Outline

- Maximum likelihood estimation
- Bayesian inference basics
Maximum Likelihood Estimation (MLE)
Goal: estimate parameters $\theta$ from observed data $\{x_1, \cdots, x_N\}$

Main idea: We should choose parameters that assign high probability to the observed data:

$$\hat{\theta} = \arg\max L(\theta; x_1, \cdots, x_N)$$
Three steps for computing MLE

1. Write down the likelihood objective:

   \[ L(\theta; x_1, \cdots, x_N) = \prod_{i=1}^{N} L(\theta; x_i) \]

2. Transform to log likelihood:

   \[ l(\theta; x_1, \cdots, x_N) = \sum_{i=1}^{N} \log L(\theta; x_i) \]

3. Compute the critical point:

   \[ \frac{\partial l}{\partial \theta} = 0 \]
**Example 1 - categorial distribution**

\( X \) is a discrete random variable with the following probability mass function (\( 0 \leq \theta \leq 1 \) is an unknown parameter):

<table>
<thead>
<tr>
<th>( X )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X) )</td>
<td>( 2\theta/3 )</td>
<td>( \theta/3 )</td>
<td>( 2(1 - \theta)/3 )</td>
<td>( (1 - \theta)/3 )</td>
</tr>
</tbody>
</table>

- The following 10 independent observations were taken from \( X \): \{3, 0, 2, 1, 3, 2, 1, 0, 2, 1\}.
- What is the MLE for \( \theta \)?
Step 1: Likelihood objective

\[ L(\theta) = P(X = 3)P(X = 0)P(X = 2)P(X = 1)P(X = 3) \]
\[ \times P(X = 2)P(X = 1)P(X = 0)P(X = 2)P(X = 1) \]
\[ = \left( \frac{2\theta}{3} \right)^2 \left( \frac{\theta}{3} \right)^3 \left( \frac{2(1 - \theta)}{3} \right)^3 \left( \frac{(1 - \theta)}{3} \right)^2 \]
Step 2: Log likelihood

\[ l(\theta) = \log L(\theta) \]

\[ = 2(\log \frac{2}{3} + \log \theta) + 3(\log \frac{1}{3} + \log \theta) \]

\[ + 3(\log \frac{2}{3} + \log (1 - \theta)) + 2(\log \frac{2}{3} + \log (1 - \theta)) \]

\[ = C + 5(\log \theta + \log (1 - \theta)) \]
Step 3: critical points

\[ \frac{\partial l}{\partial \theta} = 0 \]

\[ \rightarrow 5\left(\frac{1}{\theta} - \frac{1}{1 - \theta}\right) = 0 \]

\[ \rightarrow \hat{\theta} = 0.5 \]
Example 2 - Poisson distribution

- **X** is a discrete random variable following the Poisson distribution:
  \[ P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \]

- Suppose we observe \( N \) samples of **X**: \( \{x_1, \cdots, x_N\} \)

- What is the MLE for \( \lambda \)?
Three steps

1. Likelihood objective:

\[ L(\lambda) = \prod_{i=1}^{N} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \]

2. Log likelihood:

\[ l(\lambda) = -N\lambda + \log \lambda \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \log(x_i!) \]

3. Critical point:

\[ \frac{\partial l}{\partial \lambda} = 0 \rightarrow -N + \frac{1}{\lambda} \sum_{i=1}^{N} x_i \rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{N} x_i}{N} \]
Suppose that $X_1, \cdots, X_n$ form a random sample from a uniform distribution on the interval $(0, \theta)$, where of the parameter $\theta > 0$ but is unknown. Please find MLE of $\theta$. 
Bayesian Inference Basics
Bayesian Philosophy

- Bayesian interprets probability as degrees of beliefs.
- Bayesian treats parameters as random variables.
- Bayesian learning is updating our beliefs (probability distribution) based on observations.
MLE is the standard frequentist inference method.

Bayesian and frequentist are the two main approaches in statistical machine learning. Some of their ideological differences can be summarized as:

<table>
<thead>
<tr>
<th></th>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability is</td>
<td>relative frequency</td>
<td>degree of beliefs</td>
</tr>
<tr>
<td>Parameter $\theta$ is</td>
<td>unknown constant</td>
<td>random variable</td>
</tr>
</tbody>
</table>

---

1Han Liu and Larry Wasserman, Statistical Machine Learning, 2014
The Bayesian approach to machine learning

1. We define a model that expresses qualitative aspects of our knowledge (e.g., forms of distributions, independence assumptions). The model will have some unknown parameters.

2. We specify a prior probability distribution for these unknown parameters that expresses our beliefs about which values are more or less likely, before seeing the data.

3. We gather data.

4. We compute the posterior probability distribution for the parameters, given the observed data.

5. We use this posterior distribution to draw scientific conclusions and make predictions.

---

2 Radford M. Neal, Bayesian Methods for Machine Learning, NIPS 2004 tutorial
Computing the posterior

The posterior distribution is computed by the Bayes’ rule:

\[ P(\text{parameter}|\text{data}) = \frac{P(\text{parameter})P(\text{data}|\text{parameter})}{P(\text{data})} \]

The denominator is just the required normalizing constant. So as a proportionality, we can write:

posterior \(\propto\) prior \(\times\) likelihood
Exercise

- Suppose you have a Beta(4, 4) prior distribution on the probability $\theta$ that a coin will yield a ‘head’ when spun in a specified manner.

- The coin is independently spun ten times, and ‘heads’ appear fewer than 3 times. You are not told how many heads were seen, only that the number is less than 3.

- Calculate your exact posterior density (up to a proportionality constant) for $\theta$ and sketch it.
Questions?