Imperial College London

Lecture 11: Probability Distributions and Parameter Estimation

Recommended reading:
Bishop: Chapters 1.2, 2.1–2.3.4, Appendix B

Duncan Gillies and Marc Deisenroth

Department of Computing
Imperial College London

February 10, 2016

Key Concepts in Probability Theory

Two fundamental rules:

$$p(x) = \int p(x, y)\mathrm{d}y \qquad \text{Sum rule/Marginalization property}$$

$$p(x, y) = p(y|x)p(x) \qquad \text{Product rule}$$

Key Concepts in Probability Theory

Two fundamental rules:

$$p(x) = \int p(x, y)\mathrm{d}y \qquad \text{Sum rule/Marginalization property}$$

$$p(x, y) = p(y|x)p(x) \qquad \text{Product rule}$$

Bayes' Theorem (Probabilistic Inverse)

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})\,p(\boldsymbol{x})}{p(\boldsymbol{y})}, \qquad \boldsymbol{x} : \text{hypothesis}, \quad \boldsymbol{y} : \text{measurement}$$

Key Concepts in Probability Theory

Two fundamental rules:

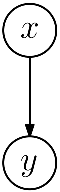
$$p(x) = \int p(x, y) \mathrm{d}y \qquad \text{Sum rule/Marginalization property}$$

$$p(x, y) = p(y|x)p(x) \qquad \text{Product rule}$$

Bayes' Theorem (Probabilistic Inverse)

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})\, p(\boldsymbol{x})}{p(\boldsymbol{y})}, \qquad \boldsymbol{x} : \text{hypothesis}, \quad \boldsymbol{y} : \text{measurement}$$

- Posterior belief
- Prior belief
- Likelihood (measurement model)
- Marginal likelihood (normalization constant)



Mean and (Co)Variance

Mean and covariance are often useful to describe properties of probability distributions (expected values and spread).

Definition

$$\mathbb{E}_x[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} =: \boldsymbol{\mu}$$

$$\mathbb{V}_x[\boldsymbol{x}] = \mathbb{E}_x[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_x[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_x[\boldsymbol{x}]^\top =: \boldsymbol{\Sigma}$$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{x,y}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_y[\boldsymbol{y}]^\top$$

Mean and (Co)Variance

Mean and covariance are often useful to describe properties of probability distributions (expected values and spread).

Definition

$$\mathbb{E}_x[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} =: \boldsymbol{\mu}$$

$$\mathbb{V}_x[\boldsymbol{x}] = \mathbb{E}_x[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_x[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_x[\boldsymbol{x}]^\top =: \boldsymbol{\Sigma}$$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{x,y}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_y[\boldsymbol{y}]^\top$$

Linear/Affine Transformations:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \text{where} \quad \mathbb{E}_x[\boldsymbol{x}] = \boldsymbol{\mu},\ \mathbb{V}_x[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{y}] =$$

$$\mathbb{V}[\boldsymbol{y}] =$$

Mean and (Co)Variance

Mean and covariance are often useful to describe properties of probability distributions (expected values and spread).

Definition

$$\mathbb{E}_x[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} =: \boldsymbol{\mu}$$

$$\mathbb{V}_x[\boldsymbol{x}] = \mathbb{E}_x[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_x[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_x[\boldsymbol{x}]^\top =: \boldsymbol{\Sigma}$$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{x,y}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_y[\boldsymbol{y}]^\top$$

Linear/Affine Transformations:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \text{where} \quad \mathbb{E}_x[\boldsymbol{x}] = \boldsymbol{\mu}, \mathbb{V}_x[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}_x[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_x[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$$

$$\mathbb{V}[\boldsymbol{y}] =$$

Mean and (Co)Variance

Mean and covariance are often useful to describe properties of probability distributions (expected values and spread).

Definition

$$\mathbb{E}_x[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} =: \boldsymbol{\mu}$$

$$\mathbb{V}_x[\boldsymbol{x}] = \mathbb{E}_x[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_x[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_x[\boldsymbol{x}]^\top =: \boldsymbol{\Sigma}$$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{x,y}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_y[\boldsymbol{y}]^\top$$

Linear/Affine Transformations:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \text{where} \quad \mathbb{E}_x[\boldsymbol{x}] = \boldsymbol{\mu}, \mathbb{V}_x[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}_x[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_x[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$$

$$\mathbb{V}[\boldsymbol{y}] = \mathbb{V}_x[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \mathbb{V}_x[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_x[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$$

Mean and (Co)Variance

Mean and covariance are often useful to describe properties of probability distributions (expected values and spread).

Definition

$$\mathbb{E}_x[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} =: \boldsymbol{\mu}$$

$$\mathbb{V}_x[\boldsymbol{x}] = \mathbb{E}_x[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top] = \mathbb{E}_x[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_x[\boldsymbol{x}]^\top =: \boldsymbol{\Sigma}$$

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] = \mathbb{E}_{x,y}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}_x[\boldsymbol{x}]\mathbb{E}_y[\boldsymbol{y}]^\top$$

Linear/Affine Transformations:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \qquad \text{where} \quad \mathbb{E}_x[\boldsymbol{x}] = \boldsymbol{\mu}, \mathbb{V}_x[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}_x[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}_x[\boldsymbol{x}] + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}$$

$$\mathbb{V}[\boldsymbol{y}] = \mathbb{V}_x[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}] = \mathbb{V}_x[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{V}_x[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top$$

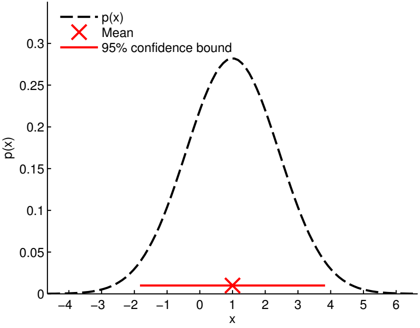
If $\boldsymbol{x}, \boldsymbol{y}$ independent: $\mathbb{V}_{x,y}[\boldsymbol{x}+\boldsymbol{y}] = \mathbb{V}_x[\boldsymbol{x}] + \mathbb{V}_y[\boldsymbol{y}]$

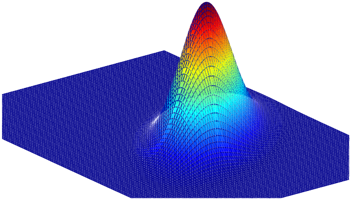
Basic Probability Distributions

The Gaussian Distribution

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Mean vector $\boldsymbol{\mu}$ ⏩ Average of the data
- Covariance matrix $\boldsymbol{\Sigma}$ ⏩ Spread of the data

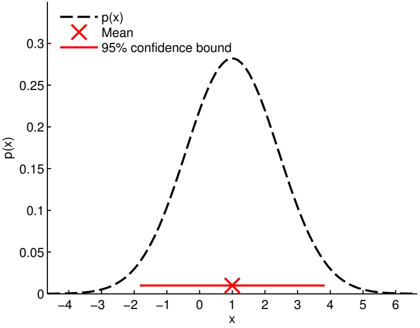




The Gaussian Distribution

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

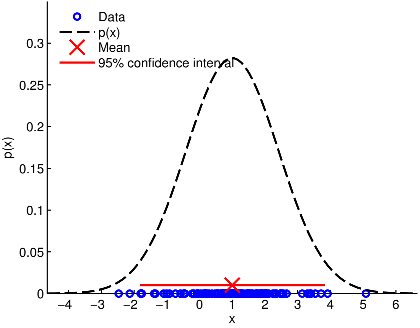
- Mean vector $\boldsymbol{\mu}$ ▶▶ Average of the data
- Covariance matrix $\boldsymbol{\Sigma}$ ▶▶ Spread of the data

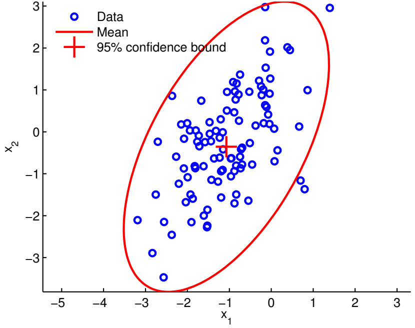


The Gaussian Distribution

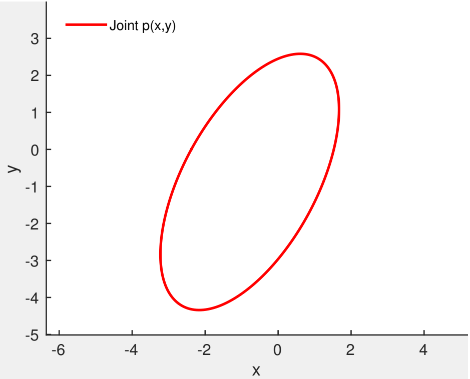
$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Mean vector $\boldsymbol{\mu}$ ⏩ Average of the data
- Covariance matrix $\boldsymbol{\Sigma}$ ⏩ Spread of the data



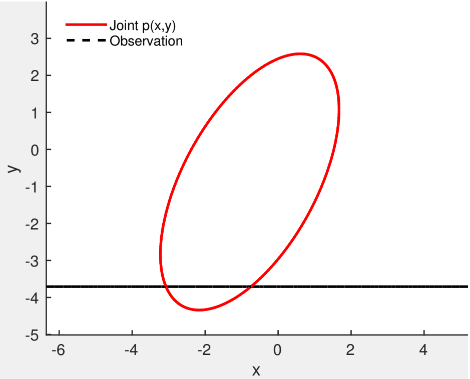


Conditional



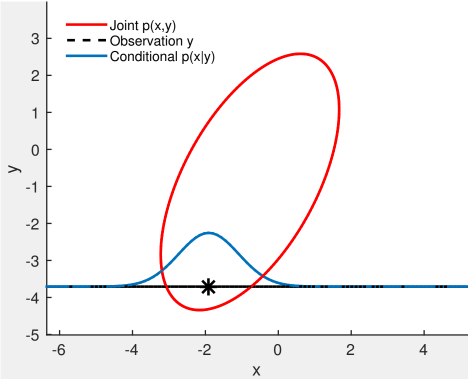
$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right)$$

Conditional



$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right)$$

Conditional



$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right)$$

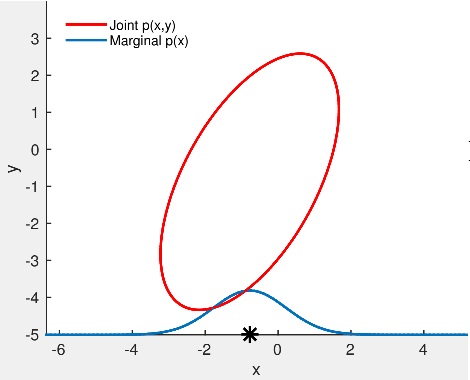
$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}\right)$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)$$

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$$

Conditional $p(\boldsymbol{x}|\boldsymbol{y})$ is also Gaussian
▶▶ Computationally convenient

Marginal

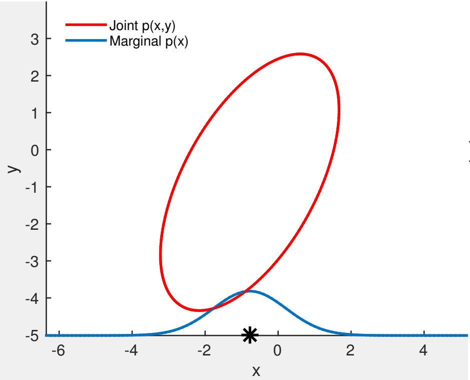


$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right)$$

Marginal distribution:

$$\begin{aligned} p(\boldsymbol{x}) &= \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \\ &= \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}) \end{aligned}$$

Marginal



$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}\right)$$

Marginal distribution:

$$\begin{aligned} p(\boldsymbol{x}) &= \int p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \\ &= \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}) \end{aligned}$$

- The marginal of a joint Gaussian distribution is Gaussian
- Intuitively: Ignore (integrate out) everything you are not interested in

Bernoulli Distribution

- Distribution for a single binary variable $x \in \{0, 1\}$
- Governed by a single continuous parameter $\mu \in [0, 1]$ that represents the probability of $x \in \{0, 1\}$.

$$
\begin{aligned}
p(x|\mu) &= \mu^x (1-\mu)^{1-x} \\
\mathbb{E}[x] &= \mu \\
\mathbb{V}[x] &= \mu(1-\mu)
\end{aligned}
$$

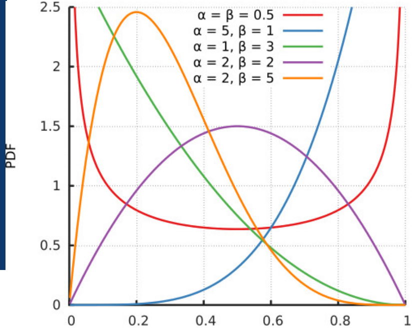
Bernoulli Distribution

- Distribution for a single binary variable $x \in \{0, 1\}$
- Governed by a single continuous parameter $\mu \in [0, 1]$ that represents the probability of $x \in \{0, 1\}$.

$$
\begin{aligned}
p(x|\mu) &= \mu^x (1-\mu)^{1-x} \\
\mathbb{E}[x] &= \mu \\
\mathbb{V}[x] &= \mu(1-\mu)
\end{aligned}
$$

- Example: Result of flipping a coin.

Beta Distribution

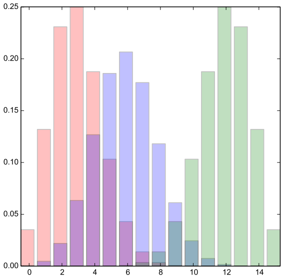


- Distribution over a continuous variable $\mu \in [0, 1]$, which is often used to represent the probability for some binary event (see Bernoulli distribution)
- Governed by two parameters $\alpha > 0,\ \beta > 0$

$$p(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha+\beta}, \qquad \mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Binomial Distribution

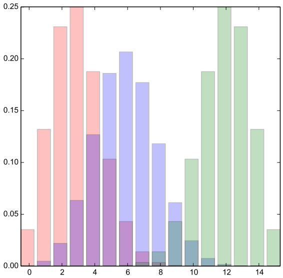


- Generalization of the Bernoulli distribution to a distribution over integers
- Probability of observing m occurrences of $x = 1$ in a set of N samples from a Bernoulli distribution, where $p(x = 1) = \mu \in [0, 1]$

$$p(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$\mathbb{E}[m] = N\mu \,, \qquad \mathbb{V}[m] = N\mu(1 - \mu)$$

Binomial Distribution



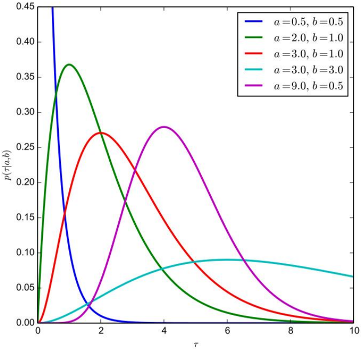
- Generalization of the Bernoulli distribution to a distribution over integers
- Probability of observing m occurrences of $x = 1$ in a set of N samples from a Bernoulli distribution, where $p(x = 1) = \mu \in [0, 1]$

$$p(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$\mathbb{E}[m] = N\mu\,, \qquad \mathbb{V}[m] = N\mu(1 - \mu)$$

Example: What is the probability of observing m heads in N experiments if the probability for observing head in a single experiment is μ?

Gamma Distribution



$$p(\tau|a,b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau)$$

$$\mathbb{E}[\tau] = \frac{a}{b}$$

$$\mathbb{V}[\tau] = \frac{a}{b^2}$$

- Distribution over positive real numbers $\tau > 0$
- Governed by parameters $a > 0$ (shape), $b > 0$ (scale)

Conjugate Priors

- Posterior $\propto$ prior $\times$ likelihood
- Specification or the prior can be tricky

Conjugate Priors

- Posterior $\propto$ prior $\times$ likelihood
- Specification or the prior can be tricky
- Some priors are (computationally) convenient
- If the posterior and the prior are of the same type (e.g., Beta), the prior is called **conjugate** ⏩ Likelihood is also involved...

Conjugate Priors

- Posterior $\propto$ prior $\times$ likelihood
- Specification or the prior can be tricky
- Some priors are (computationally) convenient
- If the posterior and the prior are of the same type (e.g., Beta), the prior is called **conjugate** ⏩ Likelihood is also involved...
- Examples:

Conjugate prior	Likelihood	Posterior
Beta	Bernoulli	Beta
Gaussian-iGamma	Gaussian	Gaussian-iGamma
Dirichlet	Multinomial	Dirichlet

Example

- Consider a Binomial random variable $x \sim \mathrm{Bin}(m|N, \mu)$ where
$$p(x|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \propto \mu^a (1-\mu)^b$$
for some constants a, b.

Example

- Consider a Binomial random variable $x \sim \text{Bin}(m|N, \mu)$ where

$$p(x|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \propto \mu^a (1-\mu)^b$$

  for some constants a, b.

- We place a Beta-prior on the parameter μ:

$$\text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \propto \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

Example

- Consider a Binomial random variable $x \sim \mathrm{Bin}(m|N, \mu)$ where

$$p(x|\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \propto \mu^a(1-\mu)^b$$

  for some constants a, b.

- We place a Beta-prior on the parameter μ:

$$\mathrm{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \propto \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- If we now observe some outcomes $\boldsymbol{x} = (x_1, \ldots, x_n)$ of a repeated coin-flip experiment with h heads and t tails, we compute the posterior distribution on μ :

Example

- Consider a Binomial random variable $x \sim \text{Bin}(m|N, \mu)$ where

$$p(x|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \propto \mu^a (1-\mu)^b$$

  for some constants a, b.
- We place a Beta-prior on the parameter μ:

$$\text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \propto \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- If we now observe some outcomes $\boldsymbol{x} = (x_1, \ldots, x_n)$ of a repeated coin-flip experiment with h heads and t tails, we compute the posterior distribution on μ :

$$p(\mu|\boldsymbol{x} = h)$$

Example

- Consider a Binomial random variable $x \sim \text{Bin}(m|N, \mu)$ where

$$p(x|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \propto \mu^a (1-\mu)^b$$

  for some constants a, b.

- We place a Beta-prior on the parameter μ:

$$\text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1} \propto \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- If we now observe some outcomes $\boldsymbol{x} = (x_1, \ldots, x_n)$ of a repeated coin-flip experiment with h heads and t tails, we compute the posterior distribution on μ :

$$p(\mu|\boldsymbol{x} = h) \propto {\color{blue}p(\boldsymbol{x}|\mu)}{\color{red}p(\mu|\alpha, \beta)}$$

Example

- Consider a Binomial random variable $x \sim \text{Bin}(m|N,\mu)$ where

$$p(x|\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \propto \mu^a(1-\mu)^b$$

for some constants a, b.

- We place a Beta-prior on the parameter μ:

$$\text{Beta}(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \propto \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- If we now observe some outcomes $\boldsymbol{x} = (x_1, \ldots, x_n)$ of a repeated coin-flip experiment with h heads and t tails, we compute the posterior distribution on μ :

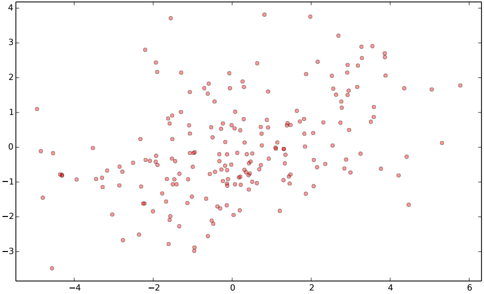
$$p(\mu|\boldsymbol{x}=h) \propto {\color{blue}p(\boldsymbol{x}|\mu)}{\color{red}p(\mu|\alpha,\beta)} = {\color{blue}\mu^h(1-\mu)^t}{\color{red}\mu^{\alpha-1}(1-\mu)^{\beta-1}}$$

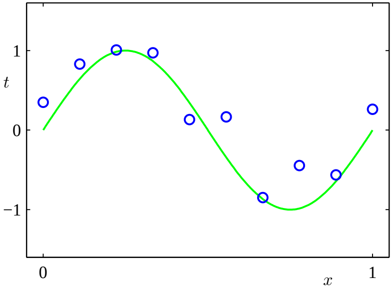
Example

- Consider a Binomial random variable $x \sim \text{Bin}(m|N, \mu)$ where
$$p(x|\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \propto \mu^a(1-\mu)^b$$
for some constants a, b.
- We place a Beta-prior on the parameter μ:
$$\text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1} \propto \mu^{\alpha-1}(1-\mu)^{\beta-1}$$
- If we now observe some outcomes $\boldsymbol{x} = (x_1, \dots, x_n)$ of a repeated coin-flip experiment with h heads and t tails, we compute the posterior distribution on μ :
$$\begin{aligned} p(\mu|\boldsymbol{x} = h) &\propto p(\boldsymbol{x}|\mu)p(\mu|\alpha, \beta) = \mu^h(1-\mu)^t\mu^{\alpha-1}(1-\mu)^{\beta-1} \\ &= \mu^{h+\alpha-1}(1-\mu)^{t+\beta-1} \propto \text{Beta}(h+\alpha, t+\beta) \end{aligned}$$

Parameter Estimation

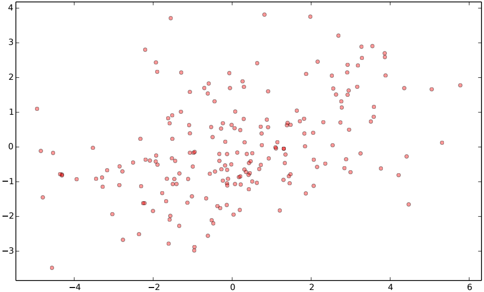
Parameter Estimation

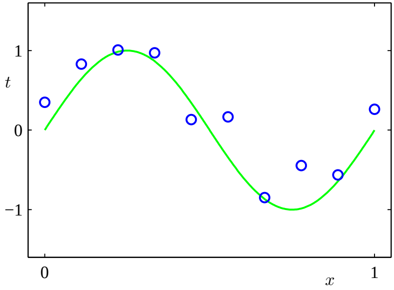




- Given a data set we want to obtain good estimates of the parameters of the model that may have generated the data ⏩ Parameter estimation problem

Parameter Estimation





- Given a data set we want to obtain good estimates of the parameters of the model that may have generated the data ⏩ Parameter estimation problem
- Example: $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^D$ are i.i.d. samples from a Gaussian
  ⏩ Find the mean and covariance of $p(\boldsymbol{x})$

Maximum Likelihood Parameter Estimation (1)

- Maximum likelihood estimation finds a point estimate of the parameters that maximizes the likelihood of the parameters

Maximum Likelihood Parameter Estimation (1)

- Maximum likelihood estimation finds a point estimate of the parameters that maximizes the likelihood of the parameters
- In our Gaussian example, we seek $\boldsymbol{\mu}, \boldsymbol{\Sigma}$:

Maximum Likelihood Parameter Estimation (1)

- Maximum likelihood estimation finds a point estimate of the parameters that maximizes the likelihood of the parameters
- In our Gaussian example, we seek $\boldsymbol{\mu}, \boldsymbol{\Sigma}$:

$$\max p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \overset{i.i.d.}{=} \max \prod_{i=1}^{N} p(\boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Maximum Likelihood Parameter Estimation (1)

- Maximum likelihood estimation finds a point estimate of the parameters that maximizes the likelihood of the parameters
- In our Gaussian example, we seek $\boldsymbol{\mu}, \boldsymbol{\Sigma}$:

$$\max p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \stackrel{i.i.d.}{=} \max \prod_{i=1}^{N} p(\boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$= \max \sum_{i=1}^{N} \log p(\boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Maximum Likelihood Parameter Estimation (1)

- Maximum likelihood estimation finds a point estimate of the parameters that maximizes the likelihood of the parameters
- In our Gaussian example, we seek $\boldsymbol{\mu}, \boldsymbol{\Sigma}$:

$$
\begin{aligned}
\max p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}) &\overset{i.i.d.}{=} \max \prod_{i=1}^{N} p(\boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \max \sum_{i=1}^{N} \log p(\boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \max -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})
\end{aligned}
$$

Maximum Likelihood Parameter Estimation (2)

$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\mu}} -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\end{aligned}$$

Maximum Likelihood Parameter Estimation (2)

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\mu}} -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\
&= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \\
\boldsymbol{\Sigma}^*_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\Sigma}} -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})^\top
\end{aligned}
$$

Maximum Likelihood Parameter Estimation (2)

$$\boldsymbol{\mu}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\mu}} -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}^*_{\mathrm{ML}} = \arg\max_{\boldsymbol{\Sigma}} -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})^\top$$

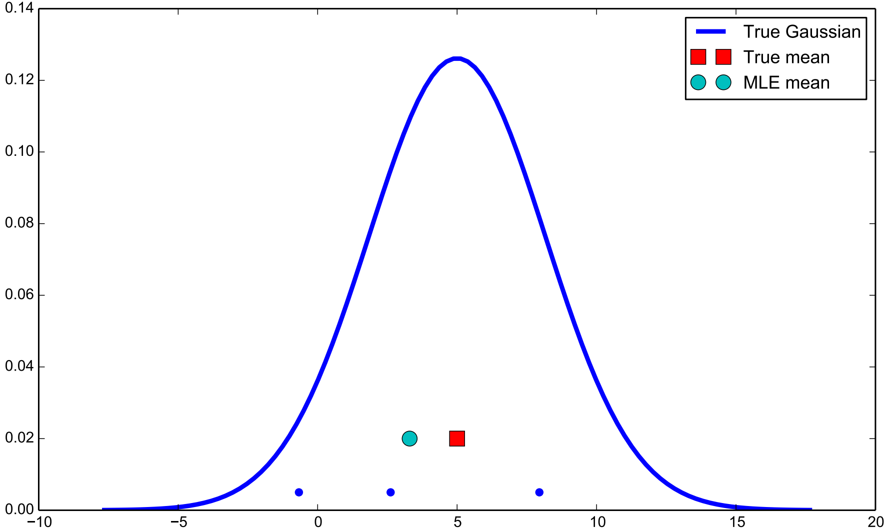
ML estimate $\boldsymbol{\Sigma}^*_{\mathrm{ML}}$ is biased, but we can get an unbiased estimate as

$$\boldsymbol{\Sigma}^* = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})^\top$$

MLE: Properties

- Asymptotic consistency: The MLE converges to the true value in the limit of infinitely many observations, plus a random error that is approximately normal
- The size of the sample necessary to achieve these properties can be quite large
- The error's variance decays in $1/N$ where N is the number of data points
- Especially, in the "small" data regime, MLE can lead to overfitting

Example: MLE in the Small-Data Regime



Maximum A Posteriori Estimation

- Instead of maximizing the likelihood, we can seek parameters that maximize the posterior distribution of the parameters

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

Maximum A Posteriori Estimation

- Instead of maximizing the likelihood, we can seek parameters that maximize the posterior distribution of the parameters

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

- MLE with an additional regularizer that comes from the prior

  ▸▸ **MAP estimator**

Maximum A Posteriori Estimation

- Instead of maximizing the likelihood, we can seek parameters that maximize the posterior distribution of the parameters

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

- MLE with an additional regularizer that comes from the prior

  ▶▶ **MAP estimator**
- Example:
  - Estimate the mean μ of a 1D Gaussian with known variance σ^2 after having observed N data points x_i.

Maximum A Posteriori Estimation

- Instead of maximizing the likelihood, we can seek parameters that maximize the posterior distribution of the parameters

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{x}|\boldsymbol{\theta})$$

- MLE with an additional regularizer that comes from the prior

  ⏩ **MAP estimator**
- Example:
  - Estimate the mean μ of a 1D Gaussian with known variance σ^2 after having observed N data points x_i.
  - Gaussian prior $p(\mu) = \mathcal{N}\left(\mu \,|\, m,\, s^2\right)$ on mean yields

$$\mu_{\text{MAP}} = \frac{Ns^2}{Ns^2 + \sigma^2}\mu_{\text{ML}} + \frac{\sigma^2}{Ns^2 + \sigma^2}m$$

Interpreting the Result

$$\mu_{\text{MAP}} = \frac{Ns^2}{Ns^2 + \sigma^2}\mu_{\text{ML}} + \frac{\sigma^2}{Ns^2 + \sigma^2}m$$

- Linear interpolation between the prior mean and the sample mean (ML estimate), weighted by their respective covariances

Interpreting the Result

$$\mu_{\text{MAP}} = \frac{Ns^2}{Ns^2 + \sigma^2}\mu_{\text{ML}} + \frac{\sigma^2}{Ns^2 + \sigma^2}m$$

- Linear interpolation between the prior mean and the sample mean (ML estimate), weighted by their respective covariances
- The more data we have seen (N), the more we believe the sample mean

Interpreting the Result

$$\mu_{\text{MAP}} = \frac{Ns^2}{Ns^2 + \sigma^2}\mu_{\text{ML}} + \frac{\sigma^2}{Ns^2 + \sigma^2}m$$

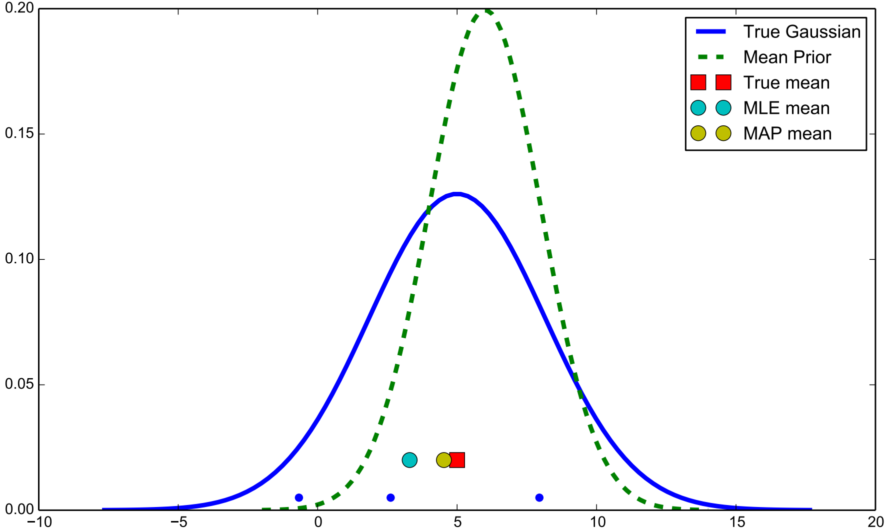
- Linear interpolation between the prior mean and the sample mean (ML estimate), weighted by their respective covariances
- The more data we have seen (N), the more we believe the sample mean
- The higher the variance s^2 of the prior, the more we believe the sample mean; MAP $\xrightarrow{s^2 \to \infty}$ MLE

Interpreting the Result

$$\mu_{\text{MAP}} = \frac{Ns^2}{Ns^2 + \sigma^2} \mu_{\text{ML}} + \frac{\sigma^2}{Ns^2 + \sigma^2} m$$

- Linear interpolation between the prior mean and the sample mean (ML estimate), weighted by their respective covariances
- The more data we have seen (N), the more we believe the sample mean
- The higher the variance s^2 of the prior, the more we believe the sample mean; MAP $\xrightarrow{s^2 \to \infty}$ MLE
- The higher the variance σ^2 of the data, the less we believe the sample mean

Example



True Gaussian
Mean Prior
True mean
MLE mean
MAP mean
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
10
15
20

Bayesian Inference (Marginalization)

An even better idea than MAP estimation:

- Instead of estimating a parameter, integrate it out (according to the posterior) when making predictions

$$p(x|\mathcal{D}) = \int p(x|\theta)p(\theta|\mathcal{D})d\theta$$

where $p(\theta|\mathcal{D})$ is the parameter posterior

Bayesian Inference (Marginalization)

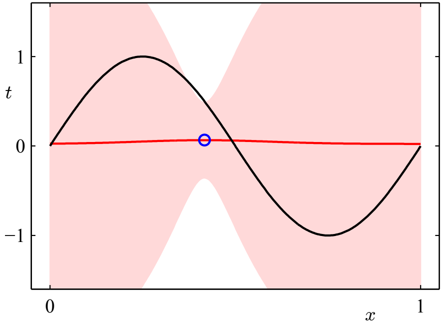
An even better idea than MAP estimation:

- Instead of estimating a parameter, integrate it out (according to the posterior) when making predictions

$$p(x|\mathcal{D}) = \int p(x|\theta)p(\theta|\mathcal{D})d\theta$$

  where $p(\theta|\mathcal{D})$ is the parameter posterior
- This integral is often tricky to solve
  ⏩ Choose appropriate distributions (e.g., conjugate distributions) or solve approximately (e.g., sampling or variational inference)
- Works well (even in the small-data regime) and is robust to overfitting

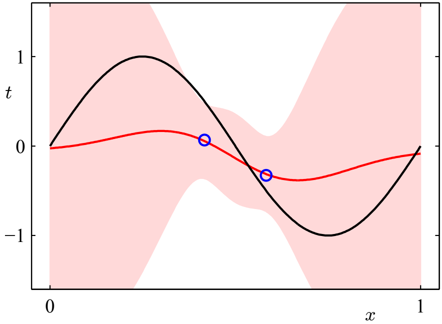
Example: Linear Regression



Adapted from PRML (Bishop, 2006)

- Blue: data
- Black: True function (unknown)
- Red: Posterior mean (MAP estimate)
- Red-shaded: 95% confidence area of the prediction

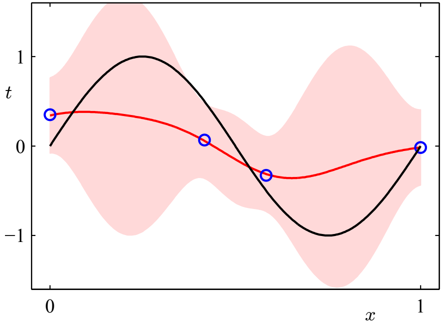
Example: Linear Regression



Adapted from PRML (Bishop, 2006)

- Blue: data
- Black: True function (unknown)
- Red: Posterior mean (MAP estimate)
- Red-shaded: 95% confidence area of the prediction

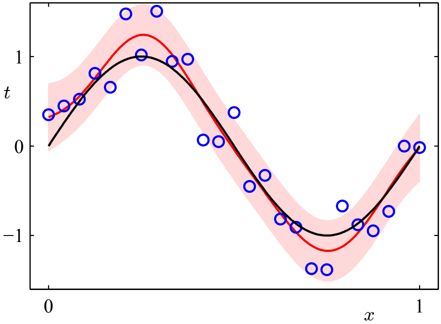
Example: Linear Regression



Adapted from PRML (Bishop, 2006)

- Blue: data
- Black: True function (unknown)
- Red: Posterior mean (MAP estimate)
- Red-shaded: 95% confidence area of the prediction

Example: Linear Regression



Adapted from PRML (Bishop, 2006)

- Blue: data
- Black: True function (unknown)
- Red: Posterior mean (MAP estimate)
- Red-shaded: 95% confidence area of the prediction

References I

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, 2006.