<1> **Lemma.** *Let $\{p_i : i = 1, \ldots, n\}$ and $\{q_i : i = 1, \ldots, n\}$ be two sets of nonnegative numbers with $\sum_i p_i = 1 = \sum_i q_i$. Then*

$$\sum_i p_i \log(p_i/q_i) \geq \frac{1}{2}\Big(\sum_i |p_i - q_i|\Big)^2$$

See the bonus question on Sheet 4 for a proof.

<2> **Corollary.** *The function $f(\mathbf{q}) = \sum_i p_i \log q_i$ is maximized over probability distributions $\mathbf{q} = (q_1, \ldots, q_n)$ when $\mathbf{q} = \mathbf{p}$.*

1. Generalized EM

The EM algorithm is an iterative procedure tha tries to maximize a function

$$G(\theta) = \sum\nolimits_{x \in \mathbb{X}} g(x, \theta)$$

where $g(x, \theta)$ is a known, strictly positive function of $x \in \mathcal{X}$ and $\theta \in \Theta$. Each iteration works by identifying the ratio $p(x, \theta) = g(x, \theta)/G(\theta)$ as a probability distribution on $\mathcal{X}$ for each fixed θ. Starting from a guess θ_0, it generates a new guess θ_1 in two steps:

E-step: Define $H(\theta) := \mathbb{E}_{\theta_0} \log g(x, \theta) = \sum_x p(x, \theta_0) \log(x, \theta)$, an expectation under the $p(\cdot, \theta_0)$ distribution.

M-step: Find θ_1 to maximize $H(\theta)$, or at least such that $H(\theta_1) > H(\theta_0)$.

These two steps lead to an increase in G:

$$H(\theta_1) = \mathbb{E}_{\theta_0} \log\big(p(x, \theta_1)G(\theta_1)\big) > \mathbb{E}_{\theta_0} \log\big(p(x, \theta_0)G(\theta_0)\big) = H(\theta_0)$$

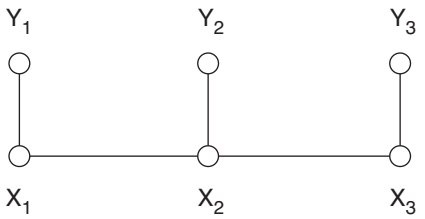
which rearranges to

$$\begin{aligned}
\log\big(G(\theta_1)/G(\theta_0)\big) &> \mathbb{E}_{\theta_0} \log\big(p(x, \theta_0)/p(x, \theta_1)\big) \\
&= \sum\nolimits_x p(x, \theta_0) \log\big(p(x, \theta_0)/p(x, \theta)\big) \\
&\geq \frac{1}{2}\sum\nolimits_x \big|p(x, \theta_0) - p(x, \theta_1)\big|^2 \qquad \text{by Lemma <1>} \\
&\geq 0
\end{aligned}$$

Thus $G(\theta_1) > G(\theta_0)$.

Repeat the two steps, starting from θ_1 to generate a θ_2. And so on.

2. A Hidden Markov model



Suppose we have a MRF on a graph with six nodes, the random variable at each node taking values 0 or 1, with joint distribution indexed by $\theta = (\alpha, \beta)$:

$$\begin{aligned}
&\mathbb{P}_\theta\{X_1 = 1\} = \tfrac{1}{2} \\
&\mathbb{P}_\theta\{X_{i+1} = X_i \mid X_i = b\} = \alpha \qquad \text{for } i = 1, 2 \text{ and } b = 0, 1 \\
&\mathbb{P}\{Y_i = X_i \mid X_i = b\} = \beta \qquad \text{for } i = 1, 2, 3 \text{ and } b = 0, 1
\end{aligned}$$

If we observe $\mathbf{X} = \mathbf{x}$ and $\mathbf{Y} = \mathbf{y}$, the method of maximum likelihood chooses $\widehat{\theta}$ to maximize

$$\begin{aligned} g(\mathbf{x}, \mathbf{y}, \theta) &= \mathbb{P}_\theta\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} = \mathbb{P}_\theta\{\mathbf{X} = \mathbf{x}\}\mathbb{P}_\theta\{\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}\} \\ &= \tfrac{1}{2}\alpha^{1\{x_2 = x_1\}}(1-\alpha)^{1\{x_2 \neq x_1\}}\alpha^{1\{x_3 = x_2\}}(1-\alpha)^{1\{x_3 \neq x_2\}} \times \textstyle\prod_{i=1}^{3} \beta^{1\{y_i = x_i\}}(1-\beta)^{1\{y_i \neq x_i\}} \end{aligned}$$

Equivalently, the method maximizes

$$\begin{aligned} &\log g(\mathbf{x}, \mathbf{y}, \theta) - \log(1/2) \\ &\qquad = M \log\alpha + (2 - M)\log(1-\alpha) + N\log\beta + (3-N)\log(1-\beta) \\ &\qquad = 2\left(\frac{M}{2}\log\alpha + \frac{2-M}{2}\log(1-\alpha)\right) + 3\left(\frac{N}{3}\log\beta + \frac{3-N}{3}\log(1-\beta)\right) \end{aligned} \qquad \text{<3>}$$

where

$$M = \textstyle\sum_{i=1}^{2} 1\{x_{i+1} = x_i\} \qquad \text{and} \qquad N = \sum_{i=1}^{3} 1\{x_i = y_i\}.$$

For example, for $\mathbf{y} = (1, 0, 1)$, the following table gives the values of the two factors for each possible $\mathbf{x}$.

$\mathbf{x}$	$2\mathbb{P}_\theta\{\mathbf{X} = \mathbf{x}\}$	$\mathbb{P}_\theta\{\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x}\}$	M	N
111	α^2	$\beta^2(1-\beta)$	2	2
000	α^2	$\beta(1-\beta)^2$	2	1
100	$\alpha(1-\alpha)$	$\beta^2(1-\beta)$	1	2
001	$\alpha(1-\alpha)$	$\beta^2(1-\beta)$	1	2
110	$\alpha(1-\alpha)$	$\beta(1-\beta)^2$	1	1
011	$\alpha(1-\alpha)$	$\beta(1-\beta)^2$	1	1
101	$(1-\alpha)^2$	β^3	0	3
010	$(1-\alpha)^2$	$(1-\beta)^3$	0	0

If both $\mathbf{x}$ and $\mathbf{y}$ are observed, the maximum likelihood estimators can be determined by a separate maximization of the last two expressions in <3>. From Corollary <2>, the maximizing values are

$$\widehat{\alpha} = M/2 \qquad \text{and} \qquad \widehat{\beta} = N/3$$

If we only observe that $\mathbf{Y} = \mathbf{y} = (1, 0, 1)$, the $\mathbf{X}$ values being hidden, the maximum likelihood estimator are chosen to maximize

Identifiable?

$$2G(\theta) = 2\mathbb{P}_\theta\{\mathbf{Y} = \mathbf{y}\} = \alpha^2\beta(1-\beta) + 2\alpha(1-\alpha)\beta(1-\beta) + (1-\alpha)^2\left(\beta^3 + (1-\beta)^3\right)$$

For EM with $g(\mathbf{x}, \theta) = \mathbb{P}_\theta\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\}$, we have

$$p(\mathbf{x}, \theta) = g(\mathbf{x}, \theta)/G(\theta) = \mathbb{P}_\theta\{\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y}\}.$$

I have omitted the dependence of g on $\mathbf{y}$ because it stays fixed throughout the EM argument.

With starting guess $\theta_0 = (\alpha_0, \beta_0)$, one EM step seeks to maximize

$$\begin{aligned} H(\theta) &= \textstyle\sum_{\mathbf{x}} p(\mathbf{x}, \theta_0)\log g(\mathbf{x}, \theta) \\ &= \mathbb{E}_{\theta_0}\left(\log g(\mathbf{X}, \theta) \mid \mathbf{Y} = \mathbf{y}\right) \\ &= \log(1/2) + \mathbb{E}_{\theta_0}\left(M \mid \mathbf{Y} = \mathbf{y}\right)\log\alpha + \mathbb{E}_{\theta_0}\left(2 - M \mid \mathbf{Y} = \mathbf{y}\right)\log(1-\alpha) \\ &\qquad + \mathbb{E}_{\theta_0}\left(N \mid \mathbf{Y} = \mathbf{y}\right)\log\beta + \mathbb{E}_{\theta_0}\left(3 - N \mid \mathbf{Y} = \mathbf{y}\right)\log(1-\beta) \end{aligned}$$

Again from Corollary <2>, the maximizing values are

$$\alpha_1 = \frac{1}{2}\mathbb{E}_{\theta_0}\left(M \mid \mathbf{Y} = \mathbf{y}\right) \qquad \text{and} \qquad \beta_1 = \frac{1}{3}\mathbb{E}_{\theta_0}\left(N \mid \mathbf{Y} = \mathbf{y}\right)$$

And so on.

In this simple example, we could just grind out the conditional expectations by explicit, brute force calulation. For an analogous problem with n instead of 3 hidden states, with n large, we would need to find a more systematic metod for organizing the calculations. See Chang §3.5.4 for the treatment of a slightly more general problem.

Possible (one person) project

Write a short report covering the following points. I would hope to follow what you had done without having to work through the fine details of any programs.

(i) Extend the Hidden Markov model to the case of n pairs (X_i, Y_i), deriving the form of the MLE (if both $\mathbf{x}$ and $\mathbf{y}$ are observed) or of the EM step (if only $\mathbf{y}$ is observed).

(ii) Write a small program to generate $\mathbf{x}$ and $\mathbf{y}$ for any specified α and β and n.

(iii) Implement the full method described in Chang §3.5.4 for estimating the transition probabilities via EM from only the observed $\mathbf{y}$. (That is, ignore the parametric specification of the transition probabilities, and estimate the probabilities usuing the method described by Joe.)

(iv) Adapt the method from (iii) to estimate α and β from observed $\mathbf{y}$. That is, try to find a method analogous to the one in §3.5.4 for calculating the necessary conditional expectations.

(v) Explain your methods and produce some informative output for an n large enough to demonstrate the need for the algorithm in (iv). You should compare the MLE and the output from (iii) and (iv) for data generated by your program from (ii).