

050/520/600.666 Information Extraction from Speech and Text

Homework # 7

Due April 14, 2005.

Generalization of the Results of Section 9.4.1

It is suggested in Section 9.4.2 that the results of Section 9.4.1 for 2-dimensional observations extend easily to d -dimensions. We will work through the details in this problem. Specifically, we will consider an HMM with output densities attached to arcs. Let \mathcal{S} denote the set of states, let the arcs be indexed by $t \in \mathcal{T}$, and let the outputs or emissions take values in \mathbb{R}^d for some finite $d > 0$. Let $L(t)$ and $R(t)$ respectively denote the origin- and destination-states of the arc t , and let p_t denote the probability of taking the arc t when the underlying Markov chain is in $L(t)$. Clearly,

$$\sum_{t: L(t)=s} p_t = 1, \quad \forall s \in \mathcal{S}. \quad (1)$$

For each non-null arc t , let the corresponding output density be a multivariate Gaussian,

$$\mathcal{N}_t(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{|\mathbf{U}_t|}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{m}_t)^T \mathbf{U}_t^{-1} (\mathbf{y} - \mathbf{m}_t) \right\}, \quad \forall \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_d \end{bmatrix} \in \mathbb{R}^d, \quad (2)$$

where \mathbf{m}_t is the mean vector and \mathbf{U}_t the covariance matrix of the emitted random vector. Note that \mathbf{y} and \mathbf{m}_t are column vectors here, while they are a row-vector in the textbook, and the arc-dependence of \mathbf{m}_t and \mathbf{U}_t is denoted via a subscript instead of writing $\mathbf{m}(t)$ and $\mathbf{U}(t)$. The free parameters of the HMM are $\theta = \{\theta_t, t \in \mathcal{T}\}$, where $\theta_t = \{p_t, \mathbf{m}_t, \mathbf{U}_t\}$, the p_t 's satisfy the sum-to-one condition (1), and the \mathbf{U}_t 's are symmetric and positive-definite.

Given the k -length observation $\mathbf{Y} = \langle \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k \rangle$ from this HMM, the EM auxiliary function may be constructed as

$$Q(\theta', \theta) = \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \log P_{\theta}(\mathbf{t}, \mathbf{Y}) = \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \log \left[\prod_{l=1}^{n(\mathbf{t})} p_{t_l} \mathcal{N}_{t_l}(\mathbf{y}_l) \right], \quad (3)$$

where $\mathbf{t} = \langle t_1, t_2, \dots, t_{n(\mathbf{t})} \rangle$ denotes any valid path through the HMM, and $n(\mathbf{t})$ denotes its length. While it is not made precise in the textbook, it is to be understood in (3) that

- $P_{\theta'}(\mathbf{t}|\mathbf{Y}) > 0$ only for paths \mathbf{t} of length $n(\mathbf{t}) \geq k$ that contain exactly k non-null arcs and $n - k$ null arcs, and hence other paths need not be considered in the sum over all \mathbf{t} ;

- the reference to the l -th output symbol \mathbf{y}_l is valid only after reindexing $\mathbf{Y} = \langle \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k \rangle$ and (re)assigning \mathbf{y}_1 to the first non-null arc of \mathbf{t} , \mathbf{y}_2 to the second non-null arc of \mathbf{t} , and so on, until \mathbf{y}_k to the last non-null arc of \mathbf{t} , while no symbols are assigned to the null arcs of \mathbf{t} ;
- $\mathcal{N}_{t_l}(\mathbf{y}_l)$ is computed via (2) for non-null arcs t_l in \mathbf{t} , but $\mathcal{N}_{t_l}(\cdot) = 1$ for all null arcs in \mathbf{t} .

Next, given a θ' , we must try to maximize $Q(\theta', \theta)$ as a function of θ . To this end, we form the Lagrangian

$$L(\theta) = Q(\theta', \theta) - \sum_{s \in \mathcal{S}} \lambda_s \sum_{t' : L(t')=s} p_{t'}, \quad (4)$$

and note that for every arc $t \in \mathcal{T}$,

$$\begin{aligned} \frac{\partial}{\partial p_t} L(\theta) &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \frac{\partial}{\partial p_t} \log P_{\theta}(\mathbf{t}, \mathbf{Y}) - \frac{\partial}{\partial p_t} \sum_{s \in \mathcal{S}} \lambda_s \sum_{t' : L(t')=s} p_{t'} \\ &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \frac{\partial}{\partial p_t} \log \left[\prod_{l=1}^{n(\mathbf{t})} p_{t_l} \mathcal{N}_{t_l}(\mathbf{y}_l) \right] - \lambda_{L(t)} \\ &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \frac{\partial}{\partial p_t} \left[\sum_{l=1}^{n(\mathbf{t})} \log p_{t_l} + \sum_{l=1}^{n(\mathbf{t})} \log \mathcal{N}_{t_l}(\mathbf{y}_l) \right] - \lambda_{L(t)} \\ &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \left[\sum_{l: t_l=t} \frac{\partial}{\partial p_t} \log p_t + 0 \right] - \lambda_{L(t)} \\ &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \sum_{l: t_l=t} \frac{1}{p_t} - \lambda_{L(t)}. \end{aligned}$$

(I) Argue carefully that for a non-null arc $t \in \mathcal{T}$,

$$\frac{\partial}{\partial p_t} L(\theta) = 0 \quad \Rightarrow \quad p_t = \frac{1}{K_{L(t)}} \sum_{l=1}^k [\alpha'_{l-1}(L(t)) p'_t \mathcal{N}'_t(\mathbf{y}_l) \beta'_l(R(t))],$$

where the parameters p'_t and \mathcal{N}'_t correspond to θ' , $\alpha'_*(\cdot)$ and $\beta'_*(\cdot)$ are the forward- and backward-probabilities computed using θ' on a k -stage trellis, with the null arcs going between vertically aligned states and only non-null arcs traversing left-to-right, and $K_{L(t)}$ is a normalization constant. Similarly show that for null arcs $t \in \mathcal{T}$,

$$\frac{\partial}{\partial p_t} L(\theta) = 0 \quad \Rightarrow \quad p_t = \frac{1}{K_{L(t)}} \sum_{l=1}^k [\alpha'_l(L(t)) p'_t \beta'_l(R(t))],$$

where the states are topologically sorted, the $\alpha'_l(\cdot)$'s are calculated top-to-bottom and the $\beta'_l(\cdot)$'s bottom-to-top. Finally, write a closed form expression for K_s , $s \in \mathcal{S}$.

Next, note that for every non-null arc $t \in \mathcal{T}$, if we let $\mathbf{m}_t = [m_{t,1} \dots m_{t,d}]^T$, then

$$\frac{\partial}{\partial m_{t,i}} L(\theta) = \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \frac{\partial}{\partial m_{t,i}} \log P_{\theta}(\mathbf{t}, \mathbf{Y}) - \frac{\partial}{\partial m_{t,i}} \sum_{s \in \mathcal{S}} \lambda_s \sum_{t' : L(t')=s} p_{t'}$$

$$\begin{aligned}
&= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \frac{\partial}{\partial m_{t,i}} \log \left[\prod_{l=1}^{n(\mathbf{t})} p_{t_l} \mathcal{N}_{t_l}(\mathbf{y}_l) \right] - 0 \\
&= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \frac{\partial}{\partial m_{t,i}} \left[\sum_{l=1}^{n(\mathbf{t})} \log p_{t_l} + \sum_{l=1}^{n(\mathbf{t})} \log \mathcal{N}_{t_l}(\mathbf{y}_l) \right] \\
&= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \left[0 + \sum_{l:t_l=t} \frac{\partial}{\partial m_{t,i}} \log \mathcal{N}_t(\mathbf{y}_l) \right] \\
&= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \sum_{l:t_l=t} \frac{\partial}{\partial m_{t,i}} \left\{ -\frac{1}{2} (\mathbf{y}_l - \mathbf{m}_t)^T \mathbf{U}_t^{-1} (\mathbf{y}_l - \mathbf{m}_t) - \log(2\pi)^{\frac{d}{2}} \sqrt{|\mathbf{U}_t|} \right\} \\
&= -\frac{1}{2} \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \sum_{l:t_l=t} \frac{\partial}{\partial m_{t,i}} \{ (\mathbf{y}_l - \mathbf{m}_t)^T \mathbf{U}_t^{-1} (\mathbf{y}_l - \mathbf{m}_t) + 0 \}.
\end{aligned}$$

The partial derivatives of $L(\theta)$ with respect to the components of the mean vector \mathbf{m}_t may therefore be compactly written as the vector

$$\begin{bmatrix} \frac{\partial}{\partial m_{t,1}} L(\theta) \\ \vdots \\ \frac{\partial}{\partial m_{t,d}} L(\theta) \end{bmatrix} = -\frac{1}{2} \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \sum_{l:t_l=t} \begin{bmatrix} \frac{\partial}{\partial m_{t,1}} \{ (\mathbf{y}_l - \mathbf{m}_t)^T \mathbf{U}_t^{-1} (\mathbf{y}_l - \mathbf{m}_t) \} \\ \vdots \\ \frac{\partial}{\partial m_{t,d}} \{ (\mathbf{y}_l - \mathbf{m}_t)^T \mathbf{U}_t^{-1} (\mathbf{y}_l - \mathbf{m}_t) \} \end{bmatrix}. \quad (5)$$

(II) Show that for any $d \times 1$ vectors $\mathbf{x} = [x_1 \dots x_d]^T$ and $\mathbf{b} = [b_1 \dots b_d]^T$, and symmetric $d \times d$ matrix \mathbf{A} , the partial derivatives of the *scalar* $\mathbf{x}^T \mathbf{b}$ with respect to the components of \mathbf{x} are

$$\begin{bmatrix} \frac{\partial}{\partial x_1} \mathbf{x}^T \mathbf{b} \\ \vdots \\ \frac{\partial}{\partial x_d} \mathbf{x}^T \mathbf{b} \end{bmatrix} = \mathbf{b},$$

and the partial derivatives of the *scalar* $\mathbf{x}^T \mathbf{A} \mathbf{x}$ with respect to the components of the vector \mathbf{x} are

$$\begin{bmatrix} \frac{\partial}{\partial x_1} \mathbf{x}^T \mathbf{A} \mathbf{x} \\ \vdots \\ \frac{\partial}{\partial x_d} \mathbf{x}^T \mathbf{A} \mathbf{x} \end{bmatrix} = \mathbf{A}^T \mathbf{x} + \mathbf{A} \mathbf{x} = 2\mathbf{A} \mathbf{x}.$$

(III) Combine the results above with (5) to show that the partial derivative of $L(\theta)$ w.r.t. \mathbf{m}_t is

$$\begin{bmatrix} \frac{\partial}{\partial m_{t,1}} L(\theta) \\ \vdots \\ \frac{\partial}{\partial m_{t,d}} L(\theta) \end{bmatrix} = \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \sum_{l:t_l=t} \mathbf{U}_t^{-1} (\mathbf{y}_l - \mathbf{m}_t),$$

and show that

$$\begin{bmatrix} \frac{\partial}{\partial m_{t,1}} L(\theta) \\ \vdots \\ \frac{\partial}{\partial m_{t,d}} L(\theta) \end{bmatrix} = \mathbf{0} \quad \Rightarrow \quad \mathbf{m}_t = \frac{\sum_{l=1}^k [\alpha'_{l-1}(L(t)) \ p'_t \mathcal{N}'_t(\mathbf{y}_l) \ \beta'_l(R(t))] \mathbf{y}_l}{\sum_{l=1}^k [\alpha'_{l-1}(L(t)) \ p'_t \mathcal{N}'_t(\mathbf{y}_l) \ \beta'_l(R(t))]}, \quad (6)$$

where, again, the parameters p'_t and \mathcal{N}'_t correspond to θ' , and $\alpha'_*(\cdot)$ and $\beta'_*(\cdot)$ are the forward- and backward-probabilities computed using θ' on a k -stage trellis, with the null arcs going between vertically aligned states and only non-null arcs traversing left-to-right. Recall that

$$\frac{\alpha'_{l-1}(L(t)) p'_t \mathcal{N}'_t(\mathbf{y}_l) \beta'_l(R(t))}{P_{\theta'}(\mathbf{Y})} = P_{\theta'}(t_l = t | \mathbf{Y}) = P_{\theta'}(\mathbf{y}_l \text{ was emitted by arc } t | \mathbf{Y}).$$

The value of \mathbf{m}_t that maximizes $L(\theta)$ in (6) may therefore be seen as a “sample mean,” where each observation \mathbf{y}_l in the sample has a fractional count — the probability that it came from arc t — and the sample size is the *expected* number of times the arc t was traversed, or as a “weighted mean,” where the weight of the sample \mathbf{y}_l is the probability that it was emitted from t .

To find the \mathbf{U}_t that maximizes $L(\theta)$, let $\mathbf{V}_t = \mathbf{U}_t^{-1}$, and $v_{t,ij}$ denote the ij -th element of \mathbf{V}_t .

$$\begin{aligned} \frac{\partial}{\partial v_{t,ij}} L(\theta) &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t} | \mathbf{Y}) \frac{\partial}{\partial v_{t,ij}} \log P_{\theta'}(\mathbf{t}, \mathbf{Y}) - \frac{\partial}{\partial v_{t,ij}} \sum_{s \in \mathcal{S}} \lambda_s \sum_{t': L(t')=s} p_{t'} \\ &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t} | \mathbf{Y}) \frac{\partial}{\partial v_{t,ij}} \log \left[\prod_{l=1}^{n(\mathbf{t})} p_{t_l} \mathcal{N}_{t_l}(\mathbf{y}_l) \right] - 0 \\ &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t} | \mathbf{Y}) \frac{\partial}{\partial v_{t,ij}} \left[\sum_{l=1}^{n(\mathbf{t})} \log p_{t_l} + \sum_{l=1}^{n(\mathbf{t})} \log \mathcal{N}_{t_l}(\mathbf{y}_l) \right] \\ &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t} | \mathbf{Y}) \left[0 + \sum_{l: t_l=t} \frac{\partial}{\partial v_{t,ij}} \log \mathcal{N}_t(\mathbf{y}_l) \right] \\ &= \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t} | \mathbf{Y}) \sum_{l: t_l=t} \frac{\partial}{\partial v_{t,ij}} \left\{ -\frac{1}{2} (\mathbf{y}_l - \mathbf{m}_t)^T \mathbf{U}_t^{-1} (\mathbf{y}_l - \mathbf{m}_t) - \log (2\pi)^{\frac{d}{2}} \sqrt{|\mathbf{U}_t|} \right\} \\ &= -\frac{1}{2} \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t} | \mathbf{Y}) \sum_{l: t_l=t} \frac{\partial}{\partial v_{t,ij}} \{ (\mathbf{y}_l - \mathbf{m}_t)^T \mathbf{U}_t^{-1} (\mathbf{y}_l - \mathbf{m}_t) + \log |\mathbf{U}_t| \} \\ &= -\frac{1}{2} \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t} | \mathbf{Y}) \sum_{l: t_l=t} \frac{\partial}{\partial v_{t,ij}} \{ (\mathbf{y}_l - \mathbf{m}_t)^T \mathbf{V}_t (\mathbf{y}_l - \mathbf{m}_t) - \log |\mathbf{V}_t| \}. \quad (7) \end{aligned}$$

(IV) Show that for any $d \times 1$ vector $\mathbf{x} = [x_1 \dots x_d]^T$ and symmetric positive-definite $d \times d$ matrix \mathbf{A} , the partial derivatives of the *scalar* $\mathbf{x}^T \mathbf{A} \mathbf{x}$ with respect to the components of the matrix \mathbf{A} are

$$\begin{bmatrix} \frac{\partial}{\partial a_{11}} \mathbf{x}^T \mathbf{A} \mathbf{x} & \dots & \frac{\partial}{\partial a_{1d}} \mathbf{x}^T \mathbf{A} \mathbf{x} \\ \vdots & \frac{\partial}{\partial a_{ij}} \mathbf{x}^T \mathbf{A} \mathbf{x} & \vdots \\ \frac{\partial}{\partial a_{d1}} \mathbf{x}^T \mathbf{A} \mathbf{x} & \dots & \frac{\partial}{\partial a_{dd}} \mathbf{x}^T \mathbf{A} \mathbf{x} \end{bmatrix} = \mathbf{x} \mathbf{x}^T,$$

and the partial derivatives of the *scalar* $\log |\mathbf{A}|$ with respect to the components of \mathbf{A} are

$$\begin{bmatrix} \frac{\partial}{\partial a_{11}} \log |\mathbf{A}| & \dots & \frac{\partial}{\partial a_{1d}} \log |\mathbf{A}| \\ \vdots & \frac{\partial}{\partial a_{ij}} \log |\mathbf{A}| & \vdots \\ \frac{\partial}{\partial a_{d1}} \log |\mathbf{A}| & \dots & \frac{\partial}{\partial a_{dd}} \log |\mathbf{A}| \end{bmatrix} = \mathbf{A}^{-1}.$$

(V) Combine the results above with (7) to show that the partial derivative of $L(\theta)$ w.r.t. \mathbf{V}_t is

$$\begin{bmatrix} \frac{\partial}{\partial v_{t,11}} L(\theta) & \dots & \frac{\partial}{\partial v_{t,1d}} L(\theta) \\ \vdots & \frac{\partial}{\partial v_{t,ij}} L(\theta) & \vdots \\ \frac{\partial}{\partial v_{t,d1}} L(\theta) & \dots & \frac{\partial}{\partial v_{t,dd}} L(\theta) \end{bmatrix} = -\frac{1}{2} \sum_{\text{All paths } \mathbf{t}} P_{\theta'}(\mathbf{t}|\mathbf{Y}) \sum_{l:t_l=t} \{(\mathbf{y}_l - \mathbf{m}_t)(\mathbf{y}_l - \mathbf{m}_t)^T - \mathbf{V}_t^{-1}\}$$

and show that the choice of \mathbf{V}_t (equivalently \mathbf{U}_t) that makes $\frac{\partial}{\partial v_{t,ij}} L(\theta) = 0$ for all i and j is

$$\begin{aligned} \mathbf{U}_t &= \frac{1}{\sum_{l=1}^k P_{\theta'}(t_l = t, \mathbf{Y})} \sum_{l=1}^k P_{\theta'}(t_l = t, \mathbf{Y}) (\mathbf{y}_l - \mathbf{m}_t)(\mathbf{y}_l - \mathbf{m}_t)^T \\ &= \frac{\sum_{l=1}^k [\alpha'_{l-1}(L(t)) \ p'_t \mathcal{N}'_t(\mathbf{y}_l) \ \beta'_l(R(t))]}{\sum_{l=1}^k [\alpha'_{l-1}(L(t)) \ p'_t \mathcal{N}'_t(\mathbf{y}_l) \ \beta'_l(R(t))]} (\mathbf{y}_l - \mathbf{m}_t)(\mathbf{y}_l - \mathbf{m}_t)^T, \end{aligned} \quad (8)$$

where, once again, the parameters p'_t and \mathcal{N}'_t correspond to θ' , and $\alpha'_*(\cdot)$ and $\beta'_*(\cdot)$ are the forward- and backward-probabilities computed using θ' on a k -stage trellis, with the null arcs going between vertically aligned states and only non-null arcs traversing left-to-right.

The inverse \mathbf{V}_t of the covariance matrix is sometimes called the *precision matrix*, and is often of interest in multivariate statistics and *factor analysis*.

(VI) Verify that the updated matrix \mathbf{U}_t of (8) is symmetric and positive-(semi)definite, thereby justifying why the Lagrangian of (4) did not impose any constraints on the components of the parameter set θ corresponding to the \mathbf{U}_t 's.

Initializations and Local Maxima in the E-M Algorithm

This problem brings out the fact that the E-M algorithm, at best, converges to a local maximum of the likelihood, and that the point of convergence depends on the starting point. You will have to write a computer program to perform the numerical computations necessary to solve this problem.

Consider a 1-state HMM (!) that produces real-valued outputs y_1, y_2, \dots, y_n , and denote the lone state by s . Let the HMM have two output-producing arcs t_1 and t_2 , both from s to s , which may be taken with probabilities $p(t_1) = \alpha_1$ and $p(t_2) = \alpha_2 = (1 - \alpha_1)$. Let the output densities be Gaussian.

$$\begin{aligned} \mathcal{N}(y|t_1) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(y-\mu_1)^2}{2\sigma_1^2}}, \\ \mathcal{N}(y|t_2) &= \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}. \end{aligned}$$

Let $\theta = [\alpha_1 \ \mu_1 \ \sigma_1^2 \ \mu_2 \ \sigma_2^2]$ denote the parameters of the HMM.

1. Write down the formula for $P_\theta(y_1, \dots, y_n)$. (Hint: this HMM is just a mixture density.)
2. Plot $P_\theta(y_1)$ for $\theta = [\frac{1}{3} \ -2 \ 1 \ 2 \ 1]$.

- Implement the E-M algorithm for updating θ for this HMM. Make your implementation flexible, so that some parameters may be tied. E.g. we will experiment with arbitrarily fixing the variances and estimating only the means and α_1 , or with setting a minimum (floor) on the estimate of the variance, etc.
- Let the observed data comprise the following $n = 25$ values.

+0.608	-1.590	+0.235	+3.949	-2.249
+2.704	-2.473	+0.672	+0.262	+1.072
-1.773	+0.537	+3.240	+2.400	-2.499
+2.608	-3.458	+0.257	+2.569	+1.415
+1.410	-2.653	+1.396	+3.286	-0.712

Set $\sigma_1^2 = \sigma_2^2 = 1$ and $\alpha_1 = \frac{1}{3}$, and obtain the sequence of iterates $[\mu_1^{(k)}, \mu_2^{(k)}]$ starting with different values $[\mu_1^{(0)}, \mu_2^{(0)}]$. Try several different starting points in each of the 4 quadrants in the region $[-4, +4] \times [-4, +4] \subset \mathbb{R}^2$.

Plot the “trajectory” of the parameters in the x - y plane for a few illustrative starting points.

- How many different points of convergence did you encounter in Step 4? Comment on the value of the (local) maxima at these points, and speculate upon their *regions of attraction*. By region of attraction of a point of convergence, we mean the *set* of starting points from which the E-M algorithm converges to that particular point.
- Use a program such as MATLAB or R to plot the surface corresponding to the likelihood of the 25 data-points of Step 4 as a function of $[\mu_1, \mu_2] \in [-4, +4] \times [-4, +4]$, with α_1, σ_1^2 and σ_2^2 fixed as above.

Comment on the trajectories you plotted in Step 5 in light of this surface plot, and revise your speculation about the regions of attraction if necessary.

- Repeat Step 4 but this time, make α_1 also a free parameter. In particular, choose the same initial values of $[\mu_1^{(0)}, \mu_2^{(0)}]$ that you used to plot the illustrative trajectories of $[\mu_1^{(k)}, \mu_2^{(k)}]$ in Step 4, and plot the (new) trajectories when $\alpha_1^{(0)} = \frac{1}{3}$, but it is updated at every iteration instead of being held fixed. Comment on the point of convergence, as well as the (locally) maximum value of likelihood attained, when the HMM has more free parameters.

What happens when you start at $[\mu_1^{(0)}, \mu_2^{(0)}]$ but an $\alpha_1^{(0)}$ that is different from $\frac{1}{3}$?

- Let all five parameters be free, and compute the maximum likelihood that is attained for one or two different starting points.