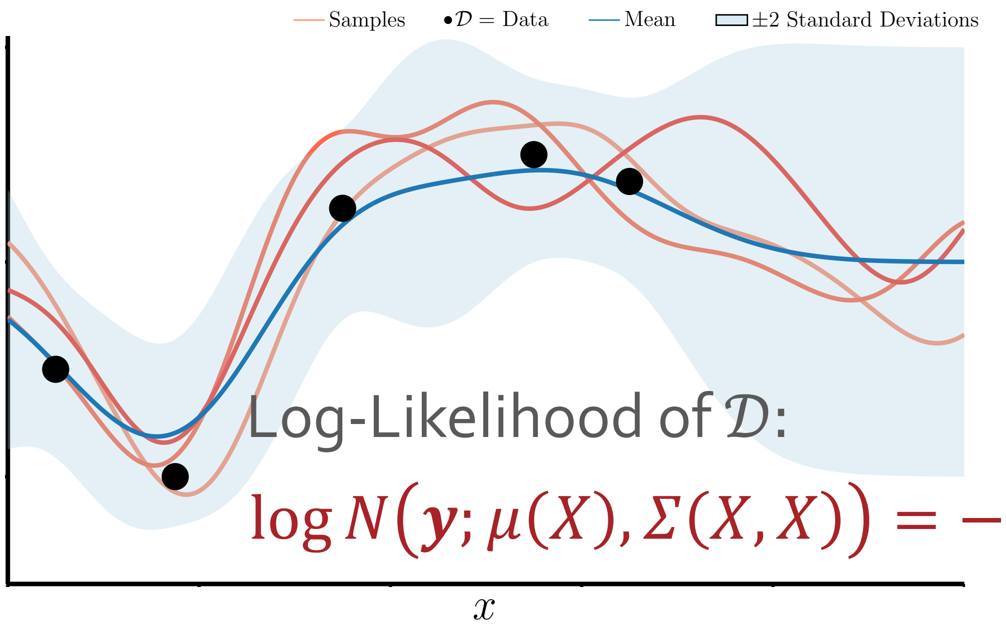
- σ^2 is another hyperparameter we can tune
  - $\sigma^2 = 0$ is a noiseless fit: the mean will always pass through the observations exactly; $\sigma^2 > 0$ allows for deviations



Log-Likelihood of $\mathcal{D}$:

$\log N\big(\boldsymbol{y}; \mu(X), \Sigma(X,X)\big) = -5.11$

$\sigma^2 = 0.1$



Log-Likelihood of $\mathcal{D}$:

$\log N\big(\boldsymbol{y}; \mu(X), \Sigma(X,X)\big) = -7.84$

$\sigma^2 = 0.5$

Noise