Principle of Maximum Likelihood

- Given parameters $\theta$ and data $X$
- The function $f(X \mid \theta)$ is the probability of observing data $X$ given parameter $\theta$. (Both $X$ and $\theta$ can be multi-dimensional.)
- Keeping $\theta$ fixed, and treating $f$ as a function of $X$, the total probability is one.
Principle of Maximum Likelihood

- The function \( L(\theta) = f(X \mid \theta) \) with \( X \) fixed and \( \theta \) unknown is called the *likelihood function*.
- The *principle of maximum likelihood* is to estimate \( \theta \) with the value \( \hat{\theta} \) that maximizes \( L(\theta) \).
- In practice, it is common to maximize the log-likelihood, \( \ell(\theta) = \ln L(\theta) \).
- This is because \( X \) often takes the form of an independent sample so that
  \[
  L(X) = \prod_{i=1}^{n} f(X_i \mid \theta), \quad \ell(\theta) = \sum_{i=1}^{n} \ln f(X_i \mid \theta)
  \]

Coin-tossing Example

- A coin has a probability \( \theta \) of being a head.
- Consider tossing the coin 100 times. The probability of each single sequence with exactly \( x \) heads is \( f(x \mid \theta) = p^x (1-p)^{100-x} \).
- Say we observe the sequence
  \[
  HHTHTHHT \ldots TTH
  \]
  where heads appear 57 times.
- The maximum likelihood estimate is the value \( \hat{\theta} \) that maximizes the function
  \[
  L(\theta) = \theta^{57} (1-\theta)^{43},
  \]
  or, equivalently that maximizes
  \[
  \ell(\theta) = 57(\ln \theta) + 43(\ln(1 - \theta)).
  \]
  Simple calculus and common sense lead to the estimate \( \hat{\theta} = 0.57 \).
Maximum-likelihood edge lengths

- For the Jukes-Cantor model, a pair of sequences have $x$ sites with observed differences and $n - x$ sites with the same base.
- The probability of any given sequence pair is
  \[
  L(d) = \left( \frac{1}{4} \right)^n \times \left( \frac{1}{4} - \frac{1}{4} e^{-\frac{4}{3}d} \right)^x \times \left( \frac{1}{4} + \frac{3}{4} e^{-\frac{4}{3}d} \right)^{n-x}
  \]
  which has the form
  \[
  L(\theta) = C \times \theta^x (1 - 3\theta)^{n-x}
  \]
  where
  \[
  \theta = \frac{1}{4} - \frac{1}{4} e^{-\frac{4}{3}d}.
  \]
- Solving the calculus problem yields $\hat{\theta} = \frac{x}{3n}$.
- Plugging in and solving for $d$ gives
  \[
  \hat{d} = -\frac{3}{4} \ln \left( 1 - \frac{4x}{3n} \right)
  \]

Computing Likelihood on a Tree

![Tree Diagram](image)
Transition Probabilities

\[ P(0.1) = \begin{bmatrix} 0.90 & 0.04 & 0.04 & 0.03 \\ 0.03 & 0.91 & 0.04 & 0.03 \\ 0.03 & 0.04 & 0.91 & 0.03 \\ 0.03 & 0.04 & 0.04 & 0.90 \end{bmatrix} \quad P(0.2) = \begin{bmatrix} 0.81 & 0.07 & 0.07 & 0.05 \\ 0.05 & 0.83 & 0.07 & 0.05 \\ 0.05 & 0.07 & 0.83 & 0.05 \\ 0.05 & 0.07 & 0.07 & 0.81 \end{bmatrix} \]

\[ P(0.4) = \begin{bmatrix} 0.67 & 0.13 & 0.13 & 0.08 \\ 0.08 & 0.71 & 0.13 & 0.08 \\ 0.08 & 0.13 & 0.71 & 0.08 \\ 0.08 & 0.13 & 0.13 & 0.67 \end{bmatrix} \]

Model Selection

12 \textit{rbcL} genes from 12 plant species

<table>
<thead>
<tr>
<th>Model</th>
<th>p</th>
<th>( \ell )</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC69</td>
<td>21</td>
<td>(-6262.01)</td>
</tr>
<tr>
<td>K80</td>
<td>22</td>
<td>(-6113.86)</td>
</tr>
<tr>
<td>HKY85</td>
<td>25</td>
<td>(-6101.76)</td>
</tr>
<tr>
<td>HKY85 + ( \Gamma_5 )</td>
<td>26</td>
<td>(-5764.26)</td>
</tr>
<tr>
<td>HKY85 + C</td>
<td>35</td>
<td>(-5624.70)</td>
</tr>
</tbody>
</table>

- The AIC criterion is to select the model with the lowest AIC score, which is
  \[ \text{AIC} = -2 \ln(\text{likelihood}) + 2 \times (\# \text{ of parameters}) \]
- AIC balances the competing goals to fit the data well (likelihood high) and keep the model simple (few parameters).
- For this data, the HKY85+C model is the best among those compared; using 9 more parameters yielded an improvement in loglikelihood of over 139, which lowered the AIC by about 130.
The Bootstrap: A brief history

The bootstrap was introduced to the world by Brad Efron, chair of the Department of Statistics at Stanford University, in 1979.

The bootstrap is one of the most widely used new method in statistics that was invented within the past 50 years.

In a special issue of Statistical Science that celebrates the 25th anniversary of the bootstrap, Brad Efron uses its application to phylogenetics as one of a small number of examples to illustrate its use and importance.

The General Bootstrap Framework

- We have a sample \( x_1, \ldots, x_n \) drawn from a distribution \( F \) from which we wish to estimate a parameter \( \theta \) using a statistic \( \hat{\theta} = T(x_1, \ldots, x_n) \). (We might think of \( \theta \) as being the median of the distribution, for example, and \( \hat{\theta} = T(x_1, \ldots, x_n) \) as the sample median.)

- If we wanted to compute the standard error of the estimate, we would ideally compute the standard deviation of \( T(X_1, \ldots, X_n) \) where \( X_i \sim \text{iid} \ F \).

- We could estimate this to any desired degree of accuracy by generating a large enough number (say \( B \)) of random samples \( X_1, \ldots, X_n \), computing \( \hat{\theta}_i = T(X_1, \ldots, X_n) \) for the \( i \)th such sample, and then computing the standard deviation of these estimates.

\[
\sqrt{\frac{\sum_{i=1}^{B} (\hat{\theta}_i - \theta)^2}{B}}
\]
The Key Idea

- Unfortunately, we cannot take multiple samples from $F$.
- However, our original sample $x_1, \ldots, x_n$ is an estimate of the distribution $F$.
- Instead of taking samples from $F$, we could sample from the estimated distribution $\hat{F}$ by sampling from our original sample with replacement.

The Procedure

- We sample $n$ values $x_1^*, \ldots, x_n^*$ with replacement from $x_1, \ldots, x_n$.
- It is very likely that some of the original $x$ values will be sampled multiple times and others will not be sampled at all.
- For each sample, compute the estimate of $\theta$ using the original statistic.
- The $i$th estimate is $\hat{\theta}_i^* = T(x_1^*, \ldots, x_n^*)$.
- Repeat this $B$ times and compute the standard deviation of the bootstrap estimates around the estimate from the original sample.

$$\sqrt{\frac{\sum_{i=1}^{B} (\hat{\theta}_i^* - \hat{\theta})^2}{B}}$$
Why it works

- If the sampling distribution of the bootstrap sample estimate $\hat{\theta}^*$ around the estimate $\hat{\theta}$ is similar to the sampling distribution of the estimate $\hat{\theta}$ around the true value $\theta$, then the bootstrap standard error will be a good estimate of the real standard error.
- The bootstrap can be used to estimate bias, variance, for confidence intervals, and for hypothesis testing in many situations.
- It does depend critically on the assumption of independence of the original sample.

Consensus Trees

- A **strict consensus tree** shows only those clades that appear in every sampled tree.
- A **majority rule consensus tree** shows all clades that appear in more than half the sample of trees.
  
  (Notice that two clades that each appear in more than half the sampled trees must appear in at least one tree together, implying that they are compatible with one another.)
- A **priority consensus tree** adds clades to the majority rule consensus tree in order of decreasing frequency in the sample provided that these clades do not conflict with a clade with higher frequency.
Dynamic Exploration of Tree Samples

- Show off Mark Derthick’s **Summary Tree Explorer**.
- Software is free and available at [http://cityscape.inf.cs.cmu.edu/phylogeny/](http://cityscape.inf.cs.cmu.edu/phylogeny/).
Interpretation of Bootstrap Proportions

What does a bootstrap proportion mean? Let me count the ways.

- **Confidence** that the clade is in the true tree.
- Bayesian posterior probability that the clade is in the true tree.
- One minus p-value for a formal hypothesis test that the clade is in the true tree.
- Rough measure of method robustness.
- Measure of repeatability of the inferences for the method at hand.
- Others?