THE EM ALGORITHM

LINUS TANG

1. Introduction

Often we observe data drawn from a distribution parameterized by some unknowns and want to estimate the unknown parameters. The Expectation Maximization (EM) algorithm is useful for situations in which the observed data is incomplete. The missing data is called latent information. The latent information can be used to assist in finding the maximum likelihood estimator of the parameters given the observed data.

While finding a closed form answer is often impossible or intractable, the EM algorithm approximates the answer numerically through the following process:

(1) Start with an initial guess for $\hat{\boldsymbol{x}}_0$ the parameters.
(2) Compute the distribution $p_{\mathsf{s}|\mathsf{y}}(\cdot \mid \boldsymbol{y}; \hat{\boldsymbol{x}}_0)$ of the latent information based on the (the observed data + that guess). Here, the latent information is denoted by s and the observed data is denoted by $\boldsymbol{y}$.
(3) Temporarily taking this distribution as ground truth, compute the expected value over of the logarithmic likelihood of an estimator $\hat{\boldsymbol{x}} = \boldsymbol{a}$ of the parameter, in terms of $\boldsymbol{a}$. (expectation step, or E-step)
(4) Take $\hat{\boldsymbol{x}}_1$ to be the estimator that maximizes this expected value. (maximization step, or M-step)
(5) Now, $\hat{\boldsymbol{x}}_1$ is a refined version of $\hat{\boldsymbol{x}}_0$. Apply steps 2 through 4 repeatedly to get a sequence of refinements, $\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \dots,$ of estimates of the parameters. Stop when the refinements become negligible.

Under certain conditions which will be described later, we can prove that this sequence of estimators converges to the maximum likelihood estimator of the parameters!

2. The Algorithm

We describe the EM Algorithm more formally here.

We observe data y drawn from a distribution $p_{\mathsf{y}}(\cdot; \boldsymbol{x})$ from a family parameterized by $\mathsf{x} \in \mathcal{X}$. The EM algorithm aims to numerically approximate the maximum likelihood estimator

$$\hat{\boldsymbol{x}}_{ML} = \operatorname*{argmax}_{\boldsymbol{x} \in \mathcal{X}} p_{\mathsf{y}}(\boldsymbol{y}; \boldsymbol{x}).$$

The EM algorithm is applicable when there is a variable s (the latent information) such that the joint distribution of (y, s) (the complete information) has a "nicer structure" than the distribution of y itself. We will see examples of this later.

The algorithm now follows the following steps.

- Choose an arbitrary estimate $\hat{\boldsymbol{x}}_0$ of the parameter. Recursively compute the sequence $\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \dots$ of estimators as follows.
- (E-step) Compute the expectation

$$U(\boldsymbol{a}; \hat{\boldsymbol{x}}_\ell) = \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}}_\ell)}[\log(p_{\mathsf{y},\mathsf{s}}(\boldsymbol{y}, \mathsf{s}; \boldsymbol{a}))].$$

- (M-step) Compute the estimator which maximizes the above expectation, namely
$$\hat{\boldsymbol{x}}_{\ell+1} = \operatorname*{argmax}_{a \in \mathcal{X}} U(\boldsymbol{a}; \hat{\boldsymbol{x}}_\ell).$$
- Under conditions which will be described later, the sequence $\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots$ of estimators generated by repeated applications of the E-step and M-step will converge to $\hat{\boldsymbol{x}}_{ML}$.

3. Example and Analysis

Let $\boldsymbol{x} = (x_1, x_2, x_3) \in \mathcal{X} = \mathbb{R}^3$ and consider the Gaussian mixture consisting of three normal distributions with variance 1 and means x_1, x_2, x_3 with respective weights $1 : 2 : 3$. In particular, this defines a family of probability distributions parameterized by $\boldsymbol{x}$, given by the probability density function

$$p_{\mathsf{y}}(y; \boldsymbol{x}) = \frac{1}{6}\left(\sum_{i=1}^{3} \frac{i}{\sqrt{2\pi}} e^{(y-x_i)^2/2}\right).$$

Suppose that we observe the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ which consists of n independent samples y_i from $p_{\mathsf{y}}(\cdot; \boldsymbol{x})$. We want to estimate

$$\hat{\boldsymbol{x}}_{ML} = \operatorname*{argmax}_{\boldsymbol{x} \in \mathcal{X}} p_{\boldsymbol{\mathsf{y}}}(\boldsymbol{y}; \boldsymbol{x}).$$

While we can write the likelihood function $p_{\boldsymbol{\mathsf{y}}}(\boldsymbol{y}; \boldsymbol{x})$ as a product of n probability densities, maximizing this product over $\boldsymbol{n}$ directly, such as by solving for critical points, is intractable for large n. We turn to the EM algorithm to get a numerical approximation of this estimator.

Since each y_i is drawn from a Gaussian mixture, we can model it by first choosing one of the constituent Gaussian distributions to draw it from. In this case, we can write

$$(\mathsf{y}_i \mid \mathsf{s}_i) \sim \mathcal{N}(x_{\mathsf{s}_i}, 1),$$

where

$$\mathbb{P}(\mathsf{s}_i = 1) = \frac{1}{6}, \mathbb{P}(\mathsf{s}_i = 2) = \frac{2}{6}, \mathbb{P}(\mathsf{s}_i = 3) = \frac{3}{6},$$

and the s_i are independent.

Loosely speaking, this latent information $\boldsymbol{\mathsf{s}} = (\mathsf{s}_1, \ldots, \mathsf{s}_n)$ is useful to work with because the probability distributions $p_{\boldsymbol{\mathsf{y}}|\boldsymbol{\mathsf{s}}}$ and $p_{\boldsymbol{\mathsf{s}}}$ are both nice.

We now perform the E-step. We have

$$\begin{aligned} p_{\mathsf{y}_i, \mathsf{s}_i}(y_i, s_i; \boldsymbol{x}) &= p_{\mathsf{y}_i|\mathsf{s}_i}(y_i \mid s_i; \boldsymbol{x}) p_{\mathsf{s}_i}(s_i; \boldsymbol{x}) \\ &= \frac{1}{\sqrt{2\pi}} e^{(y_i - x_{s_i})^2/2} \cdot \frac{s_i}{6} \end{aligned}$$

and

$$p_{\mathsf{s}_i|\boldsymbol{\mathsf{y}}}(s_i \mid \boldsymbol{y}; \boldsymbol{x}) = \frac{p_{\boldsymbol{\mathsf{y}}, \mathsf{s}_i}(\boldsymbol{y}, s_i; \boldsymbol{x})}{p_{\boldsymbol{\mathsf{y}}}(\boldsymbol{y}; \boldsymbol{x})} = \frac{\frac{1}{\sqrt{2\pi}} e^{(y_i - x_{s_i})^2/2} \cdot \frac{s_i}{6}}{\sum_{s=1}^{3} \left(\frac{1}{\sqrt{2\pi}} e^{(y_i - x_s)^2/2} \cdot \frac{s}{6}\right)}. \tag{1}$$

We now perform the M-step. We want to choose $\boldsymbol{a}$ to maximize the expected value of $\log(p_{\boldsymbol{\mathsf{y}}, \boldsymbol{\mathsf{s}}}(\boldsymbol{y}, \boldsymbol{\mathsf{s}}); \boldsymbol{a})$ when $\boldsymbol{\mathsf{s}}$ is distributed according to equation 1.

We have

$$\begin{aligned}
\hat{\boldsymbol{x}}_{\ell+1} &= \operatorname*{argmax}_{\boldsymbol{a}} \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}}_\ell)}[\log(p_{\mathsf{y},\mathsf{s}}(\boldsymbol{y},\mathbf{s});\boldsymbol{a})] \\
&= \operatorname*{argmax}_{\boldsymbol{a}} \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}}_\ell)}\left[\log \prod_{i=1}^{n} e^{(a_{\mathsf{s}_i}-y_i)^2}\right] \\
&= \operatorname*{argmax}_{\boldsymbol{a}} \sum_{i=1}^{n} \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}}_\ell)}[(a_{\mathsf{s}_i}-y_i)^2] \\
&= \operatorname*{argmax}_{\boldsymbol{a}} \sum_{s=1}^{3}\sum_{i=1}^{n} p_{\mathsf{s}_i|\mathsf{y}}(s \mid \boldsymbol{y};\hat{\boldsymbol{x}}_\ell)\mathbb{E}[(a_s-y_i)^2].
\end{aligned}$$

We can choose a_1, a_2, a_3 individually to minimize the contributions from $s = 1, 2, 3$, respectively. We get that

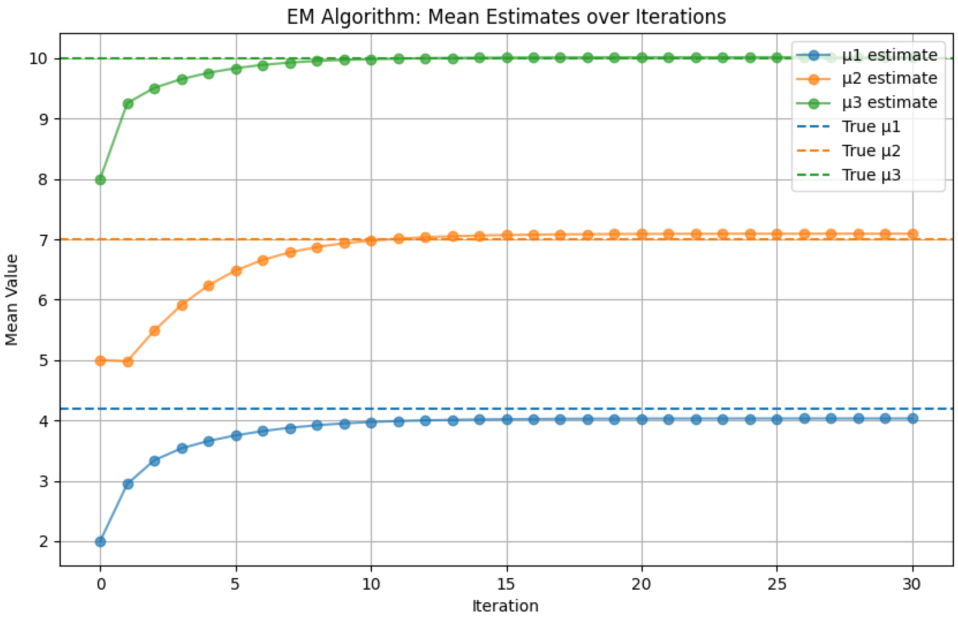
$$\hat{x}_{\ell,s} = \frac{\sum_{i=1}^{n} p_{\mathsf{s}_i|\mathsf{y}}(s \mid \boldsymbol{y};\hat{\boldsymbol{x}}_\ell)y_i}{\sum_{i=1}^{n} p_{\mathsf{s}_i|\mathsf{y}}(s \mid \boldsymbol{y};\hat{\boldsymbol{x}}_\ell)}.$$

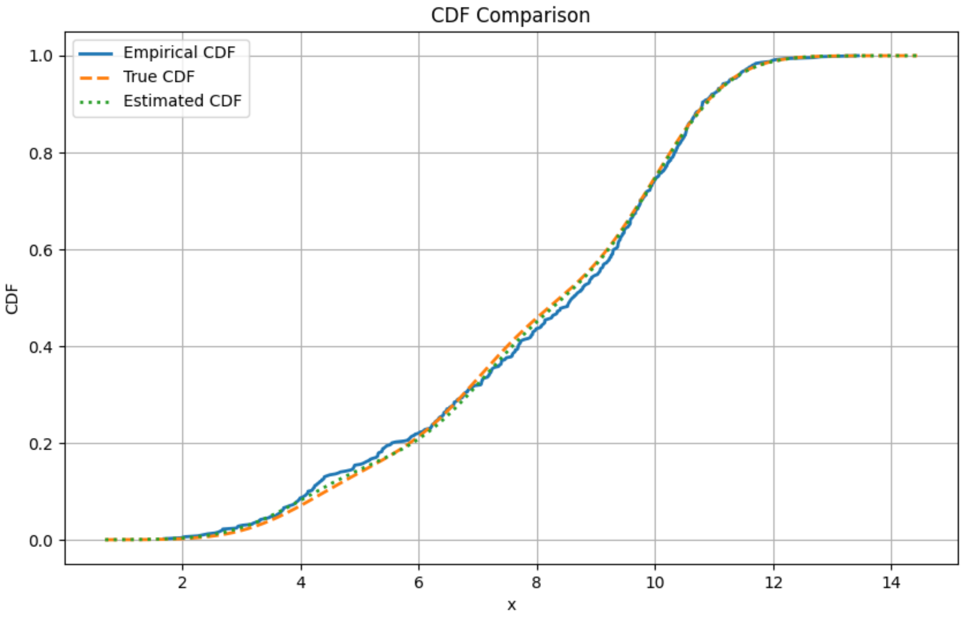
And we can calculate $p_{\mathsf{s}_i|\mathsf{y}}(s \mid \boldsymbol{y};\hat{\boldsymbol{x}}_\ell)$ by equation 1.

Repeating the E-step and M-step gives a series of estimators that converges to the maximum likelihood estimator.

Remark. With some work, the EM algorithm can also handle a much more general problem, in which Gaussian distributions are instead multivariate normal distributions, and their mean vectors, covariance matrices, and weights in the mixture are all unknown parameters!

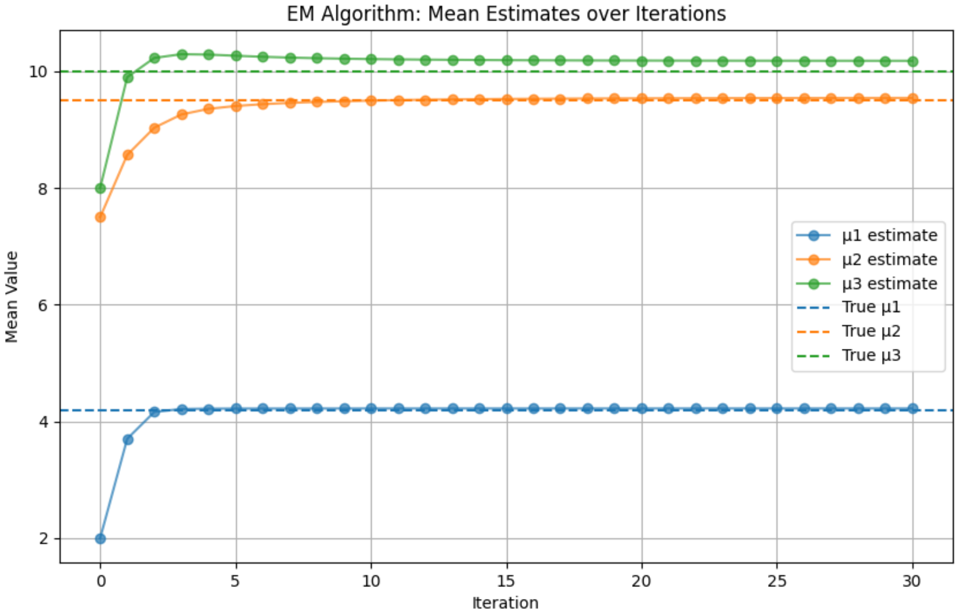
The graphs below show the convergence of estimators found by the EM algorithm run on this model, where the observed data consists of $n = 500$ samples, and where the true means are $(x_1, x_2, x_3) = (4.2, 7.0, 10.0)$. Within just 15 iterations, the EM algorithm is quite close to its stable configuration.

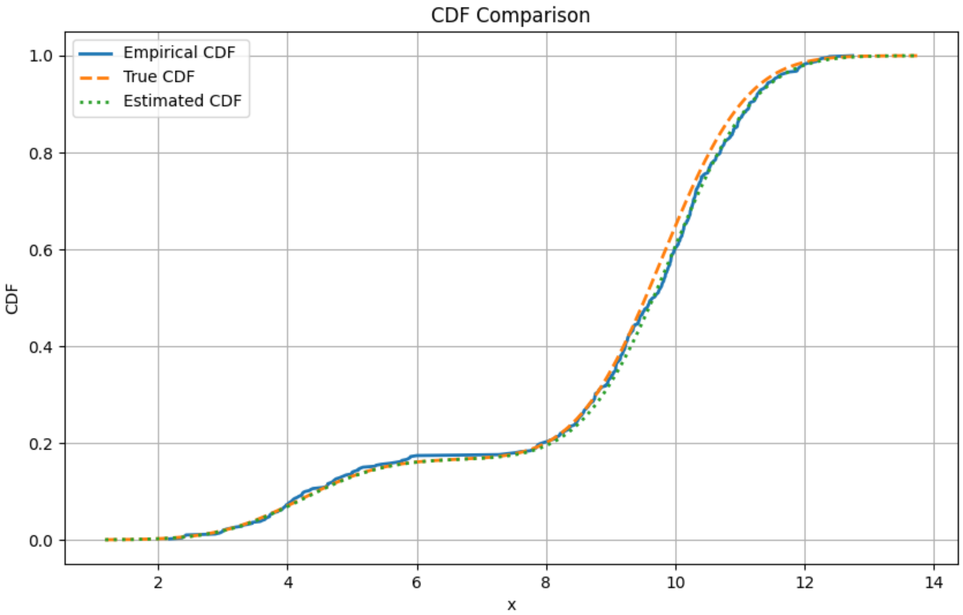




The rate of convergence can depend on the true parameters, as shown by running the same algorithm with means $(x_1, x_2, x_3) = (4.2, 9.0, 10.0)$, results graphed below. In each case, we have set the initial guess $\hat{\boldsymbol{x}}_0$ to have each mean approximately 2 less than the true mean, so that rates of convergence can be visually compared more easily.

One thing to note is that the estimators are converging to the maximum likelihood estimator rather than the true values of the parameters. (This is natural because the former is a feature of the observed data and the latter is not.) In fact, we can see that the implied CDF of the EM estimate fits that of the observed data better than the true CDF does.





4. Intuition and conditions for Convergence

Here we give intuition for why we might expect the EM algorithm to converge to the maximum likelihood estimator for well-behaved families of probability distributions. Among other things, we will show that the likelihood increases at each step (unless a fixed point has been reached).

Proposition 1. The log likelihood of the estimator increases at each iteration (unless a fixed point has been reached).

Proposition 2. The maximum likelihood estimator is a fixed point of the algorithm. Symbolically,

$$\hat{\boldsymbol{x}}_{ML} = \underset{a \in \mathcal{X}}{\operatorname{argmax}}\, U(\boldsymbol{a}; \hat{\boldsymbol{x}}_{ML}).$$

Proofs of these propositions are saved for the end of the section. These two properties of the EM algorithm make it plausible that it converges to the maximum likelihood estimator for a broad class of models.

Note, however, that the algorithm may have multiple fixed points, in which case the EM algorithm is not guaranteed to converge to the maximum likelihood estimator. Specifically, some but not necessarily all fixed points of the algorithm are stationary points of the likelihood $p_{\mathbf{y}}(\boldsymbol{y}; \cdot)$ (i.e. a point where all of its first-order partial derivatives are 0), some but not necessarily all stationary points of the likelihood are local maxima, and some but not necessarily all local maxima are global maxima.

Hence, the analysis of the EM algorithm can get messy. As a practical note, these "traps" often can be avoided by running the EM algorithm with many different initial estimates, then choosing the result with the greatest likelihood.

Here we summarize sufficient conditions for various properties of the limit of the EM algorithm, proven by C.F. Jeff Wu.

(1) If the likelihood $p_{\mathbf{y}}(\boldsymbol{y}; \cdot)$ is bounded above, then $p_{\mathbf{y}}(\boldsymbol{y}; \hat{\boldsymbol{x}}_\ell)$ converges.
(2) If $U(\hat{\boldsymbol{x}}'; \hat{\boldsymbol{x}})$ is continuous in $\hat{\boldsymbol{x}}'$ and $\hat{\boldsymbol{x}}$, then $p_{\mathbf{y}}(\boldsymbol{y}; \hat{\boldsymbol{x}}_\ell)$ converges to a stationary value.

(3) Note that (1) and (2) are statements about convergence in likelihood, rather than about convergence of the estimators themselves. If in addition to (2) we have $\lim_{\ell\to\infty}\|\hat{\boldsymbol{x}}_{\ell+1}-\hat{\boldsymbol{x}}_\ell\|=0$, then the estimators converge to a stationary point.

(4) If there exists a function $\sigma:\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ such that

$$(\forall \hat{\boldsymbol{x}}', \hat{\boldsymbol{x}} \in \mathcal{X})(U(\hat{\boldsymbol{x}}';\hat{\boldsymbol{x}}) - U(\hat{\boldsymbol{x}};\hat{\boldsymbol{x}}) \geq \sigma(\|\hat{\boldsymbol{x}}' - \hat{\boldsymbol{x}}\|))$$

and $(\forall\{t_k\}\in\mathbb{R}_{\geq 0})(\lim_{k\to\infty}\sigma(t_k)=0 \Rightarrow \lim_{k\to\infty}t_k=0)$, then (3) is satisfied.

(5) If the likelihood has a unique maximum which is its only stationary point, and the partial derivatives of $U(\hat{\boldsymbol{x}}';\hat{\boldsymbol{x}})$ with respect to the components of $\hat{\boldsymbol{x}}'$ are continuous in $\hat{\boldsymbol{x}}'$ and $\hat{\boldsymbol{x}}$, then the EM algorithm converges to the maximum likelihood estimator.

Proofs of Propositions 1 and 2.

By definition,

$$\hat{\boldsymbol{x}}_{\ell+1} = \operatorname*{argmax}_{\boldsymbol{\alpha}\in\mathcal{X}} U(\boldsymbol{\alpha};\hat{\boldsymbol{x}}_\ell),$$

so

$$U(\hat{\boldsymbol{x}}_{\ell+1};\hat{\boldsymbol{x}}_\ell) \geq U(\hat{\boldsymbol{x}}_\ell;\hat{\boldsymbol{x}}_\ell).$$

Now we have

$$\begin{aligned}\log p_{\mathsf{y}}(\boldsymbol{y};\hat{\boldsymbol{x}}_{\ell+1}) &= U(\hat{\boldsymbol{x}}_{\ell+1};\hat{\boldsymbol{x}}_\ell) + D_{\mathrm{KL}}\left(p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}}_\ell)\parallel p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}}_{\ell+1})\right)\\ &\geq U(\hat{\boldsymbol{x}}_{\ell+1};\hat{\boldsymbol{x}}_\ell)\\ &\geq U(\hat{\boldsymbol{x}}_\ell;\hat{\boldsymbol{x}}_\ell)\\ &= \log p_{\mathsf{y}}(\boldsymbol{y};\hat{\boldsymbol{x}}_\ell),\end{aligned}$$

so the likelihood of the estimator is nondecreasing, proving Proposition 1.

Proposition 2 follows quickly. If $\hat{\boldsymbol{x}}_\ell = \arg\max_{\boldsymbol{x}}\log p_{\mathsf{y}}(\boldsymbol{y};\boldsymbol{x})$, then from Proposition 1 we have

$$\log p_{\mathsf{y}}(\boldsymbol{y};\hat{\boldsymbol{x}}_{\ell+1}) \geq \log p_{\mathsf{y}}(\boldsymbol{y};\hat{\boldsymbol{x}}_\ell) = \max_{\boldsymbol{x}}\log p_{\mathsf{y}}(\boldsymbol{y};\boldsymbol{x})$$

so

$$\hat{\boldsymbol{x}}_{\ell+1} = \arg\max_{\boldsymbol{x}}\log p_{\mathsf{y}}(\boldsymbol{y};\boldsymbol{x}),$$

as desired.

5. Simplifications for Exponential Families

When $p_{\mathsf{y},\mathsf{s}}(\cdot,\cdot;\cdot)$ is an exponential family, i.e.

$$p_{\mathsf{y},\mathsf{s}}(\boldsymbol{y},\boldsymbol{s};\boldsymbol{x}) = \exp\left(\sum_{i=1}^{K} x_i t(\boldsymbol{y},\boldsymbol{s}) - \alpha(\boldsymbol{x}) + \beta(\boldsymbol{y},\boldsymbol{s})\right),$$

the expressions involved in the EM algorithm simplify nicely. (Here, we have written the exponential family in terms of its natural parameter $\boldsymbol{x}$.) The key property is that

$$\begin{aligned}
\frac{\partial}{\partial x_k}\alpha(\boldsymbol{x}) &= \frac{\partial}{\partial x_k}\log\left(\int \exp\left(\sum_{i=1}^{K} x_i t(\boldsymbol{y},\boldsymbol{s}) + \beta(\boldsymbol{y},\boldsymbol{s})\right) d\boldsymbol{y} d\boldsymbol{s}\right) \\
&= \frac{\frac{\partial}{\partial x_k}\int \exp\left(\sum_{i=1}^{K} x_i t(\boldsymbol{y},\boldsymbol{s}) + \beta(\boldsymbol{y},\boldsymbol{s})\right) d\boldsymbol{y} d\boldsymbol{s}}{\int \exp\left(\sum_{i=1}^{K} x_i t(\boldsymbol{y},\boldsymbol{s}) + \beta(\boldsymbol{y},\boldsymbol{s})\right) d\boldsymbol{y} d\boldsymbol{s}} \\
&= \frac{\int \frac{\partial}{\partial x_k}\exp\left(\sum_{i=1}^{K} x_i t(\boldsymbol{y},\boldsymbol{s}) + \beta(\boldsymbol{y},\boldsymbol{s})\right) d\boldsymbol{y} d\boldsymbol{s}}{\int \exp\left(\sum_{i=1}^{K} x_i t(\boldsymbol{y},\boldsymbol{s}) + \beta(\boldsymbol{y},\boldsymbol{s})\right) d\boldsymbol{y} d\boldsymbol{s}} \qquad (2) \\
&= \frac{\int t_k(\boldsymbol{y},\boldsymbol{s}) \exp\left(\sum_{i=1}^{K} x_i t(\boldsymbol{y},\boldsymbol{s}) + \beta(\boldsymbol{y},\boldsymbol{s})\right) d\boldsymbol{y} d\boldsymbol{s}}{\int \exp\left(\sum_{i=1}^{K} x_i t(\boldsymbol{y},\boldsymbol{s}) + \beta(\boldsymbol{y},\boldsymbol{s})\right) d\boldsymbol{y} d\boldsymbol{s}} \\
&= \mathbb{E}_{p_{\mathsf{s},\mathsf{y}}(\cdot,\cdot;\boldsymbol{x})}[t_k(\mathsf{y},\mathsf{s})].
\end{aligned}$$

In particular,

$$\begin{aligned}
U(\boldsymbol{x};\boldsymbol{x}') &= \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\boldsymbol{x}')}[\log(p_{\mathsf{y},\mathsf{s}}(\boldsymbol{y},\mathsf{s};\boldsymbol{x}))] \\
&= \left(\sum_{i=1}^{K} x_i \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\boldsymbol{x}')}[t_i(\boldsymbol{y},\mathsf{s})]\right) - \alpha(\boldsymbol{x}) + \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\boldsymbol{x}')}[\beta(\boldsymbol{y},\mathsf{s})],
\end{aligned}$$

so

$$\begin{aligned}
\frac{\partial}{\partial x_k}U(\boldsymbol{x};\boldsymbol{x}') &= \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\boldsymbol{x}')}[t_k(\boldsymbol{y},\mathsf{s})] - \frac{\partial}{\partial x_k}\alpha(\boldsymbol{x}) \\
&= \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\boldsymbol{x}')}[t_k(\boldsymbol{y},\mathsf{s})] - \mathbb{E}_{p_{\mathsf{s},\mathsf{y}}(\cdot,\cdot;\boldsymbol{x})}[t_k(\mathsf{y},\mathsf{s})].
\end{aligned}$$

Since we defined

$$\hat{\boldsymbol{x}}_{\ell+1} = \operatorname*{argmax}_{\boldsymbol{x}} U(\boldsymbol{x};\hat{\boldsymbol{x}}_\ell),$$

it follows that

$$\frac{\partial}{\partial \hat{x}_{\ell+1,k}}U(\hat{\boldsymbol{x}}_{\ell+1};\hat{\boldsymbol{x}}) = 0,$$

so the EM algorithm updates its estimate according to

$$\mathbb{E}_{p_{\mathsf{s},\mathsf{y}}(\cdot,\cdot;\hat{\boldsymbol{x}}_{\ell+1})}[t_k(\mathsf{y},\mathsf{s})] = \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}}_\ell)}[t_k(\boldsymbol{y},\mathsf{s})].$$

From this identity, we show as follows that every fixed point of the EM algorithm is a stationary point of the likelihood $p_{\mathsf{y}}(\boldsymbol{y};\cdot)$.

Indeed, any fixed point $\hat{\boldsymbol{x}}$ of the EM algorithm must satisfy

$$\mathbb{E}_{p_{\mathsf{s},\mathsf{y}}(\cdot,\cdot;\hat{\boldsymbol{x}})}[t_k(\mathsf{y},\mathsf{s})] = \mathbb{E}_{p_{\mathsf{s}|\mathsf{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}})}[t_k(\boldsymbol{y},\mathsf{s})]$$

for $k = 1,\ldots,K$.

Note that the conditional distribution of the complete data on the observed data is also exponential. In particular,

$$p_{\mathsf{s}|\mathsf{y}}(\boldsymbol{s}|\boldsymbol{y};\boldsymbol{x}) = \frac{p_{\mathsf{y},\mathsf{s}}(\boldsymbol{y},\boldsymbol{s};\boldsymbol{x})}{p_{\mathsf{y}}(\boldsymbol{y};\boldsymbol{x})} = \exp\{\boldsymbol{x}^T t(\boldsymbol{y},\boldsymbol{s}) - (\alpha(\boldsymbol{x}) + \log p_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{x})) + \beta(\boldsymbol{y},\boldsymbol{s})\}.$$

It now suffices to show that

$$\frac{\partial}{\partial x_k} \log p_{\mathsf{y}}(\boldsymbol{y}; \boldsymbol{x}) = 0$$

for $k = 1, \dots, K$. Indeed,

$$\begin{aligned}
\left.\frac{\partial}{\partial x_k} \log p_{\boldsymbol{y}}(\boldsymbol{y}; \boldsymbol{x})\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} &= -\left.\frac{\partial}{\partial x_k} \log p_{\boldsymbol{s}|\boldsymbol{y}}(\boldsymbol{s}|\boldsymbol{y}; \boldsymbol{x})\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} + \left.\frac{\partial}{\partial x_k} \log p_{\boldsymbol{y},\boldsymbol{s}}(\boldsymbol{y}, \boldsymbol{s}; \boldsymbol{x})\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \\
&= \left.\frac{\partial}{\partial x_k}(\alpha(\boldsymbol{x}) + \log p_{\boldsymbol{y}}(\boldsymbol{y}; \boldsymbol{x}))\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} - \left.\frac{\partial}{\partial x_k}\alpha(\boldsymbol{x})\right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \\
&= \mathbb{E}_{p_{\boldsymbol{s}|\boldsymbol{y}}(\cdot|\boldsymbol{y};\hat{\boldsymbol{x}})}[t_k(\boldsymbol{y}, \boldsymbol{s})] - \mathbb{E}_{p_{\boldsymbol{y},\boldsymbol{s}}(\cdot,\cdot;\hat{\boldsymbol{x}})}[t_k(\boldsymbol{y}, \boldsymbol{s})] \\
&= 0,
\end{aligned}$$

where in the second-to-last line we apply the key property (2) to both exponential distributions.

So, every fixed point $\hat{\boldsymbol{x}}$ of the EM algorithm is a stationary point of the likelihood function (when the distribution of the complete data lies in an exponential family parameterized by x).

6. References

- P. J. Bickel; K. A. Doksum. Mathematical Statistics: Basic Ideas and Selected Topics
- A. P. Dempster; N. M. Laird; D. B. Rubin. `https://www.ece.iastate.edu/~namrata/EE527_Spring08/Dempster77.pdf`
- M. Haugh `https://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf`
- T. K. Moon `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=543975`
- C. F. J. Wu `https://projecteuclid.org/journals/annals-of-statistics/volume-11/issue-1/On-the-Convergence-Properties-of-the-EM-Algorithm/10.1214/aos/1176346060.full`
- 6.7800 Course Notes, Fall 2025