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:

\[ p(x) = \int p(x, y) dy \]  \quad \text{Sum rule/Marginalization property}

\[ p(x, y) = p(y|x) p(x) \]  \quad \text{Product rule}
Key Concepts in Probability Theory

Two fundamental rules:

\[ p(x) = \int p(x, y) dy \quad \text{Sum rule/Marginalization property} \]

\[ p(x, y) = p(y|x)p(x) \quad \text{Product rule} \]

Bayes’ Theorem (Probabilistic Inverse)

\[ p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \quad x: \text{hypothesis, } y: \text{measurement} \]
Key Concepts in Probability Theory

Two fundamental rules:

\[ p(x) = \int p(x, y) \, dy \]  
\[ p(x, y) = p(y|x) \, p(x) \]

- Sum rule/Marginalization property
- Product rule

Bayes’ Theorem (Probabilistic Inverse)

\[ p(x|y) = \frac{p(y|x) \, p(x)}{p(y)}, \quad x : \text{hypothesis}, \quad y : \text{measurement} \]

- 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**

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

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

\[
\text{Cov}[x, y] = \mathbb{E}_{x,y}[xy^\top] - \mathbb{E}_x[x]\mathbb{E}_y[y]^\top
\]
Mean and (Co)Variance

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

**Definition**

\[
\begin{align*}
\mathbb{E}_x[x] &= \int x p(x) \, dx =: \mu \\
\mathbb{V}_x[x] &= \mathbb{E}_x[(x - \mu)(x - \mu)^\top] = \mathbb{E}_x[xx^\top] - \mathbb{E}_x[x]\mathbb{E}_x[x]^\top =: \Sigma \\
\text{Cov}[x, y] &= \mathbb{E}_{x,y}[xy^\top] - \mathbb{E}_x[x]\mathbb{E}_y[y]^\top
\end{align*}
\]

**Linear/Affine Transformations:**

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

\[
\mathbb{E}[y] = \\
\mathbb{V}[y] =
\]
Mean and (Co)Variance

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

**Definition**

\[
E_x[x] = \int x p(x) \, dx =: \mu \\
V_x[x] = E_x[(x - \mu)(x - \mu)^\top] = E_x[xx^\top] - E_x[x]E_x[x]^\top =: \Sigma \\
Cov[x, y] = E_{x,y}[xy^\top] - E_x[x]E_y[y]^\top
\]

**Linear/Affine Transformations:**

\[
y = Ax + b, \quad \text{where} \quad E_x[x] = \mu, \ V_x[x] = \Sigma \\
E[y] = E_x[Ax + b] = A E_x[x] + b = A \mu + b \\
V[y] = \]

Mean and (Co)Variance

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

**Definition**

\[
\mathbb{E}_x[x] = \int x p(x) \, dx =: \mu
\]

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

\[
\text{Cov}[x, y] = \mathbb{E}_{x,y}[xy^\top] - \mathbb{E}[x] \mathbb{E}[y][y]^\top
\]

**Linear/Affine Transformations:**

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

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

\[
\mathbb{V}[y] = \mathbb{V}_x[Ax + b] = \mathbb{V}_x[Ax] = A \mathbb{V}_x[x] A^\top = A \Sigma A^\top
\]
Mean and (Co)Variance

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

**Definition**

\[
\mathbb{E}_x[x] = \int x p(x) \, dx =: \mu
\]

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

\[
\text{Cov}[x, y] = \mathbb{E}_{x,y}[xy^\top] - \mathbb{E}_x[x]\mathbb{E}_y[y]^\top
\]

**Linear/Affine Transformations:**

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

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

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

If \(x, y\) independent: \(\mathbb{V}_{x,y}[x + y] = \mathbb{V}_x[x] + \mathbb{V}_y[y]\)
Basic Probability Distributions
The Gaussian Distribution

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

- **Mean vector** \( \mu \) ➞ Average of the data
- **Covariance matrix** \( \Sigma \) ➞ Spread of the data
The Gaussian Distribution

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

- **Mean vector** $\mu$ ➤ Average of the data
- **Covariance matrix** $\Sigma$ ➤ Spread of the data

![Graph of a Gaussian distribution](image)

![Graph of a multivariate Gaussian distribution](image)
The Gaussian Distribution

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

- **Mean vector** \( \mu \) ➤ Average of the data
- **Covariance matrix** \( \Sigma \) ➤ Spread of the data

Probability Distributions and Parameter Estimation

IDAPI, Lecture 11 February 10, 2016
Conditional

\begin{equation}
p(x, y) = \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)
\end{equation}
Conditional

\[ p(x, y) = \mathcal{N} \left( \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right) \]
Conditional

\[ p(x, y) = \mathcal{N}\left( \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right) \]

\[ p(x|y) = \mathcal{N}\left( \mu_{x|y}, \Sigma_{x|y} \right) \]

\[ \mu_{x|y} = \mu_x + \Sigma_{xy} \Sigma_{yy}^{-1} (y - \mu_y) \]

\[ \Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \]

Conditional \( p(x|y) \) is also Gaussian

\[ \text{Computationally convenient} \]
Joint $p(x, y) = \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$

Marginal distribution:

$$p(x) = \int p(x, y) d y = \mathcal{N}\left(\mu_x, \Sigma_{xx}\right)$$
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 $x \in \{0, 1\}$.

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

$$\mathbb{E}[x] = \mu$$

$$\text{Var}[x] = \mu(1 - \mu)$$
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 $x \in \{0, 1\}$.

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

$$\mathbb{E}[x] = \mu$$

$$\mathbb{V}[x] = \mu (1 - \mu)$$

- 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 $\alpha > 0, \beta > 0$

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

$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta}, \quad \mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$
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 $p(x = 1) = \mu \in [0, 1]$

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

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

\[
p(x = 1) = \mu \in [0, 1]
\]

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

\[
E[m] = N \mu, \quad V[m] = N \mu(1 - \mu)
\]

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

<table>
<thead>
<tr>
<th>Conjugate prior</th>
<th>Likelihood</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta</td>
<td>Bernoulli</td>
<td>Beta</td>
</tr>
<tr>
<td>Gaussian-iGamma</td>
<td>Gaussian</td>
<td>Gaussian-iGamma</td>
</tr>
<tr>
<td>Dirichlet</td>
<td>Multinomial</td>
<td>Dirichlet</td>
</tr>
</tbody>
</table>
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 $a, b$. 
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 $\mu$:

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

- Consider a Binomial random variable \( x \sim \text{Bin}(m|N, \mu) \) 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 \( \mu \):
  
  \[
  \text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1} \propto \mu^{\alpha-1} (1-\mu)^{\beta-1}
  \]
- If we now observe some outcomes \( x = (x_1, \ldots, x_n) \) of a repeated coin-flip experiment with \( h \) heads and \( t \) tails, we compute the posterior distribution on \( \mu \):
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 $\mu$:
  \[
  \text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \propto \mu^{\alpha-1} (1 - \mu)^{\beta-1}
  \]

- If we now observe some outcomes $x = (x_1, \ldots, x_n)$ of a repeated coin-flip experiment with $h$ heads and $t$ tails, we compute the posterior distribution on $\mu$:
  \[ p(\mu|x = h) \]
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 $\mu$:
  
  $$\text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \propto \mu^{\alpha-1} (1 - \mu)^{\beta-1}$$

- If we now observe some outcomes $x = (x_1, \ldots, x_n)$ of a repeated coin-flip experiment with $h$ heads and $t$ tails, we compute the posterior distribution on $\mu$:
  
  $$p(\mu|x = h) \propto p(x|\mu)p(\mu|\alpha, \beta)$$
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 \( \mu \):
  
  \[
  \text{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \propto \mu^{\alpha-1} (1 - \mu)^{\beta-1}
  \]

- If we now observe some outcomes \( x = (x_1, \ldots, x_n) \) of a repeated coin-flip experiment with \( h \) heads and \( t \) tails, we compute the posterior distribution on \( \mu \):
  
  \[
p(\mu|x = h) \propto p(x|\mu)p(\mu|\alpha, \beta) = \mu^h(1 - \mu)^t \mu^{\alpha-1} (1 - \mu)^{\beta-1}
  \]
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 $\mu$:

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

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

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

$$= \mu^{h+\alpha-1}(1-\mu)^{t+\beta-1} \propto \text{Beta}(h + \alpha, t + \beta)$$
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**
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: $x_1, \ldots, x_N \in \mathbb{R}^D$ are i.i.d. samples from a Gaussian. Find the mean and covariance of $p(x)$. 
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 $\mu, \Sigma$: 
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 $\mu, \Sigma$:

$$\max p(x_1, \ldots, x_n | \mu, \Sigma) \overset{i.i.d.}{=} \max \prod_{i=1}^N p(x_i | \mu, \Sigma)$$
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 $\mu, \Sigma$:

$$\max \ p(x_1, \ldots, x_n|\mu, \Sigma) \overset{\text{i.i.d.}}{=} \max \prod_{i=1}^{N} p(x_i|\mu, \Sigma)$$

$$= \max \sum_{i=1}^{N} \log p(x_i|\mu, \Sigma)$$
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 $\mu, \Sigma$:

$$\max p(x_1, \ldots, x_n | \mu, \Sigma) \overset{i.i.d.}{=} \max \prod_{i=1}^{N} p(x_i | \mu, \Sigma)$$

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

$$= \max -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^{\top} \Sigma^{-1} (x_i - \mu)$$
Maximum Likelihood Parameter Estimation (2)

\[ \mu_{ML} = \arg \max_{\mu} - \frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \]

\[ = \frac{1}{N} \sum_{i=1}^{N} x_i \]
Maximum Likelihood Parameter Estimation (2)

\[
\mu_{\text{ML}} = \arg \max_{\mu} -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu)
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} x_i
\]

\[
\Sigma^*_{\text{ML}} = \arg \max_{\Sigma} -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu)
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{\text{ML}})(x_i - \mu_{\text{ML}})^\top
\]
Maximum Likelihood Parameter Estimation (2)

\[ \mu_{ML} = \arg \max_\mu \left( -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \right) \]

\[ = \frac{1}{N} \sum_{i=1}^{N} x_i \]

\[ \Sigma^*_{ML} = \arg \max_\Sigma \left( -\frac{ND}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^\top \Sigma^{-1} (x_i - \mu) \right) \]

\[ = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{ML}) (x_i - \mu_{ML})^\top \]

ML estimate \( \Sigma^*_{ML} \) is biased, but we can get an unbiased estimate as

\[ \Sigma^* = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_{ML}) (x_i - \mu_{ML})^\top \]
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

\[
\theta^* = \arg \max_{\theta} p(\theta|x) = \arg \max_{\theta} \log p(\theta) + \log p(x|\theta)
\]
Maximum A Posteriori Estimation

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

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

- 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

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

- MLE with an additional regularizer that comes from the prior

  - MAP estimator

- Example:
  - Estimate the mean \( \mu \) of a 1D Gaussian with known variance \( \sigma^2 \) 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

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

- MLE with an additional regularizer that comes from the prior
  - **MAP estimator**

- Example:
  - Estimate the mean \( \mu \) of a 1D Gaussian with known variance \( \sigma^2 \) after having observed \( N \) data points \( x_i \).
  - Gaussian prior \( p(\mu) = \mathcal{N}(\mu | m, s^2) \) 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

\[ \mu_{\text{MAP}} = \frac{N s^2}{N s^2 + \sigma^2} \mu_{\text{ML}} + \frac{\sigma^2}{N s^2 + \sigma^2} m \]

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

\[ \mu_{MAP} = \frac{N s^2}{Ns^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{Ns^2 + \sigma^2} m \]

- 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

\[ \mu_{\text{MAP}} = \frac{N s^2}{N s^2 + \sigma^2} \mu_{\text{ML}} + \frac{\sigma^2}{N s^2 + \sigma^2} m \]

- Linear interpolation between the \textcolor{blue}{prior mean} and the \textcolor{red}{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; \(\text{MAP} \xrightarrow{s^2 \to \infty} \text{MLE}\)
Interpreting the Result

\[ \mu_{\text{MAP}} = \frac{N s^2}{N s^2 + \sigma^2} \mu_{\text{ML}} + \frac{\sigma^2}{N s^2 + \sigma^2} m \]

- 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 \( s^2 \xrightarrow{\infty} \) MLE
- The higher the variance \( \sigma^2 \) of the data, the less we believe the sample mean
Example
Bayesian Inference (Marginalization)

An even better idea than MAP estimation:

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

\[
p(x|D) = \int p(x|\theta) p(\theta|D) d\theta
\]

where \( p(\theta|D) \) 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

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

Adapted from PRML (Bishop, 2006)
Example: Linear Regression

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

Adapted from PRML (Bishop, 2006)
Example: Linear Regression

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

Adapted from PRML (Bishop, 2006)
Example: Linear Regression

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

Adapted from PRML (Bishop, 2006)
References I