Maximum Likelihood Estimation of a binomial Distribution

If we model methylation as a binomial distribution then we can call p the probability of a read at a particular site to be methylated or not. Let C_l be the number of methylated observations in sample l and n_l be the total observations. So the likelihood is:

$$\prod_l \binom{n_l}{C_l} p_l^C (1-p)^{n_l - c_l}$$

and the log-likelihood is:

$$\sum_l \log \binom{n_l}{C_l} + C_l \log(p) + (n_l - C_l)\log(1-p)$$

Taking the derivative and setting equal to zero we get:

$$\sum_l \frac{C_l}{p} - \frac{n_l - C_l}{1-p} = 0 \Rightarrow \sum_l C_l(1-p) - p(n_l - C_l) = 0$$
$$\Rightarrow \sum_l C_l - C_l p - p n_l + p C_l = 0 \Rightarrow \sum_l C_l - n_l p = 0$$
$$\Rightarrow p = \sum_l \frac{C_l}{n_l}$$

Calculating the sufficient statistics

Let k be the hidden state. Each hidden state has it's own binomial destitution parameters.
Let C_k/ T_k be the number of methylated/unmethylated observations of all sites which have hidden state k. We have shown that the maximum likelihood estimator of the parameters of the binomial distribution of hidden state k are

$$p = \frac{C_k}{C_k + T_k}$$

Thus the sufficient statistics for finding the ML estimator are C_k and T_k. Since this is unlabelled data and we do not know which sites belong to which hidden states, we can estimate $\mathbf{E}[C_k]$ and $\mathbf{E}[T_k]$.
Let x^l be the set of sequences from which we want to learn our HMM parameters (i.e. each x^l is a WGBS sample).
Then x^{l_j} refers to read j in sample l
$x_i^{l_j}$ refers to read j in sample l at CpG site i.

Thus:

$$C_k = \sum_{l=1}^{M} \sum_{j=1}^{R} \sum_{i=1}^{N} \mathbb{1}\{S_i = k, x_i^{l_j} = C\}$$

and so:

$$\mathbf{E}[C_k] = \sum_{l=1}^{M} \sum_{j=1}^{R} \sum_{i=1}^{N} P(S_i = k, x_i^{l_j} = C | x^{l_j})$$

Where the conditional is provided since x is observed.

$$\begin{aligned}
\mathbf{E}[C_k] &= \sum_{l=1}^{M} \sum_{j=1}^{R} \sum_{i=1}^{N} P(x_i^{l_j} = C | x^{l_j}, S_i = k) * P(S_i = k | x^{l_j}) \\
&= \sum_{l=1}^{M} \sum_{j=1}^{R} \sum_{i=1}^{N} \mathbb{1}\{x_i^{l_j} = C\} * \frac{F_k(i) B_k(i)}{P(x^{l_j})} \\
&= \sum_{l=1}^{M} \sum_{i=1}^{N} C_i^l \frac{F_k(i) B_k(i)}{P(x^{l_j})}
\end{aligned}$$

Where C_i^l is the number of methylated observations in sample l at site i and $F_k(i)$ and $B_k(i)$ are the forwards and backwards tables from the Baum-Welch algorithm learned in class and $p(x)$ is the likelihood of the sample.