EM

1.1 EM - Reminder and Motivation

Previously, we introduced the EM algorithm which is used for parameter estimation where some variables are hidden. We have seen it iteratively finds parameters that maximize a lower bound of the log-likelihood and estimates our hidden variables. Here we will want to see why taking the parameters that maximize the lower bound of the log likelihood, indeed improve the log likelihood each iteration.

1.2 Jensen’s inequality

Let $f : (a, b) \rightarrow \mathbb{R}$ be a real convex function. Jensen’s inequality states that

$$f(x) \text{ is convex} \rightarrow \mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

1.2.1 Convex functions - a reminder

Let $f$ be a differentiable function. We say that $f(x)$ is convex iff $\frac{\partial^2 f(x)}{\partial x^2} \geq 0$. For $f : (x_1, \ldots, x_n) \rightarrow \mathbb{R}$ we say that $f$ is convex iff the Hessian matrix, defined s.t. $H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$, is positive-semidefinite.
A matrix $H$ is positive-semidefinite if it is symmetric and $\forall y \neq 0$ it follows that $yHy^T \geq 0$.

Jensen’s inequality can also apply to concave functions, with a change of sign:

$$f(x) \text{ is concave} \rightarrow \mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$$

Note: Straight lines apply to both inequalities, resulting in an equality.

### 1.3 EM - background - what do we want to find?

We search for the value of $\theta$, the set of parameters for our model (that is $\tau, e$). We have $X_j$, which is the set of results of the emissions, and the problem is that we don’t know the hidden states $S^j$! Had we known, the parameter estimation would have been:

$$\hat{\tau}_{kl} = \frac{N_{kl}}{\sum_m N_{km}}$$

$$\hat{e}_{kx} = \frac{N_{kx}}{\sum_y N_{ky}}$$

And the sufficient statistic:

$$N_{kl} = \sum_{j=1}^{N} \sum_{i=2}^{N_j} \{ S^j_{i-1} = k, S^j_i = l \}$$

Where $N$ is the number of the number of sequences. So, to evaluate the parameters, $e$ and $\tau$, that maximize the log likelihood - we have to compute the sufficient statistic $N_{kl}$ for all $k, l \in S$.

Let us remember how LL was defined:

$$LL(\theta) = \sum_{j=1}^{N} \log P(x^j|\theta) = \sum_{j=1}^{N} \sum_{s^j} P(x^j, s^j|\theta) = \sum_{j=1}^{N} \log \left[ \sum_{s^j} \frac{P(x^j, s^j|\theta)}{Q(s^j)} \right]$$

for every distribution $Q$ over $s^j$.

$$= \sum_{j=1}^{N} \log \mathbb{E}_Q[\frac{P(x^j, s^j|\theta)}{Q(s^j)}]$$

$\geq \sum_{j=1}^{N} \mathbb{E}_Q[\log \frac{P(x^j, s^j|\theta)}{Q(s^j)}]$
Note: $s^j$ is a series of the hidden states for sequence $x^j$.

What we have found is a lower bound on $LL(\theta)$. We will want choose such a distribution $Q$ that will maximize the lower bound. By Gibb’s Inequality, the $Q$ for which $LL(\theta)$ is optimal satisfies in our case:

$\hat{Q}(s) \propto P(x, s|\theta)$ so the expectation is of a constant value.

$Q(s)$ is a probability distribution function, so the sum over it is 1:

$$\sum_s \hat{Q}(s) = \sum_s cP(x, s|\theta) = \frac{P(x, s|\theta)}{\sum_s P(x, s|\theta)} = P(s|x, \theta)$$

So we got that the optimal distribution (which maximizes the lower bound of the log likelihood of the data) over the hidden states, is equal to posterior of the hidden states.

Let’s substitute $Q$ in the lower bound:

$$LL(\theta) \geq \sum_{j=1}^N \mathbb{E}_Q[\log \frac{P(x^j, s^j|\theta)}{Q(s^j)}] = \sum_{j=1}^N \mathbb{E}_Q[\log \frac{P(x^j, s^j|\theta)}{P(s^j|x^j, \theta)}]$$

$$= \sum_{j=1}^N \mathbb{E}_Q[\log P(x^j|\theta)] = \mathbb{E}_Q[\sum_{j=1}^N \log P(x^j|\theta)] = \mathbb{E}_Q[LL(\theta)]$$

So we got:

$$LL(\theta) \geq \mathbb{E}_Q[LL(\theta)]$$

This means the expectation over the distribution $Q$ as we previously defined it of the log likelihood, is a lower bound to the log-likelihood we are trying to maximize. How this helps our maximization algorithm? We will start with an arbitrary $\theta$, and for each iteration we will maximize $\mathbb{E}_Q[LL(\theta)]$ over $Q$. That is, find $\mathbb{E}_Q[LL(\theta)]$. Then, we recalculate $\theta$, and repeat the process.

Note: We want to use EM (Expectation Maximization) like in this situation, when we have a set of hidden variables $s$ which we don’t know, and when it’s easy to express $LL$ given $x, s$.

### 1.4 Algorithm Baum-Welch

First, let’s remember the definitions of Forward, Backward, Transition and Emission:

- $F_i^j(i+1) = P(x_1^i, \ldots, x_i^j, x_{i+1}^j, s_{i+1} = l)$
- $B_i^j(i-1) = P(x_i^j, \ldots, x_n^j|s_{i-1} = l)$
- $\tau_{kl} = P(s_{i+1} = l|s_i = k)$
- $e_i(x_{i+1}) = P(x_{i+1}|s_{i+1} = l)$

Now let us describe the Baum-Welch Algorithm, which is the EM algorithm when applied to an HMM.
Algorithm 1 Baum-Welch

Init: Start with an arbitrary $\theta^{(0)}$

Loop: for $t = 1\ldots\infty$

Loop: for seq $j = 1\ldots N$

E-step:

$$N_{kl}^{(t)} = \sum_{j=1}^{N} \sum_{i=2}^{N_j} \mathbb{E}[1\{S_{i-1}^j = k, S_i^j = l\}] =$$

$$N_{kl}^{(t)} = \sum_{j=1}^{N} \sum_{i=2}^{N_j} P(S_{i-1}^j = k, S_i^j = l|x^j, \theta) =$$

$$N_{kl}^{(t)} = \sum_{j=1}^{N} \sum_{i=2}^{N_j} F_{j}^{(i-1)}(j_{t-1}) e_{j(t-1)}(x_i) B_k^l$$

$$N_{kl}^{(t)} = \cdots = \sum_{j=1}^{N} \sum_{i=2}^{N_j} P(s_i^j = k|x^j, \theta^{(t-1)}) =$$

$$N_{kl}^{(t)} = \sum_{j=1}^{N} \sum_{i=2}^{N_j} \frac{F_{j}^{(i-1)}(j_{t-1}) e_{j(t-1)}(x_i) B_k^l}{P(x^j)}$$

End-step: Estimate $\hat{\theta}^{(t)}$ (the MLE)

$$\hat{\tau}_{kl}^{(t)} = \frac{N_{kl}^{(t)}}{\sum_{m} N_{km}^{(t)}}$$

$$\hat{e}_{kx}^{(t)} = \frac{N_{kx}^{(t)}}{\sum_{y} N_{ky}^{(t)}}$$

Stop if $LL^{(t)} - LL^{(t-1)} \leq \text{threshold}$

end

2 Evolutionary trees

2.1 Evolution

Evolution is a process of change, not necessarily in the genome, but also in cellular differentiation, populations, specific genes, etc.

Questions that might interest us, with regard to evolution, are:

- Reconstruction of intermediate stages of the process.
- Chronological order of stages.
- Rate of the process.
- The probability of changes.
- The causes of the process.
We will refer to evolution in the context of evolutionary trees.

**Differentiation** events are the splits in the tree. A differentiation event for example:

![Evolutionary tree diagram](image)

We assume inheritance or derivation from ancestor (unlike bacteria where we have horizontal gene transfer) and we note that differentiation is not enough but the process of evolution also requires conservation of changes.

### 2.2 Species trees

**Species:** The largest group of organisms in which two individuals can produce fertile offspring.

#### 2.2.1 Speciation events

Consider the normal distribution of a trait (e.g. height) over a population:

![Normal distribution graphs](image)

We will call an occurrence of a complete division of the species a *speciation event*:

![Speciation event](image)

We view the speciation process as a single event.

We will discuss 3 types of trees: Rooted, Unrooted and Ultrametric.

#### 2.2.2 Rooted trees

- The leaves are species that exist today.
- The internal nodes are extinct species.
- For every pair of nodes we define *LCA* - the least common ancestor. Every sub-tree defined by an LCA is called *clade*.
- The tree is binary.
2.2.3 Unrooted trees

- We’ll use it when we cannot know where to place the most ancient ancestor.
- The degree of every internal node is 3.
- We may talk about the distance between species, but not about time (who came first...)
- We can also ‘hang’ the tree from some edge and add a node as the root.

2.2.4 Ultrametric trees

- Directed, rooted.
- Has the ultrametric property: \( \forall v \in \text{Internal}, \forall l_1, l_2 \in \text{Leaves}(v) : d(v, l_1) = d(v, l_2) \). In words, all the leaves of \( v \) are equally distant from it.
- We assume that \( \text{branchLength} = \text{time} \ast \text{mutationRate} \). Because the time that passed since the speciation event is the same time:

![Ultrametric Trees Diagram]

This means that in ultrametric trees - the mutation rate is constant for all species. This is also called the Molecular Clock Hypothesis.

2.3 Molecular aspect of trees

Molecular changes occur in the genome, in 3 main manners: gene duplication, gene deletion and whole genome duplication.

2.3.1 Gene duplication

![Gene Duplication Diagram]

A complete gene \( A \) is duplicated and can be found in another area in the genome. We will look at the evolutionary tree form the point of view of gene \( A \).
2.3.2 Gene deletion

Gene deletion may occur, for example, if a termination codon is deleted or formed. Every gene that is not expressed is considered deleted.

2.3.3 Whole gene duplication

This is a rare event.

*Orthologous genes* are genes in different organisms that originated from the same gene. For example, the histone component H3 is highly preserved. H3 is human and H3 in yeast are orthologous.

*Paralogous Genes* are genes in the same organism that were derived from the same gene. For example, the two H3 variants in yeast, H3.1, H3.2 are paralogous.

Now, let’s look at the similarity histogram of pairs of genes in the same organism: we take all pairs of genes and calculate the percentage of difference between them. Create a histogram according to the percentage of difference. We will see such a graph:
The graph goes this way because genes accumulate mutations over time.
A similar graph would be achieved for plotting the function similarity against
the percentage of difference.

Researches made such a Count vs. % difference histogram for the genome of
Arabidopsis, and this is what they saw:

The peak in the beginning of the graph is because we see a lot of pairs of
genes with a relatively low % difference. This boost in the pairs with such a
small difference indicates a whole genome duplication event.

2.3.4 Gene trees vs. Species trees

Now, let’s see the difference between the 2 types of trees. Let $T$ be a species
tree (in black) and let $T'$ be a gene tree (in blue, marked in pink/orange), the
root of which is an ancestor gene.
Suppose we have deletion events of $a'$ and $b$. Then the tree constructed for this gene:

Then suppose also $d$ and $c'$ were deleted. So we get:

Assuming the gene tree constructed for a single gene gives us a wrong species tree construction.