10-701: Introduction to Machine Learning Lecture 6 – MLE & MAP

Henry Chai & Zack Lipton

9/18/23

Front Matter

- Announcements:
  - HW1 released 9/6, due 9/20 (Wednesday) at 11:59 PM
  - HW2 released 9/20 (Wednesday), due 10/4 at 11:59 PM
- Recommended Readings:
  - Mitchell, Estimating Probabilities
  - Murphy, Sections 15.1 & 15.2

Probabilistic Learning

- Previously:
  - (Unknown) Target function, $c^*: \mathcal{X} \to \mathcal{Y}$
  - Classifier, $h : \mathcal{X} \to \mathcal{Y}$
  - Goal: find a classifier, h, that best approximates c^*
- Now:
  - (Unknown) Target *distribution*, $y \sim p^*(Y|\boldsymbol{x})$
  - Distribution, $p(Y|\boldsymbol{x})$
  - Goal: find a distribution, p, that best approximates p^*

Likelihood

- Given N independent, identically distribution (iid) samples $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ of a random variable X
  - If X is discrete with probability mass function (pmf) $p(X|\theta)$, then the *likelihood* of $\mathcal{D}$ is

$$L(\theta) = \prod_{n=1}^{N} p(x^{(n)}|\theta)$$

  - If X is continuous with probability density function (pdf) $f(X|\theta)$, then the *likelihood* of $\mathcal{D}$ is

$$L(\theta) = \prod_{n=1}^{N} f(x^{(n)}|\theta)$$

Log-Likelihood

- Given N independent, identically distribution (iid) samples $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ of a random variable X
  - If X is discrete with probability mass function (pmf) $p(X|\theta)$, then the *log-likelihood* of $\mathcal{D}$ is

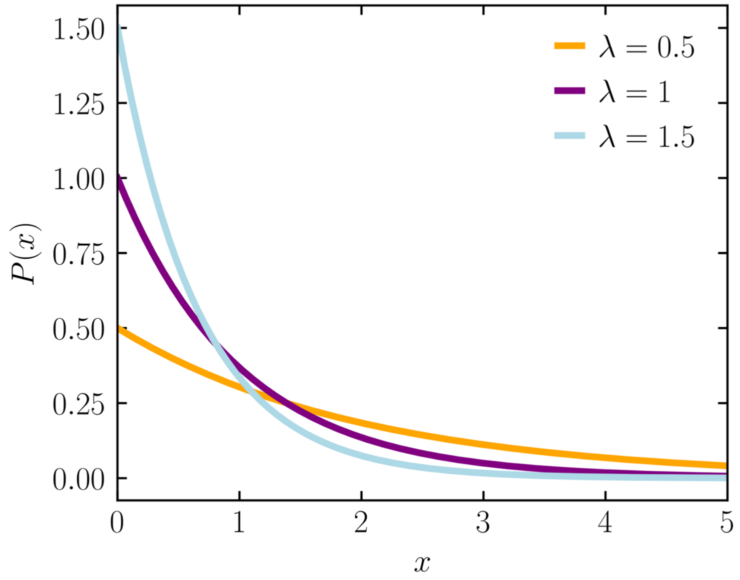
$$\ell(\theta) = \log \prod_{n=1}^{N} p(x^{(n)}|\theta) = \sum_{n=1}^{N} \log p(x^{(n)}|\theta)$$

  - If X is continuous with probability density function (pdf) $f(X|\theta)$, then the *log-likelihood* of $\mathcal{D}$ is

$$\ell(\theta) = \log \prod_{n=1}^{N} f(x^{(n)}|\theta) = \sum_{n=1}^{N} \log f(x^{(n)}|\theta)$$

Maximum Likelihood Estimation (MLE)

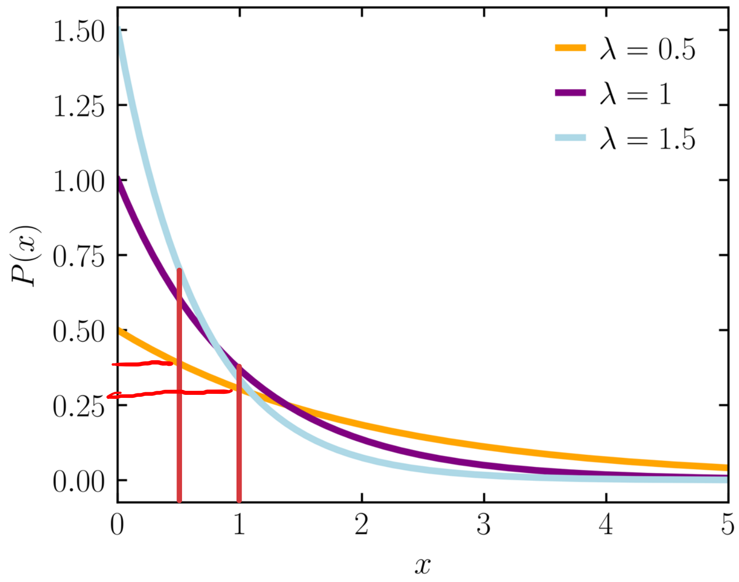
- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*
- Example: the exponential distribution



Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_probability_density.svg

Maximum Likelihood Estimation (MLE)

- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*
- Example: the exponential distribution

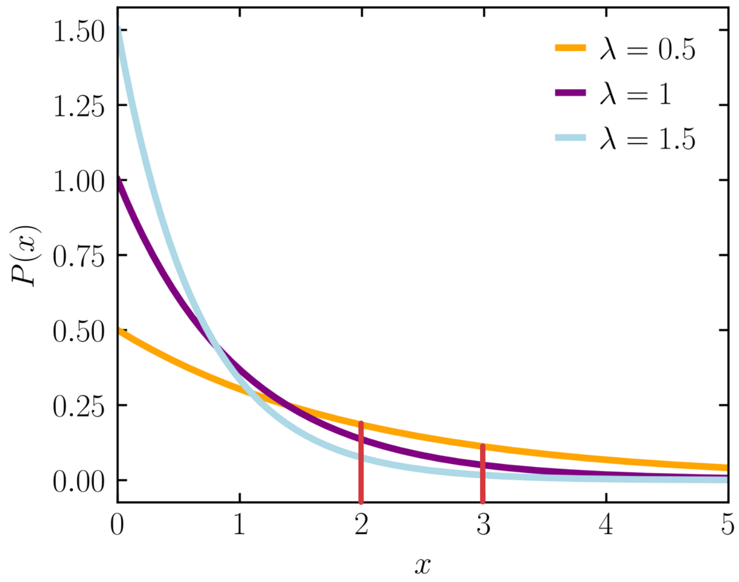


$\{x^{(1)} = 0.5,$
$x^{(2)} = 1\}$

Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_probability_density.svg

Maximum Likelihood Estimation (MLE)

- Insight: every valid probability distribution has a finite amount of probability mass as it must sum/integrate to 1
- Idea: set the parameter(s) so that the likelihood of the samples is maximized
- Intuition: assign as much of the (finite) probability mass to the observed data *at the expense of unobserved data*
- Example: the exponential distribution



$\{x^{(1)} = 2, x^{(2)} = 3\}$

Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_probability_density.svg

Exponential Distribution MLE

- The pdf of the exponential distribution is

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

- Given N iid samples $\{x^{(1)}, \dots, x^{(N)}\}$, the likelihood is

$$L(\lambda) = \prod_{n=1}^{N} \lambda e^{-\lambda x^{(n)}}$$

Exponential Distribution MLE

- The pdf of the exponential distribution is

$$f(x|\lambda) = \lambda e^{-\lambda x}$$

- Given N iid samples $\{x^{(1)}, \dots, x^{(N)}\}$, the log-likelihood is

$$\ell(\lambda) = \log \prod_{n=1}^{N} \lambda e^{-\lambda x^{(n)}} = \sum_{n=1}^{N} \log\left(\lambda e^{-\lambda x^{(n)}}\right)$$

$$= \sum_{n=1}^{N} \left(\log \lambda + \log e^{-\lambda x^{(n)}}\right)$$

$$= \sum_{n=1}^{N} \left(\log \lambda - \lambda x^{(n)}\right)$$

$$= N \log \lambda - \sum_{n=1}^{N} \lambda x^{(n)}$$

$$\frac{\partial \ell}{\partial \lambda} = \frac{N}{\lambda} - \sum_{n=1}^{N} x^{(n)} \rightarrow \frac{N}{\hat{\lambda}} - \sum_{n=1}^{N} x^{(n)} = 0 \rightarrow \hat{\lambda} = \frac{N}{\sum_{n=1}^{N} x^{(n)}}$$

Bernoulli Distribution MLE

- A Bernoulli random variable takes value 1 with probability ϕ and value 0 with probability $1 - \phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1 - \phi)^{1-x}$$

Coin Flipping MLE

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x (1-\phi)^{1-x}$$

- Given N iid samples $\{x^{(1)}, \ldots, x^{(N)}\}$, the log-likelihood is

$$\ell(\phi) = \sum_{n=1}^{N} \log\left(\phi^{x^{(n)}} \cdot (1-\phi)^{1-x^{(n)}}\right)$$

$$= \sum_{n=1}^{N} \left(x^{(n)} \log \phi + (1-x^{(n)}) \log(1-\phi)\right)$$

$(\log(a^b) = b \log a)$

let N_i be the # of i's in the samples

$$= N_1 \log \phi + N_0 \log(1-\phi)$$

Coin Flipping MLE

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

- The partial derivative of the log-likelihood is

$$\ell(\phi) = N_1 \log \phi + N_0 \log(1-\phi)$$

$$\Rightarrow \frac{\partial \ell}{\partial \phi} = \frac{N_1}{\phi} - \frac{N_0}{1-\phi}$$

$$\Rightarrow \frac{N_1}{\hat{\phi}} - \frac{N_0}{1-\hat{\phi}} = 0$$

$$\Rightarrow \frac{N_1}{\hat{\phi}} = \frac{N_0}{1-\hat{\phi}} \Rightarrow N_1 - N_1\hat{\phi} = N_0\hat{\phi}$$

$$\Rightarrow N_1 / (N_1 + N_0) = \hat{\phi}$$

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: $\hat{\Theta}_{MLE} = \arg\max_{\Theta} P(D \mid \Theta)$

MAP: $\hat{\Theta}_{MAP} = \arg\max_{\Theta} P(\Theta \mid D)$ (posterior)

$= \arg\max_{\Theta} \frac{P(D \mid \Theta) P(\Theta)}{P(D)}$

$= \arg\max_{\Theta} P(D \mid \Theta) \, P(\Theta)$

likelihood: $P(D \mid \Theta)$; prior: $P(\Theta)$

Coin Flipping MAP

- A Bernoulli random variable takes value 1 (or heads) with probability ϕ and value 0 (or tails) with probability $1-\phi$
- The pmf of the Bernoulli distribution is

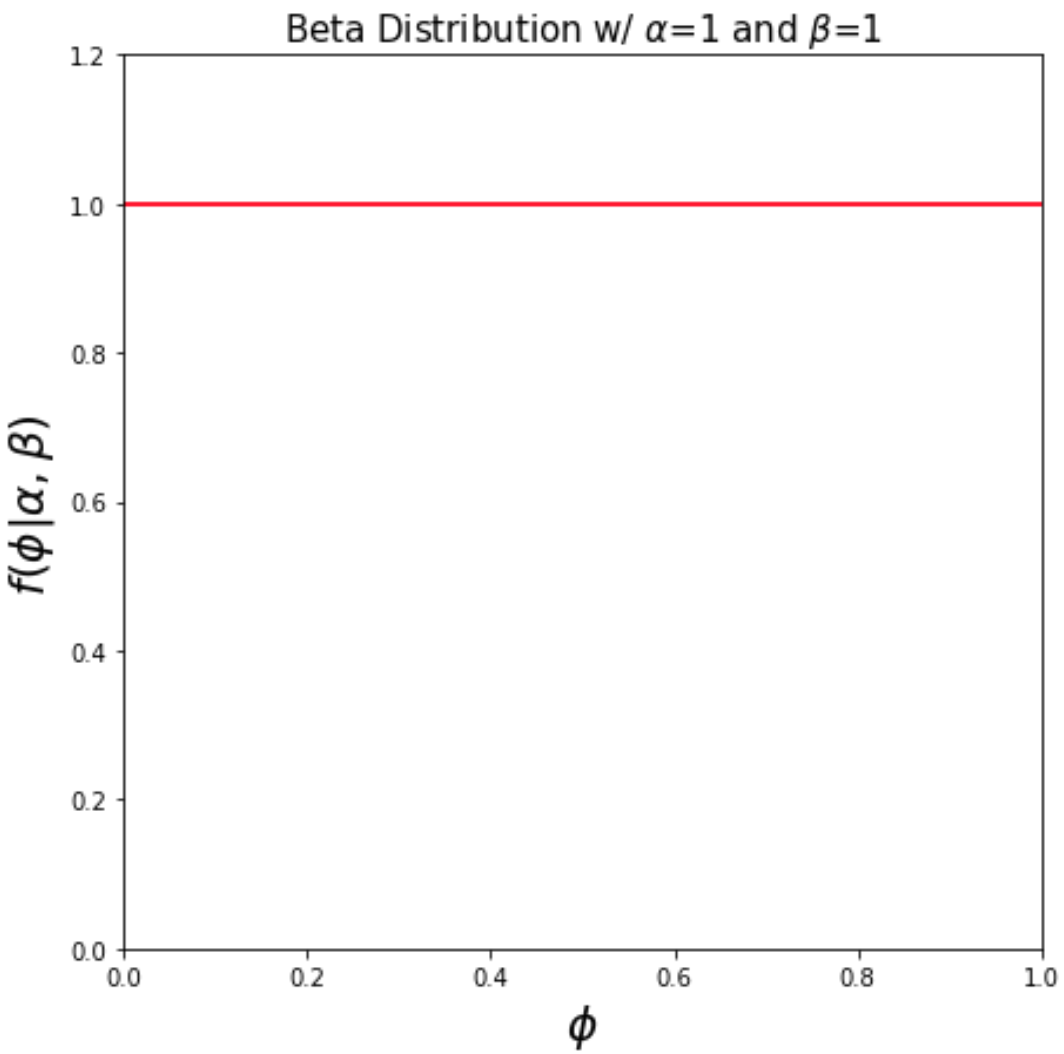
$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

- Assume a Beta prior over the parameter ϕ, which has pdf

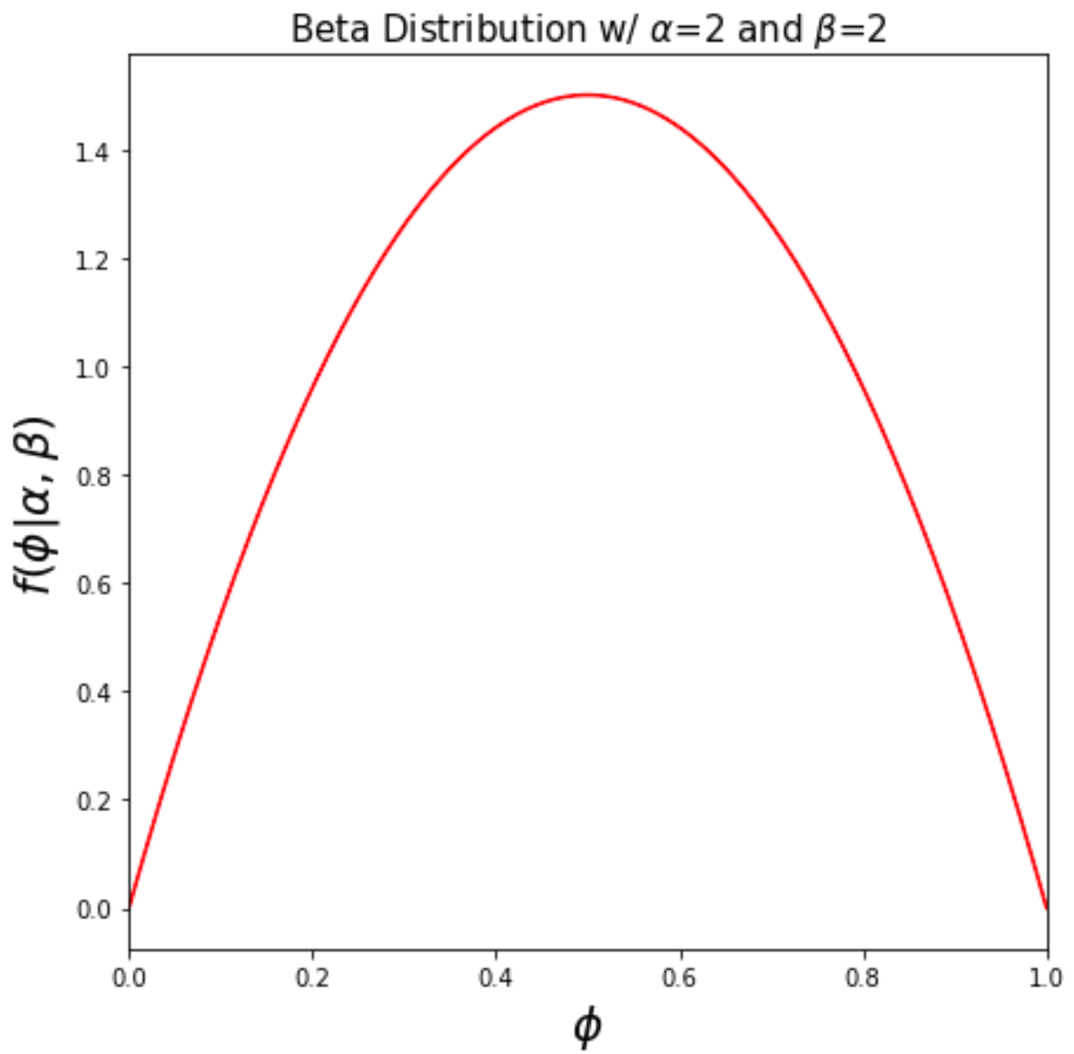
$$f(\phi|\alpha,\beta) = \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where $\mathrm{B}(\alpha,\beta) = \int_0^1 \phi^{\alpha-1}(1-\phi)^{\beta-1}d\phi$ is a normalizing constant to ensure the distribution integrates to 1

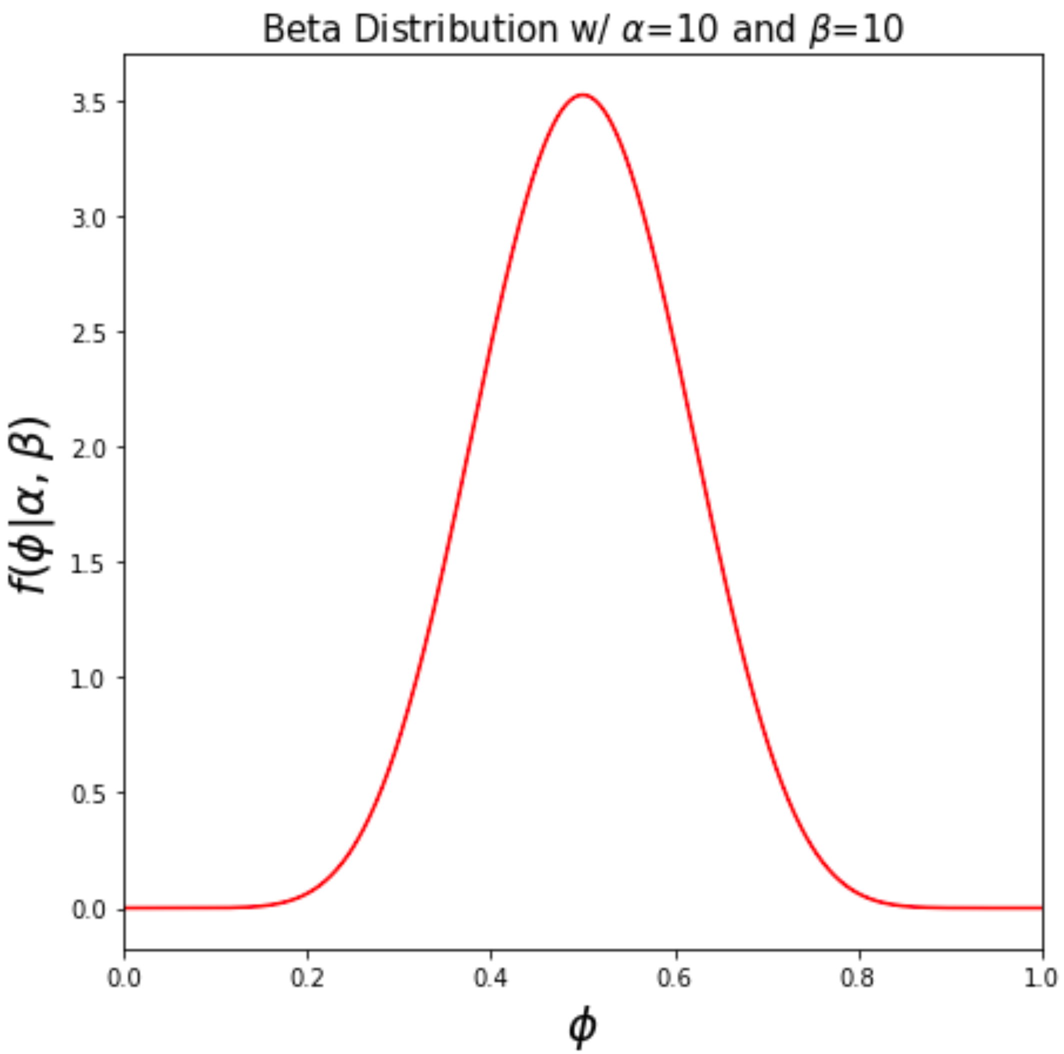
Beta Distribution



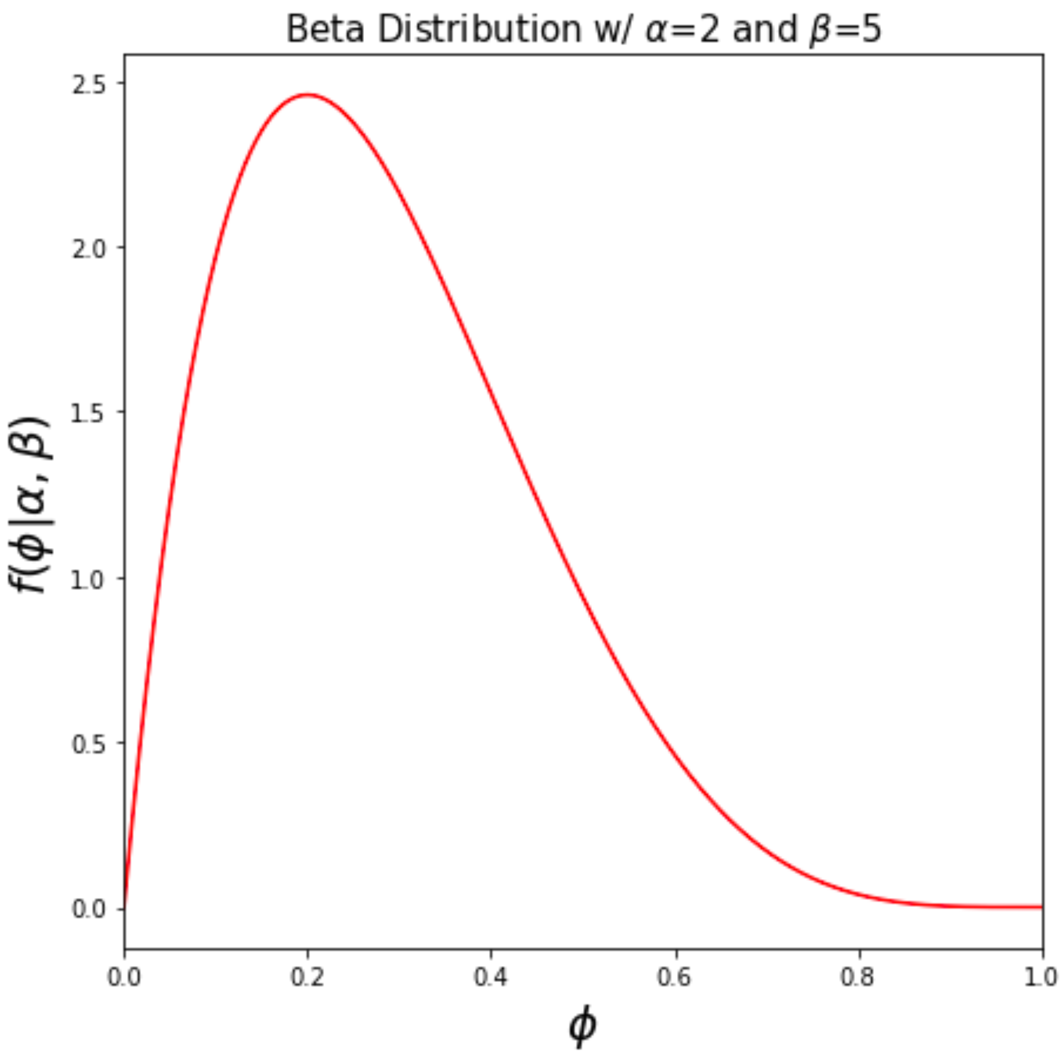
Beta Distribution



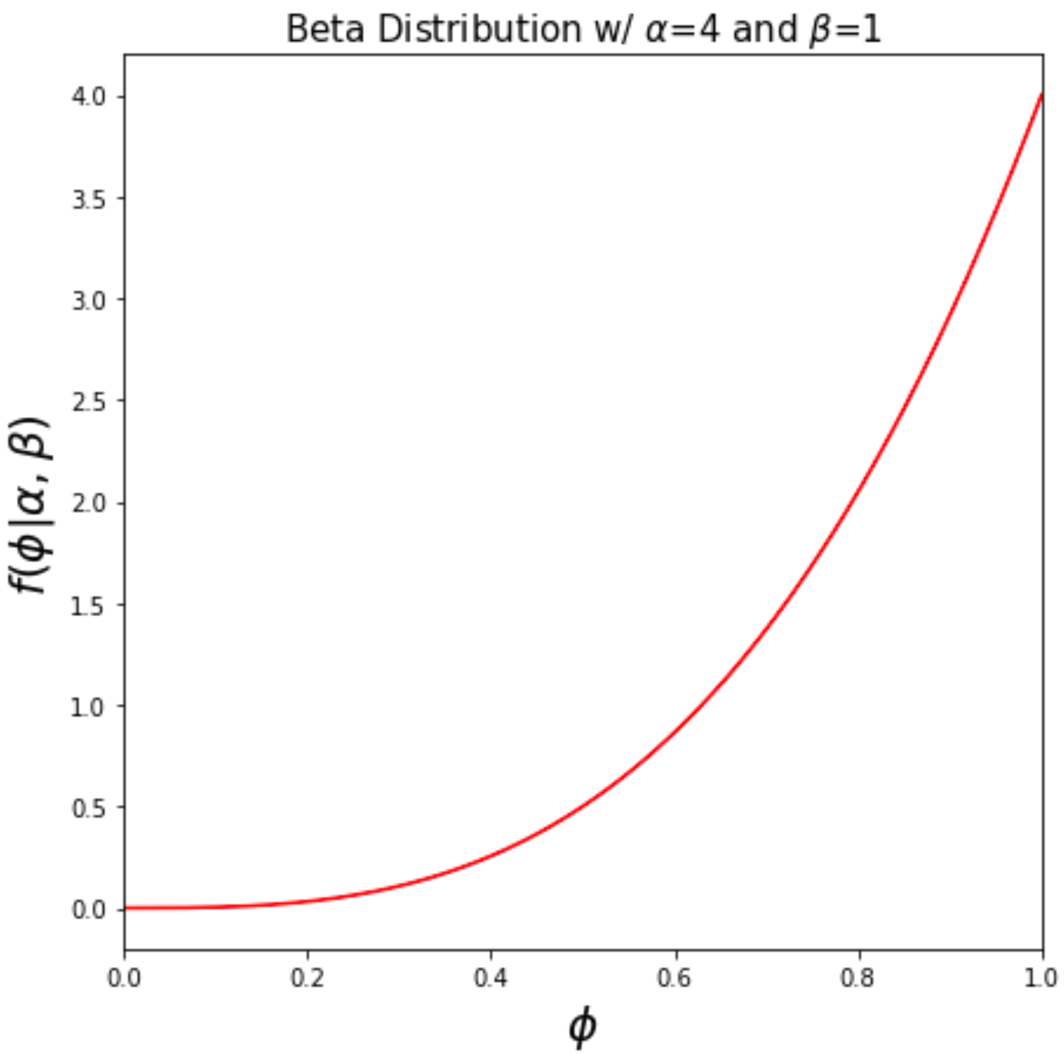
Beta Distribution



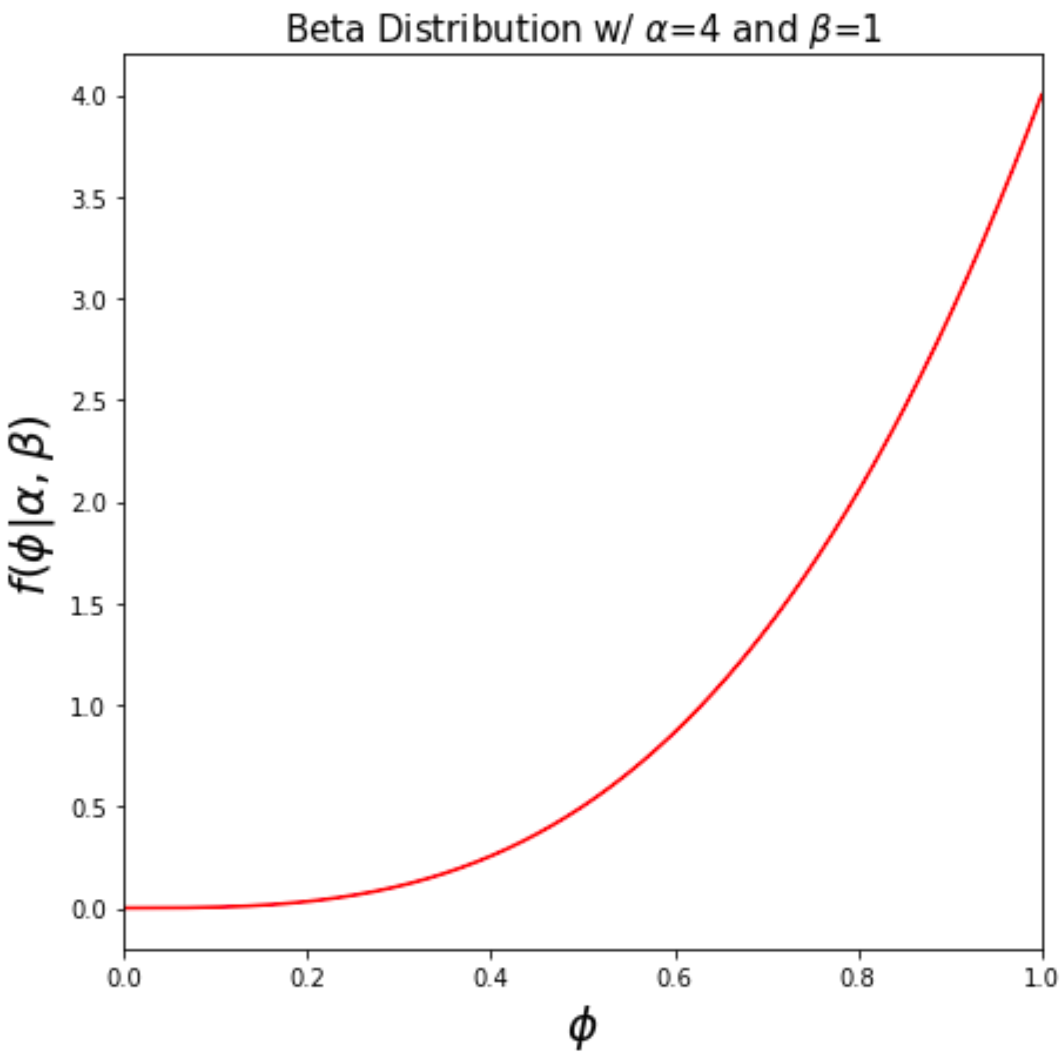
Beta Distribution



Beta Distribution



Okay, but why should we use this strange distribution as a prior?



Conjugate Priors

- For a given likelihood function $p(\mathcal{D}|\theta)$, a prior $p(\theta)$ is called a *conjugate prior* if the resulting posterior distribution $p(\theta|\mathcal{D})$ is in the same family as $p(\theta)$ i.e., $p(\theta|\mathcal{D})$ and $p(\theta)$ are the same type of random variable just with different parameters
  - We like conjugate priors because they are mathematically convenient
  - However, we do not **have** to use a conjugate prior if it doesn't align with our actual prior belief.

Example: Beta-Binomial Conjugacy

$$f(\phi|x,\alpha,\beta) = \frac{p(x|\phi)f(\phi|\alpha,\beta)}{p(x|\alpha,\beta)}$$

$$p(x|\alpha,\beta) = \int_0^1 p(x|\phi)\, f(\phi|\alpha,\beta)\, d\phi$$

$$= \int_0^1 \phi^x (1-\phi)^{1-x} \frac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha,\beta)}\, d\phi$$

$$B(\alpha,\beta) = \int_0^1 \phi^{\alpha-1}(1-\phi)^{\beta-1}\, d\phi$$

$$= \frac{1}{B(\alpha,\beta)} \int_0^1 \phi^{(\alpha+x-1)} (1-\phi)^{\beta-x}\, d\phi$$

$$p(x|\alpha,\beta) = \frac{B(\alpha+x, \beta+(1-x))}{B(\alpha,\beta)}$$

Example: Beta-Binomial Conjugacy

$$f(\phi|x,\alpha,\beta) = \frac{p(x|\phi)f(\phi|\alpha,\beta)}{p(x|\alpha,\beta)} = \frac{p(x|\phi)f(\phi|\alpha,\beta)}{\int p(x|\phi)f(\phi|\alpha,\beta)d\phi}$$

$$f(\phi|x,\alpha,\beta) = \frac{\phi^{x}(1-\phi)^{1-x}\,\dfrac{\phi^{\alpha-1}(1-\phi)^{\beta-1}}{B(\alpha,\beta)}}{\left(\dfrac{B(\alpha+x,\beta+(1-x))}{B(\alpha,\beta)}\right)}$$

$$= \frac{\phi^{\alpha+x-1}(1-\phi)^{\beta+(1-x)-1}}{B(\alpha+x,\beta+(1-x))}$$

A Beta distribution w/ parameters $\alpha+x$ & $\beta+(1-x)$

Beta-Binomial MAP

- Given N iid samples $\{x^{(1)}, \ldots, x^{(N)}\}$, the log-posterior is

$$\ell(\phi) = \log\left(f\left(\phi \mid x^{(1)}, \ldots, x^{(N)}, \alpha, \beta\right)\right)$$

$$= \log\left(f_{post}\left(\phi \mid \alpha + \sum_{n=1}^{N} x^{(n)}, \beta + \sum_{n=1}^{N}\left(1 - x^{(n)}\right)\right)\right)$$

$$= \log\left(f_{post}\left(\phi \mid \alpha + N_1, \beta + N_0\right)\right)$$

$$= \log\left(\frac{\phi^{\alpha + N_1 - 1}(1-\phi)^{\beta + N_0 - 1}}{B(\alpha + N_1, \beta + N_0)}\right)$$

$$\rightarrow = (\alpha + N_1 - 1)\log\phi + (\beta + N_0 - 1)\log(1-\phi) - \log\left(B(\alpha + N_1, \beta + N_0)\right)$$

Beta-Binomial MAP

- Given N iid samples $\{x^{(1)}, \ldots, x^{(N)}\}$, the partial derivative of the log-posterior is

$$\frac{\partial \ell}{\partial \phi} = \frac{(\alpha + N_1 - 1)}{\phi} - \frac{(\beta + N_0 - 1)}{1 - \phi}$$

$$\vdots$$

$$\hat{\phi} = \frac{\alpha + N_1 - 1}{(\alpha + N_1 - 1) + (\beta + N_0 - 1)}$$

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 2$ and $\beta = 5$, then

$$\phi_{MAP} = \frac{10+2-1}{11+16} = \frac{11}{17} < \frac{10}{12}$$

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 101$ and $\beta = 101$, then

$$\phi_{MAP} = \frac{10+100}{10+100+2+100} = \frac{110}{212} \approx \frac{1}{2}$$

Coin Flipping MAP: Example

- Suppose $\mathcal{D}$ consists of ten 1's or heads ($N_1 = 10$) and two 0's or tails ($N_0 = 2$):

$$\phi_{MLE} = \frac{10}{10+2} = \frac{10}{12}$$

- Using a Beta prior with $\alpha = 1$ and $\beta = 1$, then

$$\phi_{MAP} = \phi_{MLE}$$

M(C)LE for Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

then given $X = \begin{bmatrix} 1 & \boldsymbol{x}^{(1)^T} \\ 1 & \boldsymbol{x}^{(2)^T} \\ \vdots & \vdots \\ 1 & \boldsymbol{x}^{(N)^T} \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$, the MLE of $\boldsymbol{\omega}$ is

$$\hat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \log P(\boldsymbol{y} | X, \boldsymbol{\omega})$$

$$\vdots$$

$$= (X^T X)^{-1} X^T \boldsymbol{y}$$

MAP for Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and **independent, identical** Gaussian priors on the weights …

$$\omega_d \sim N(0, s^2) \rightarrow \boldsymbol{\omega} \sim N(\boldsymbol{0}, s^2 I_{D+1})$$

then, the MAP of $\boldsymbol{\omega}$ is the ridge regression solution!

$$\widehat{\boldsymbol{\omega}} = \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \log P(\boldsymbol{\omega} | X, \boldsymbol{y})$$

$$\vdots$$

$$= (X^T X + \lambda(s^2) I_{D+1})^{-1} X^T \boldsymbol{y}$$

Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \Sigma)$$

then the distribution over $\boldsymbol{y}$ is $\quad y = X\omega + \epsilon$

$$\boldsymbol{y} \sim N(X\mathbf{0} + \mathbf{0} = \mathbf{0}, X\Sigma X^T + \sigma^2 I)$$

Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \text{ …}$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \Sigma)$$

then the *joint* distribution over $\boldsymbol{y}$ and $\boldsymbol{\omega}$ is

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\omega} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} X\Sigma X^T + \sigma^2 I & ??? \\ ??? & \Sigma \end{bmatrix}\right)$$

$$\mathrm{Cov}(y, w) = \mathrm{Cov}(Xw + \epsilon, w)$$

$$= X\,\mathrm{Cov}(w, w) = X\Sigma$$

Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \Sigma)$$

then the *joint* distribution over $\boldsymbol{y}$ and $\boldsymbol{\omega}$ is

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\omega} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} X\Sigma X^T + \sigma^2 I & \Sigma X^T \\ X\Sigma & \Sigma \end{bmatrix} \right)$$

Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\boldsymbol{0}, \Sigma)$$

then the *conditional* distribution of $\boldsymbol{\omega}$ given $\boldsymbol{y}$ is

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

where

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

$$\Sigma_{POST} = \Sigma - \Sigma X^T (X \Sigma X^T + \sigma^2 I)^{-1} X \Sigma$$

Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \Sigma)$$

then the *conditional* distribution of $h(\boldsymbol{x}') = \boldsymbol{x}'^T \boldsymbol{\omega}$ given $\boldsymbol{y}$ is

$$h(\boldsymbol{x}') \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

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

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

Kernelized Bayesian Linear Regression

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \Sigma)$$

then the *conditional* distribution of $h(\boldsymbol{x}') = \boldsymbol{x}'^T \boldsymbol{\omega}$ given $\boldsymbol{y}$ is

$$h(\boldsymbol{x}') \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

$$K(\boldsymbol{a}, \boldsymbol{b}) = \Phi(\boldsymbol{a})^T \Sigma \Phi(\boldsymbol{b})$$

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

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

Kernelized Bayesian Linear Regression = Gaussian Process (GP)

- If we assume a linear model with additive Gaussian noise

$$y = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\boldsymbol{\omega}^T \boldsymbol{x}, \sigma^2) \ldots$$

and a **general** (zero-mean) Gaussian prior on the weights …

$$\boldsymbol{\omega} \sim N(\mathbf{0}, \Sigma)$$

then the *conditional* distribution of $h(\boldsymbol{x}') = \boldsymbol{x}'^T \boldsymbol{\omega}$ given $\boldsymbol{y}$ is

$$h(\boldsymbol{x}') \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}_{PRED}, \Sigma_{PRED})$$

where

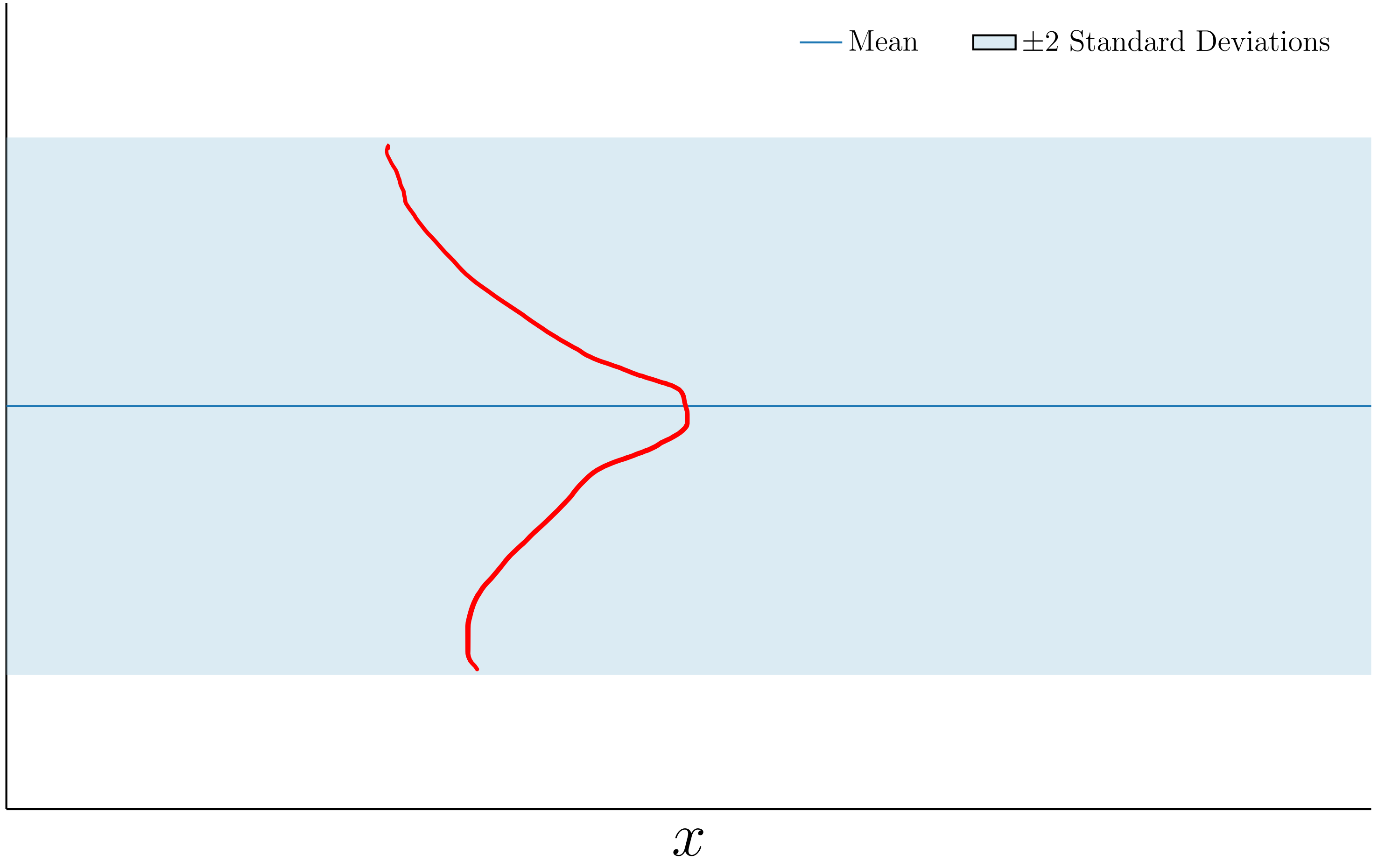
$$K(\boldsymbol{a}, \boldsymbol{b}) = \Phi(\boldsymbol{a})^T \Sigma \Phi(\boldsymbol{b})$$

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

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

Gaussian Process (GP)

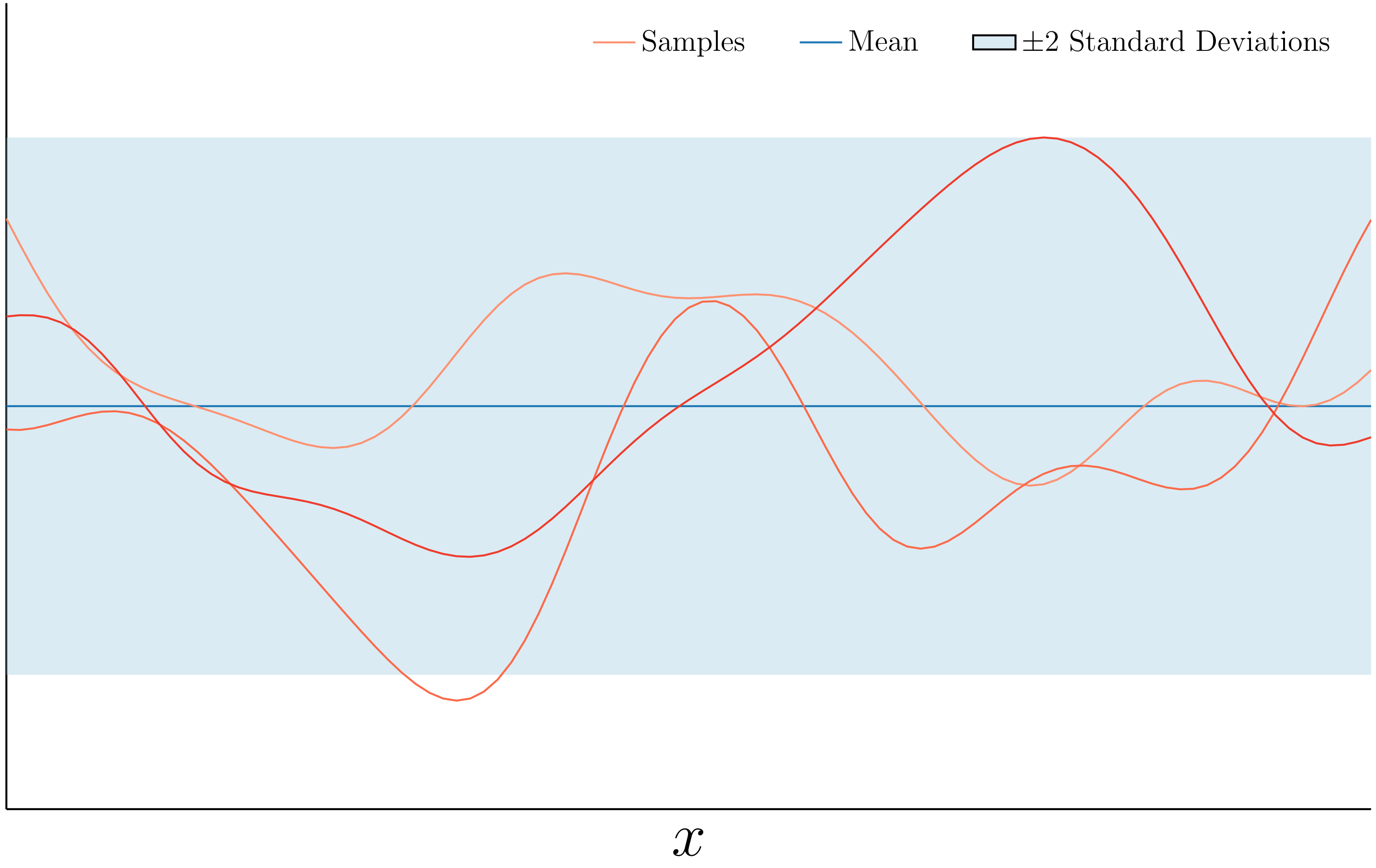
$$f \sim \mathcal{GP}(m(x) = 0, K(x, x') = \exp(-(x - x')^2))$$



$$f \sim \mathcal{GP}(m, K) \rightarrow f(x) \sim \mathcal{N}\big(m(x), K(x, x)\big)$$

Gaussian Process (GP)

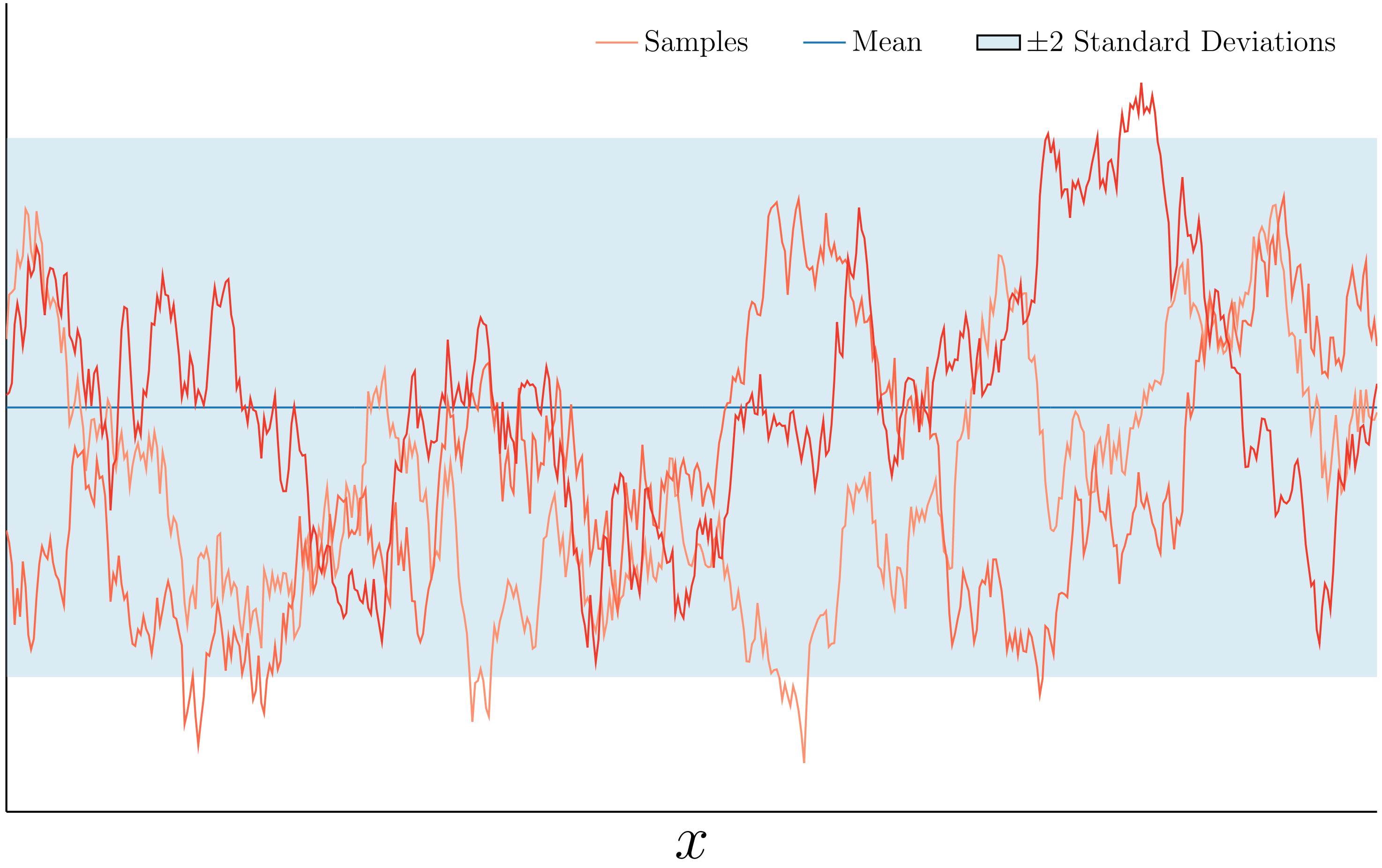
$$f \sim \mathcal{GP}(m(x) = 0, K(x, x') = \exp(-(x - x')^2))$$



$$f \sim \mathcal{GP}(m, K) \rightarrow f(x) \sim \mathcal{N}\big(m(x), K(x, x)\big)$$

Gaussian Process (GP)

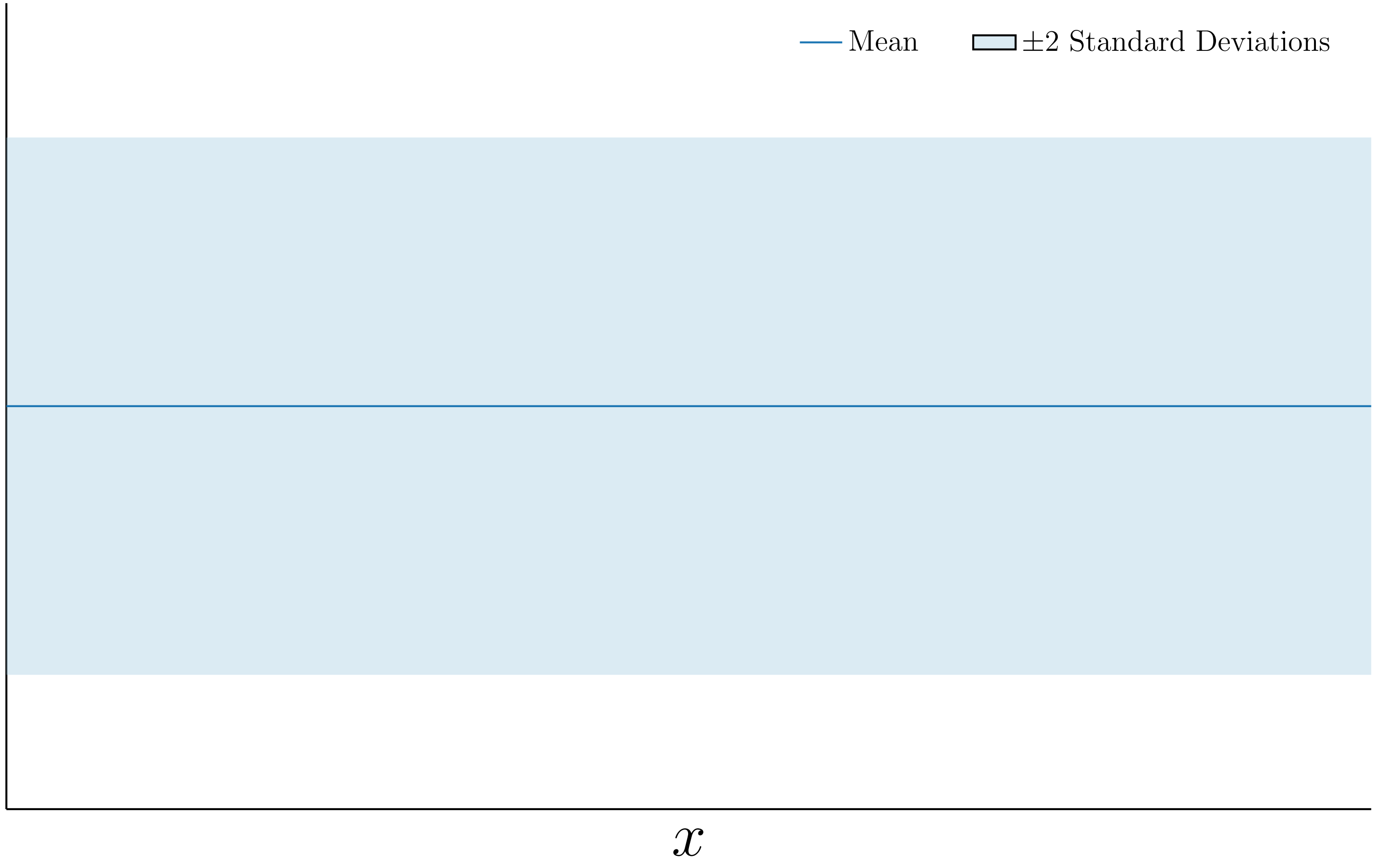
$$f \sim \mathcal{GP}(m(x) = 0, K(x, x') = \exp(-|x - x'|))$$



$$f \sim \mathcal{GP}(m, K) \rightarrow f(x) \sim \mathcal{N}\big(m(x), K(x, x)\big)$$

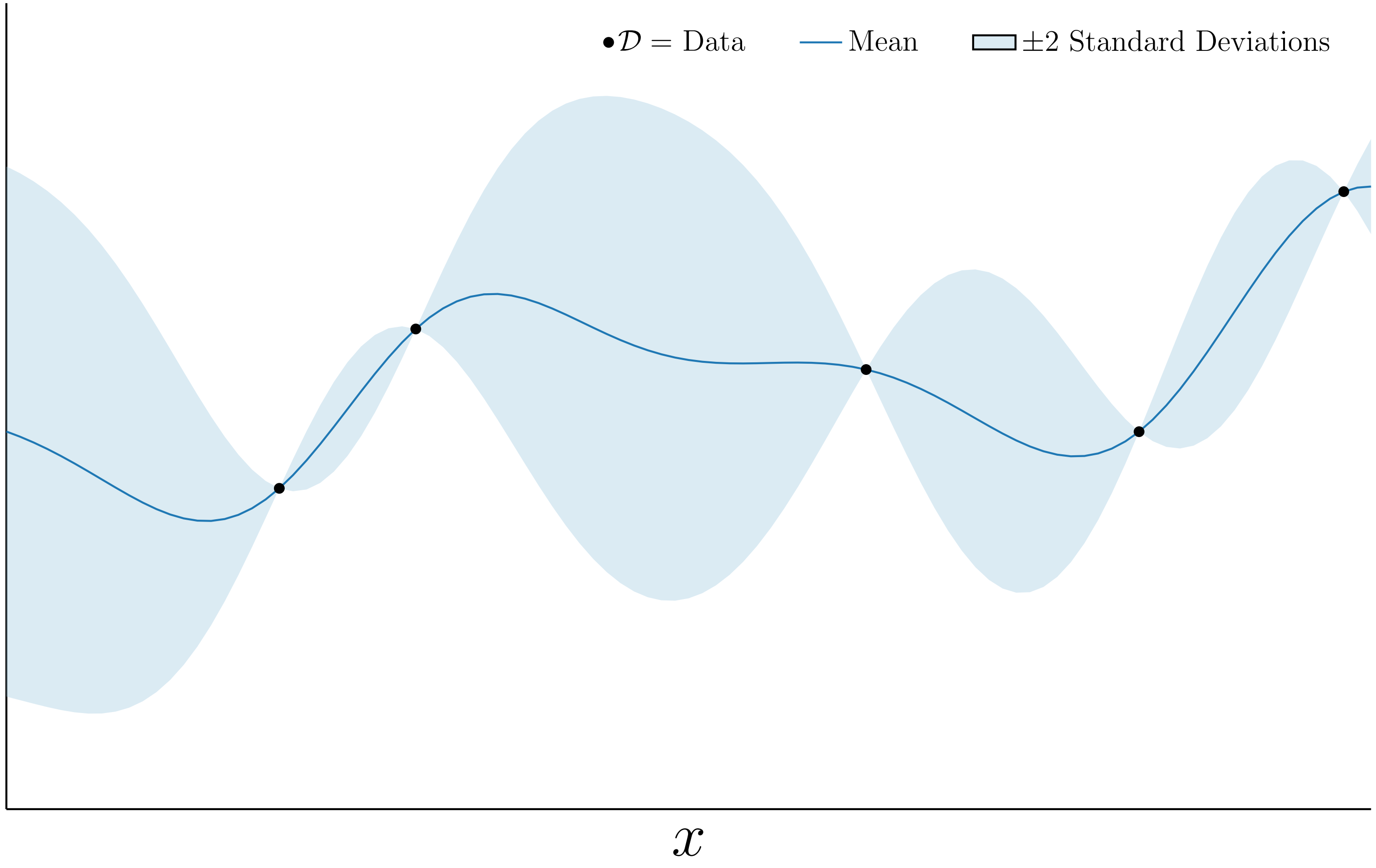
GP Prior

$$f \sim \mathcal{GP}(m(x) = 0, K(x, x') = \exp(-(x - x')^2))$$



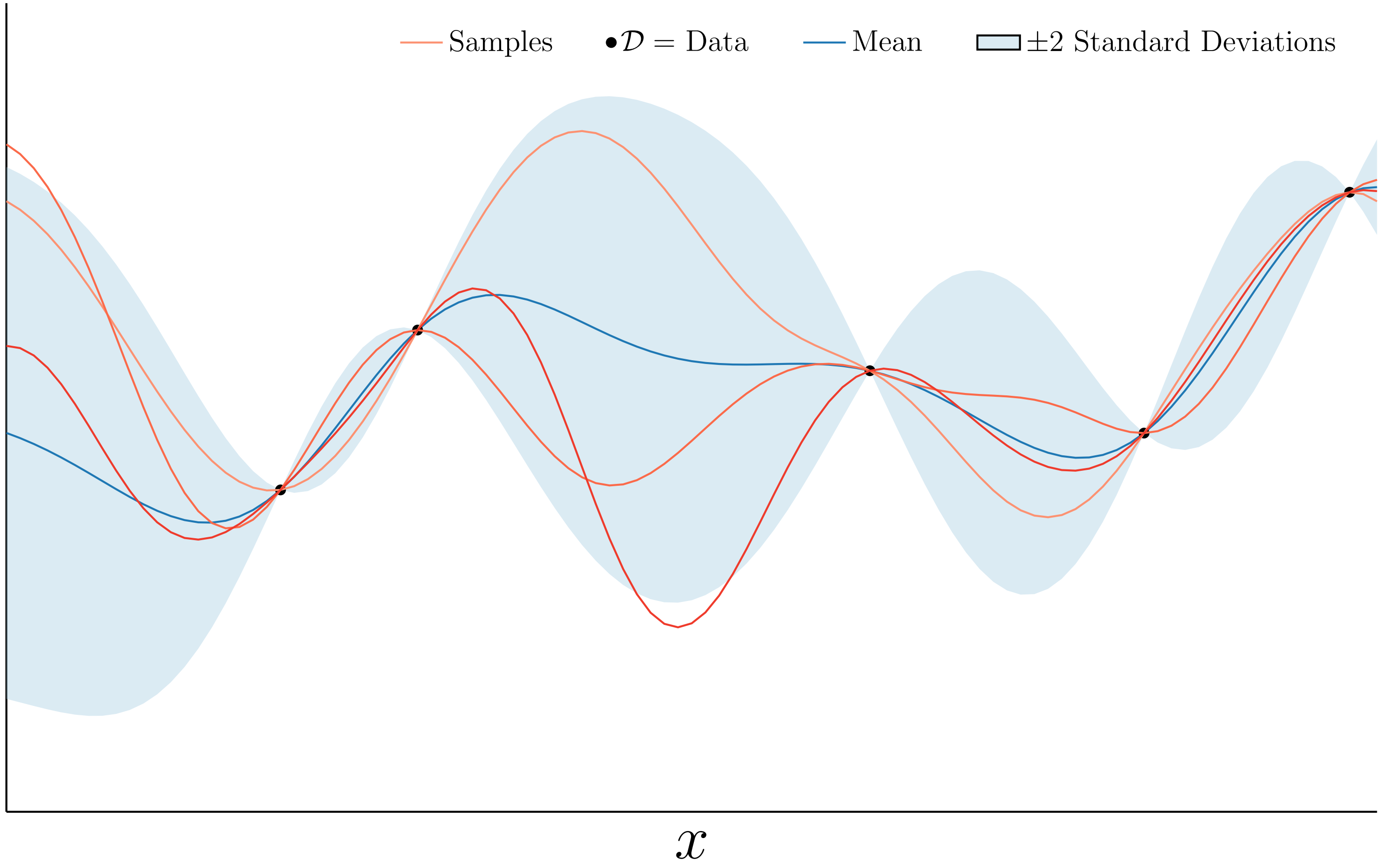
GP Posterior

$$f \mid \mathcal{D} \sim \mathcal{GP}(m_{\mathcal{D}}, K_{\mathcal{D}})$$



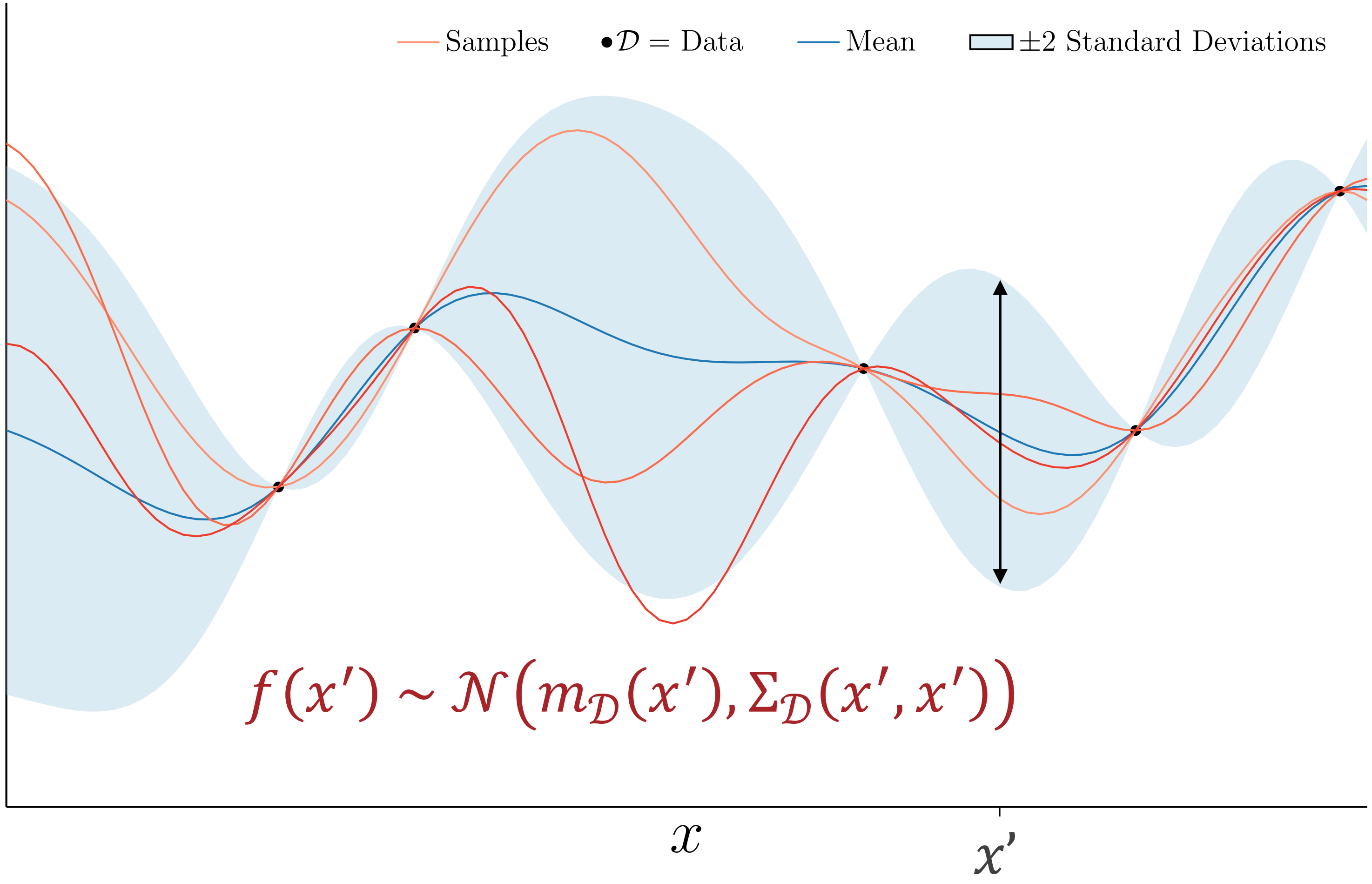
GP Posterior

$$f \mid \mathcal{D} \sim \mathcal{GP}(m_{\mathcal{D}}, K_{\mathcal{D}})$$



GP Posterior

$$f \mid \mathcal{D} \sim \mathcal{GP}(m_{\mathcal{D}}, K_{\mathcal{D}})$$



$$f(x') \sim \mathcal{N}\left(m_{\mathcal{D}}(x'), \Sigma_{\mathcal{D}}(x', x')\right)$$

Key Takeaways

- Two ways of estimating the parameters of a probability distribution given samples of a random variable:
  - Maximum likelihood estimation – maximize the (log-)likelihood of the observations
  - Maximum a posteriori estimation – maximize the (log-)posterior of the parameters conditioned on the observations
    - Requires a prior distribution, drawn from background knowledge or domain expertise
- Linear/ridge regression can be interpreted as MLE/MAP estimators under certain likelihood/prior models
  - A Gaussian process is the kernelization of Bayesian linear regression or MAP estimation for linear regression