1 Reminder

We define the alignment score of two DNA strands as \( \sum_{i=1}^{n} \sigma(s_i, t_i) \) such that if we choose \( \sigma \) wisely we get:

\[
\sum_{i=1}^{n} \sigma(s_i, t_i) = \sum_{i=1}^{n} \log \frac{p_1(s_i, t_i)}{p_0(s_i) p_0(t_i)} = LLR
\]

Where \( H_0 \) describes independent sequences with base distribution \( p_0 \), and \( H_1 \) is the hypothesis the sequences are derived from a common ancestor with joint base probability \( p_1 \).

2 Parameter Estimation

2.1 Pins Dropping Example

We start with the following example: we spread a box of pins on the floor. Each pin can either fall on his curved part, the Head, or on the pointy part, the Tail. We then define \( \{H, T\} \). We also have the given series of observations: \( D = \{\alpha_1, \alpha_2, \ldots\} \) with the distribution function \( P_\theta : \sum \to \mathbb{R}^+ \), \( P_\theta(H) = \theta \), \( P_\theta(T) = 1 - \theta \).

The likelihood is defined as follows:

\[
L(\theta : D) = P(D|\theta)
\]

In our case we have : \( L(\theta : D) = P(D|\theta) = \prod_{i=1}^{n} P_\theta(\alpha_i) \)

We want to estimate \( \theta \) according to this kind of series of events.

For example: \( D = \{H, H, T, H\} \implies L(\theta : D) = \theta^3(1 - \theta) \)

2.2 Variables

Define the variables \( N_H D \) and \( N_T D \) that stands for how many times we have H and T, respectively: \( N_H D = \sum_{i=1}^{n} 1(\alpha_i = H) \); \( N_T D = \sum_{i=1}^{n} 1(\alpha_i = T) \)

\[
\implies L(D) = \theta^{N_H(D)}(1 - \theta)^{N_T(D)}
\]
3 **MLE - Maximum Likelihood Estimators**

In this section we present the frequentists’ way - no prior knowledge. We will estimate $\theta$ according to the observed data only. We are looking for the $\theta$ that maximizes the Likelihood.

$$\hat{\theta}_{ML} = \arg\max L(\theta : D)$$

$$L(\theta : D) = \theta^{NH}(1 - \theta)^{NT}$$

![Figure 1: Likelihood as a function of $\theta$](image)

Then the LogLikelihood is - $LL = NH \log(\theta) + NT \log(1 - \theta)$.

An important notice is that $\hat{\theta}_{ML}$ also achieves maximum for the LL, and it’s usually easier to get the derivative of the LL function (because log in monotonic increasing). By comparing the derivative to 0 we get the of $\theta$ that maximizes the likelihood.

$$\frac{\partial LL}{\partial \theta} = \frac{NH}{\theta} - \frac{NT}{1 - \theta} = 0 \implies \theta \ast NT = NH - \theta \ast NH$$

$$\theta(N_T + N_H) = N_H$$

$$\hat{\theta} = \frac{N_H}{N_T + N_H}$$

In our case of $D = \{H, H, T, H\}$ we receive $\hat{\theta} = \frac{N_H}{N_T + N_H} = \frac{3}{4}$

4 **Lagrange Multipliers**

4.1 **Definition**

In many cases we would like to solve problems with some constraints on the different parameters or on the solution. We use the Lagrange Multipliers as a function with a set of constraints, such that each constraint needs to be equal to zero, as follow:

$$\arg\max_{\vec{x} \in \mathbb{R}^k} f(\vec{x}) \text{ s.t. } \{c_1(x) = 0, c_2(x) = 0, ..., c_l(x) = 0\}$$

The Lagrangian then will be: $J(\vec{x}, \vec{\lambda}) = f(\vec{x}) - \sum_i \lambda_i c_i(x)$
When all the constraints are satisfied (equal to zero) then the Lagrangian will be the original function \(f(x)\). So, by comparing the derivative of \(f(x)\) to zero, we get the \(\theta\) that maximizes its value (depending on the function’s properties, needs to make sure that there is a maximum for the function).

### 4.2 Example - Rolling the Dice

\(\theta: \{\theta_1, ..., \theta_6\}\) for \(\theta_i\) a side of the die. The constraint will be: \(c_1(\theta) = \sum \theta_i - 1 = 0\)

D - series of die rolling: \(L(\theta : D) = \theta_1^{N_1} \cdot \theta_2^{N_2} \cdot ... \cdot \theta_6^{N_6}\)

The problem: to enlarge the product we can just keep enlarging the values of \(\theta_i\).

The solution: the constraint keeps it a legal distribution function. A violation of the constraint will drive us away from the maximum of the function.

In the end of the process we want to get the multinomial ML estimator \(\hat{\theta}_{ML} = \{\hat{\theta}_i = \frac{N_i}{N}\}\)

### 5 Sufficient Statistics

#### 5.1 Definition

We want to know what parameters are important for us, in order to describe the whole data, without saving it all.

Formally, we describe the Sufficient Statistics (SS) vector \(\vec{S}\) as one that act as follows:

\[ \forall D_1, \forall D_2 \text{ if } \vec{S}(D_1) = \vec{S}(D_2) \text{ then } \forall \theta L(D_1) = L(D_2) \]

The SS (not the nazi version) is derived from the data, and the likelihood is independent of the data, given the sufficient statistic.

\[ \vec{D} \rightarrow \vec{S} \rightarrow L \]

#### 5.2 Example

Going back to the dice rolling, we defined the general Lagrangian:

\[ J(\vec{x}, \vec{\lambda}) = f(\vec{x}) - \sum_i \lambda_i c_i(\vec{x}) \]

Now we get:

\[ J(\vec{x}, \vec{\lambda}) = \sum_i N_i \cdot \log(\theta_i) - \lambda (\sum_i \theta_i - 1) \]

\[ \frac{\partial J}{\partial \lambda} = -\sum_i \theta_i + 1 = 0 \implies \sum_i \theta_i = 1 \]

\[ \frac{\partial J}{\partial \theta_i} = \frac{N_i}{\theta_i} - \lambda = 0 \implies \theta_i = \frac{N_i}{\lambda} \]

We place \(\theta_i\) in the constraint so we get:
5.3 Alignment Problem

We go back to the DNA alignment problem. We have:

\[ \sigma(s_i, t_i) = \log \frac{p_1(s_i, t_i)}{p_0(s_i)p_0(t_i)} \]

for \( p_0 \) - probability for some letter in \( \sum \); and \( p_1 \) - distribution over the four letters as shown in the table, when all the non-X cells add up to 1:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( p_1(A, A) )</td>
<td>( p_1(A, C) )</td>
<td>( p_1(A, G) )</td>
<td>( p_1(A, T) )</td>
<td>( p_1(A, \cdot) )</td>
</tr>
<tr>
<td>C</td>
<td>X</td>
<td>( p_1(C, C) )</td>
<td>( p_1(C, G) )</td>
<td>( p_1(C, T) )</td>
<td>( p_1(C, \cdot) )</td>
</tr>
<tr>
<td>G</td>
<td>X</td>
<td>X</td>
<td>( p_1(G, G) )</td>
<td>( p_1(G, T) )</td>
<td>( p_1(G, \cdot) )</td>
</tr>
<tr>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>( p_1(T, T) )</td>
<td>( p_1(T, \cdot) )</td>
</tr>
<tr>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

The data we need in order to estimate \( p_1 \) needs to be a series of columns of correct alignments, of sequences with a common ancestor. Our problem is that we use \( \theta \) in order to get a correct alignment, but use correct alignments in order to estimate \( \theta \). We shall start with a general \( \theta \), and will improve it with every iteration on the data, until reaching convergence. An article from 92’ by Henikoff & Henikoff\(^4\) presented a way of doing so, showing that the values in the table are changing correlatively to the evolutionary distance.

6 References