

## BCB/GDCB/STAT/COM S 568 Spring 2009

### Lecture Notes on Expectation Maximization

February, 2009

#### Preliminaries

In the derivation of the EM-step below we are going to make use of the following result:

Let  $f(x_1, x_2, \dots, x_r) = \sum_{i=1}^r w_i \log x_i$ , where  $\sum_{i=1}^r w_i = w$  and  $\sum_{i=1}^r x_i = 1$ . Then  $f(x_1, x_2, \dots, x_r)$  assumes its maximum when  $x_i = \frac{w_i}{w}$  for  $i = 1, 2, \dots, r$ .

**Proof.** Using the method of Lagrange multipliers from multivariate calculus, the  $x_i$  are known to be the solution of the equations

$$\frac{\partial f}{\partial x_i} - \lambda \frac{\partial g}{\partial x_i} = 0,$$

where  $g(x_1, x_2, \dots, x_r) = \sum_{i=1}^r w_i$  and  $\lambda$  is some constant. Here, the differential equations give  $\frac{w_i}{x_i} - \lambda = 0$ , or  $x_i = \frac{w_i}{\lambda}$ . Summing over  $i$  gives  $\lambda = w$ , which proves the assertion.

#### Motif Finding

A common problem in sequence analysis is that of finding a *motif* that is common to a set of sequences. A typical example would be the task of identifying a specific protein binding site in a set of unaligned DNA fragments (e.g., data derived from ChIP-chip or ChIP-seq experiments). We shall specify a general probabilistic model for the sequences and then show how to best parameterize the model.

Let there be  $n$  sequences  $S_1, S_2, \dots, S_n$  of lengths  $L_1, L_2, \dots, L_n$ , respectively, consisting of letters drawn from an alphabet  $A$  ( $\{A, C, G, T\}$  for DNA). We assume the motif is exactly of length  $W$  and may or may not occur in any one of the sequences. In each position  $i$  of the motif, let the letter  $j$  be observed with probability  $P_{ij}$ . If the motif occurs, then we allow for the possibility that the sequences upstream (to the left) and downstream (to the right) are of different average composition. Precisely, we assume that in the left context, letters are drawn independently with probability distribution  $P_{Lj}, j \in A$ . Similarly, there is a right probability distribution  $P_{Rj}, j \in A$ .

To completely specify the model, let the probability that sequence  $S_s$  contains the motif be  $P_{\sigma_s}$ . Let  $\theta = \{P_{Lj}, P_{ij}, P_{Rj}, P_{\sigma_s}\}$  denote a particular set of parameters. Given these parameters, the likelihood of sequence  $S_s$  is  $g_\theta(S_s)$ , and if the sequences are independent, then the likelihood of the entire set of sequences is simply  $g_\theta(S) = \prod_s g_\theta(S_s)$ . It is our goal to find a parameter estimate  $\hat{\theta}$  such that the data likelihood  $g_\theta(S)$  is maximized. In other words, we are looking for

the best possible model fit for the observed sequences. This is a difficult maximization problem that we don't know how to solve in general. However, we will give an iterative procedure (a particular example of what is known in statistics as Expectation Maximization) that for any parameter set  $\theta$  produces a parameter set  $\theta'$  that gives at least as large a likelihood of the data as our original parameter set (and hopefully larger). Iterating this procedure until the increase in likelihood becomes marginal would give a parameter set that represents at least a local maximum of the likelihood function. Consistent results from many random starting points should in practice produce the global maximum.

For simplicity of notation, we consider a single sequence  $S_s$ . The general case of  $n$  sequences follows immediately and only requires replacing sums over  $k$  below by double sums over  $s$  and  $k$ . If we know where the motif starts in each sequence (the *hidden information*), then the likelihood of the single sequence is easy to compute. Let  $k$  be the starting position of the motif in sequence  $S_s$ , then

$$l_\theta(S_s | k) = \left( \prod_j P_{L_j}^{n_{L_s k j}} \right) \left( \prod_{i=0}^{w-1} \prod_j P_{ij}^{I(S_s, k+i=j)} \right) \left( \prod_j P_{R_j}^{n_{R_s k j}} \right)$$

where  $n_{L_s k j}$  is the number of occurrences of letter  $j$  in positions 1 through  $k-1$ ,  $n_{R_s k j}$  is the number of occurrences of letter  $j$  in positions  $k+w$  to  $L_s$ , and  $I(S_s, i=j)$  is an indicator function that evaluates to 1 if the  $i$ th position of the sequence  $S_s$  is occupied by letter  $j$ .

When the location  $k$  of the motif is unknown, we must integrate over all possibilities

$$g_\theta(S_s) = \sum_k l_\theta(S_s | k) w_{s\theta}(k) := \sum_k h_\theta(S_s, k)$$

where  $w_{s\theta}(k)$  is the model (a priori) probability for the motif to occur at position  $k$  in  $S_s$ . For unaligned sequences the motif position is unknown, and thus we may specify uniform probabilities:

$$w_{s\theta}(k) = \begin{cases} \frac{P_{\sigma_s}}{L_s - w + 1} & k = 1, 2, \dots, L_s - w + 1 \\ \frac{1 - P_{\sigma_s}}{2} & k = 0, L_s - w + 2 \end{cases}$$

Here,  $k = 0$  represents the case that the sequence is all right context, and  $k = L_s - w + 2$  represents the case that the sequence is all left context.

Our goal is to find parameters  $\theta'$  such that  $g_{\theta'}(S_s) \geq g_\theta(S_s)$ . Let's introduce

$$w_\theta(k | S_s) = \frac{h_\theta(S_s, k)}{g_\theta(S_s)}$$

and write

$$\log g_\theta(S_s) = \log h_\theta(S_s, k) - \log w_\theta(k | S_s).$$

The lefthand side notation clearly shows that  $g_\theta(S_s)$  is independent of  $k$ , and thus taking the expectation with respect to the  $w_\theta(k | S_s)$  gives

$$\log g_\theta(S_s) = \sum_k w_\theta(k | S_s) \log g_\theta(S_s) = \sum_k w_\theta(k | S_s) \log h_\theta(S_s, k) - \sum_k w_\theta(k | S_s) \log w_\theta(k | S_s).$$

We can derive the analogous equation for parameter set  $\theta'$ , and then our task is seen to reduce to finding a set  $\theta'$  such that the following expression becomes non-negative:

$$\log g_{\theta'}(S_s) - \log g_{\theta}(S_s) = \sum_k w_{\theta}(k | S_s) \log \frac{h_{\theta'}(S_s, k)}{h_{\theta}(S_s, k)} - \sum_k w_{\theta}(k | S_s) \log \frac{w_{\theta'}(k | S_s)}{w_{\theta}(k | S_s)}$$

Using the result from our Preliminaries, note first that the second term in the above equation is already non-negative. To see this, write  $-\sum_k w_{\theta}(k | S_s) \log \frac{w_{\theta'}(k | S_s)}{w_{\theta}(k | S_s)}$  as  $\sum_k w_{\theta}(k | S_s) \log w_{\theta}(k | S_s) - \sum_k w_{\theta}(k | S_s) \log w_{\theta'}(k | S_s)$ , equate the  $w_{\theta'}(k | S_s)$  with the unknowns  $x_i$  and the  $w_{\theta}(k | S_s)$  with the  $w_i$  in our Preliminary notation, and recall that the sum is maximal for  $w_{\theta'}(k | S_s) = w_{\theta}(k | S_s)$ . Thus we'd have our task accomplished if we could find  $\theta'$  such that  $\sum_k w_{\theta}(k | S_s) \log h_{\theta'}(S_s, k)$  is maximal over all choices of  $\theta'$  (as then clearly the first time would also be non-negative).

Writing  $h_{\theta'}(S_s, k)$  explicitly, we have

$$\begin{aligned} \sum_k w_{\theta}(k | S_s) \log h_{\theta'}(S_s, k) &= \sum_k w_{\theta}(k | S_s) \\ &\times \left[ \sum_j n_{Lskj} \log P'_{Lj} + \sum_{i=0}^w \sum_j I(S_{s,i+k} = j) \log P'_{ij} + \sum_j n_{Rskj} \log P'_{Rj} + \log w_{s\theta'}(k) \right] \end{aligned}$$

which consists of the independent terms

$$\begin{aligned} &\sum_j \left[ \sum_k w_{\theta}(k | S_s) n_{Lskj} \right] \log P'_{Lj} \\ &\sum_{i=0}^{w-1} \sum_j \left[ \sum_k w_{\theta}(k | S_s) I(S_{s,i+k} = j) \right] \log P'_{ij} \\ &\sum_j \left[ \sum_k w_{\theta}(k | S_s) n_{Rskj} \right] \log P'_{Rj} \\ &\sum_k w_{\theta}(k | S_s) \log w_{s\theta'}(k) \end{aligned}$$

that are all of the form of the function discussed in the Preliminaries. Thus, maximization is achieved for

$$\begin{aligned} P'_{Lj} &= \frac{\sum_k w_{\theta}(k | S_s) n_{Lskj}}{\sum_l \sum_k w_{\theta}(k | S_s) n_{Lskl}} \\ P'_{ij} &= \frac{\sum_k w_{\theta}(k | S_s) \sum_{i=k}^{k+w} I(S_{si} = j)}{\sum_l \sum_k w_{\theta}(k | S_s) \sum_{i=k}^{k+w} I(S_{si} = l)} \\ P'_{Rj} &= \frac{\sum_k w_{\theta}(k | S_s) n_{Rskj}}{\sum_l \sum_k w_{\theta}(k | S_s) n_{Rskl}} \\ w_{s\theta'}(k) &= w_{\theta}(k | S_s) \end{aligned}$$