EECS 553: Machine Learning (ECE) **University of Michigan**

The Expectation-Maximization Algorithm

Winter 2023 **Clayton Scott**

1 The EM Algorithm in General

The EM algorithm is not specific to Gaussian mixture models, and can be used to perform maximum likelihood estimation for a variety of latent variable models. We begin by stating the EM algorithm in a general setting.

Let $\underline{\boldsymbol{X}}$ be the random variables associated to the observed data, and let $\underline{\boldsymbol{x}}$ denote the actual observation. Let $f(\underline{\boldsymbol{x}};\boldsymbol{\theta})$ be the pdf/pmf of $\underline{\boldsymbol{X}}$, where $\boldsymbol{\theta}$ is the parameter vector to be estimated. The objective is to maximize the likelihood

$$L(\boldsymbol{\theta};\underline{\boldsymbol{x}}) := f(\underline{\boldsymbol{x}};\boldsymbol{\theta}),$$

or equivalently the log-likelihood

$$\ell(\boldsymbol{\theta};\underline{\boldsymbol{x}}) := \log f(\underline{\boldsymbol{x}};\boldsymbol{\theta}),$$

with respect to $\boldsymbol{\theta}$.

Let $\underline{\boldsymbol{Z}}$ be denote the latent (unobserved) variables. The random variables $\underline{\boldsymbol{X}}$ and $\underline{\boldsymbol{Z}}$ are assumed to be jointly distributed. Let $f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}};\boldsymbol{\theta})$ denote the conditional pdf/pmf of $\underline{\boldsymbol{Z}}$ given $\underline{\boldsymbol{X}} = \underline{\boldsymbol{x}}$. The *complete data likelihood* is $L(\boldsymbol{\theta};\underline{\boldsymbol{x}},\underline{\boldsymbol{z}}) := f(\underline{\boldsymbol{x}};\boldsymbol{\theta})f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}};\boldsymbol{\theta})$, and the *complete-data log-likelihood* is $\ell(\boldsymbol{\theta};\underline{\boldsymbol{x}},\underline{\boldsymbol{z}}) = \log L(\boldsymbol{\theta};\underline{\boldsymbol{x}},\underline{\boldsymbol{z}})$. The EM algorithm is as follows.

> Initialize $\boldsymbol{\theta}_0$
> $t \leftarrow 0$
> Repeat
> **E-step**: Compute
> $$Q(\boldsymbol{\theta};\boldsymbol{\theta}_t) = \mathbb{E}_{\underline{\boldsymbol{Z}}\sim f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}};\boldsymbol{\theta}_t)}[\ell(\boldsymbol{\theta};\underline{\boldsymbol{x}},\underline{\boldsymbol{Z}})].$$
> **M-step**: Solve
> $$\boldsymbol{\theta}_{t+1} \leftarrow \arg\max_{\boldsymbol{\theta}} \ Q(\boldsymbol{\theta};\boldsymbol{\theta}_t)$$
> $t \leftarrow t+1$
> Until convergence criterion satisfied

The basic idea behind the EM algorithm is that $\ell(\boldsymbol{\theta};\underline{\boldsymbol{x}},\underline{\boldsymbol{z}})$ cannot be computed because $\underline{\boldsymbol{z}}$ is unobserved, and so the uncertainty in $\underline{\boldsymbol{Z}}$ (given $\underline{\boldsymbol{X}} = \underline{\boldsymbol{x}}$) is "averaged out," yielding a computable proxy for the log likelihood. It is important to keep in mind that the expected complete data log-likelihood is distinct from the original log-likelihood. Indeed, the entire reason for the EM algorithm is that direct maximization of the log-likelihood is difficult or intractable, wheras the expected complete-data log-likelihood is can me maximized efficiently in many problems of interest.

An important property of the EM algorithm is the following *ascent* or *monotonicity* property:

Theorem 1. *For each* $t = 0, 1, 2, \ldots$

$$\ell(\boldsymbol{\theta}_{t+1};\ \underline{\boldsymbol{x}}) \geq \ell(\boldsymbol{\theta}_t;\ \underline{\boldsymbol{x}})$$

Note that this result applies to the original likelihood, which is the function we really want to maximize, even though the EM algorithm maximizes a different objective. Figure 1 shows an example of the likelihood $\ell(\boldsymbol{\theta};\underline{\boldsymbol{x}})$ as a function of iteration, illustrating this property.

To prove this result, we will show that EM in a minorize-maximize algorithm.

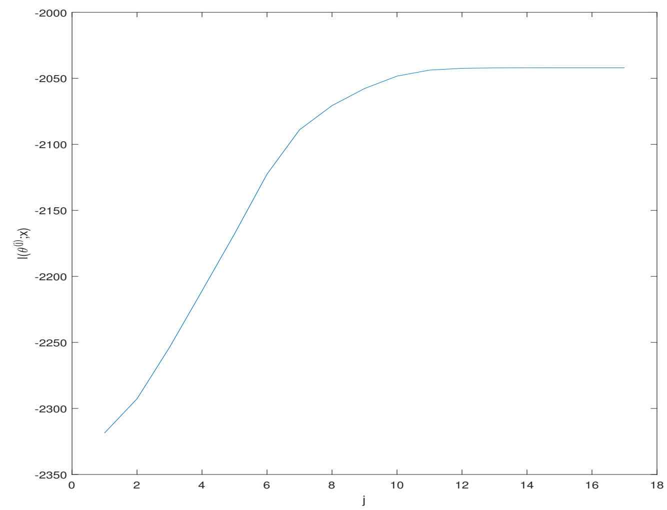

Figure 1: The log-likelihood increases monotonically.

2 EM as an MM Algorithm

We previously studied majorize-minimize algorithms for minimization problems. The idea behind minorize-maximize algorithms is the same idea but applied to maximization problems. Let $J(\boldsymbol{\theta})$ denote the objective function to be maximized. The general minorize-maximize algorithm is as follows.

> Initialize $\boldsymbol{\theta}_0$
> $t \leftarrow 0$
> Repeat
>
> **Minorize:** Find a function $J_t(\boldsymbol{\theta})$ such that
>
> $$\begin{aligned} J(\boldsymbol{\theta}_t) &= J_t(\boldsymbol{\theta}_t) \\ J(\boldsymbol{\theta}) &\geq J_t(\boldsymbol{\theta}) \qquad \forall \boldsymbol{\theta} \end{aligned}$$
>
> **Maxmize:** Solve
>
> $$\boldsymbol{\theta}_{t+1} \leftarrow \arg\max_{\boldsymbol{\theta}} J_t(\boldsymbol{\theta})$$
>
> $t \leftarrow t+1$
>
> Until convergence

MM algorithms satisfy a monotonicity property. The descent property was shown for majorize-minimize algorithms earlier, and so an analogous ascent property automatically holds for minorize-maximize algorithms, because we can always map between maximization and minimization problems by multiplying the objective by -1. Therefore, for any minorize-maximize algorithm, we know that $J(\boldsymbol{\theta}_{t+1}) \geq J(\boldsymbol{\theta}_t)$ for all t.

To establish the ascent property for EM algorithm, it suffices to show that it is an instance of a minorize-maximize algorithm. We will do this by showing that the function

$$J_t(\boldsymbol{\theta}) := Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) + \ell(\boldsymbol{\theta}_t; \underline{\boldsymbol{x}}) - Q(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t) \tag{1}$$

minorizes $J(\boldsymbol{\theta}) := \ell(\boldsymbol{\theta}; \underline{\boldsymbol{x}})$. The details are given in the next section. Noting that only the first term of J_t involves $\boldsymbol{\theta}$, we see that the minorize step is equivalent to the E-step, and the maximize step is the same as the M-step.

3 Proof of Ascent Property ⚠

It is clear from the definitions of J and J_t that $J(\boldsymbol{\theta}_t) = J_t(\boldsymbol{\theta})$, and so it remains to show

$$\ell(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) + \ell(\boldsymbol{\theta}_t; \underline{\boldsymbol{x}}) - Q(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t). \tag{2}$$

for all $\boldsymbol{\theta}$. To prove this inequality, we need the following lemma.

Lemma 1. *Let p and q both be pdfs, or both be pmfs, on $\mathbb{R}^d$. Let $\boldsymbol{Y}$ be a random variable with pdf/pmf p. Then*

$$\mathbb{E}_{\boldsymbol{Y}\sim p}[\log q(\boldsymbol{Y})] \leq \mathbb{E}_{\boldsymbol{Y}\sim p}\,[\log p(\boldsymbol{Y})]$$

and equality is attained iff p and q define the same distribution.

Proof. Jensen's inequality states that for any scalar random variable R and concave function ϕ, $\mathbb{E}[\phi(R)] \leq \phi(\mathbb{E}[R])$ and if ϕ is strictly concave, equality holds iff R is a constant random variable. Assuming that p and q and both pdfs, we apply Jensen's inequality with $R = \log \frac{q(\boldsymbol{Y})}{p(\boldsymbol{Y})}$ and $\phi(r) = \log(r)$, yielding

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Y}\sim p}\left[\log\left(\frac{q(\boldsymbol{Y})}{p(\boldsymbol{Y})}\right)\right] &\leq \log\left[\mathbb{E}_{\boldsymbol{Y}\sim p}\left(\frac{q(\boldsymbol{Y})}{p(\boldsymbol{Y})}\right)\right] \\
&= \log\left(\int \frac{q(\boldsymbol{y})}{p(\boldsymbol{y})}\; p(\boldsymbol{y})\; d\boldsymbol{y}\right) \\
&= \log\left(\int q(\boldsymbol{y})\; d\boldsymbol{y}\right) \\
&= \log(1) \\
&= 0
\end{aligned}$$

Since log is strictly concave, equality hold iff $p(\boldsymbol{y}) = q(\boldsymbol{y})$ almost everywhere, in other words, p and q define the same distribution. The same argument hold for pmfs, replacing integrals with summations. ☐

To show (2), we will apply the lemma with $p(\underline{\boldsymbol{z}}) = f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}}; \boldsymbol{\theta}_t)$ and $q(\underline{\boldsymbol{z}}) = f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}}; \boldsymbol{\theta})$. Thus,

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}) - \ell(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) &= \mathbb{E}_{\underline{\boldsymbol{Z}}\sim f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}};\boldsymbol{\theta}_t)}\left[\log\left(\frac{L(\boldsymbol{\theta}; \underline{\boldsymbol{x}}, \underline{\boldsymbol{Z}})}{f(\underline{\boldsymbol{x}}; \boldsymbol{\theta})}\right)\right] \\
&= \mathbb{E}_{\underline{\boldsymbol{Z}}\sim f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}};\boldsymbol{\theta}_t)}\left[\log\left(f(\underline{\boldsymbol{Z}}|\underline{\boldsymbol{x}}; \boldsymbol{\theta})\right)\right] \\
&\leq \mathbb{E}_{\underline{\boldsymbol{Z}}\sim f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}};\boldsymbol{\theta}_t)}\left[\log\left(f(\underline{\boldsymbol{Z}}|\underline{\boldsymbol{x}}; \boldsymbol{\theta}_t)\right)\right] \\
&= \mathbb{E}_{\underline{\boldsymbol{Z}}\sim f(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}};\boldsymbol{\theta}_t)}\left[\log\left(\frac{L(\boldsymbol{\theta}_t; \underline{\boldsymbol{x}}, \underline{\boldsymbol{Z}})}{f(\underline{\boldsymbol{x}}; \boldsymbol{\theta}_t)}\right)\right] \\
&= Q(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t) - \ell(\boldsymbol{\theta}_t; \underline{\boldsymbol{x}}).
\end{aligned}$$

Exercises

1. (★) Consider two probability distributions on the same domain, either both continuous with pdfs p and q, or both discrete with pmfs p and q. The *Kullback-Leibler divergence* between the two distributions is defined by

   $$D_{KL}(p\|q) := \mathbb{E}_{\boldsymbol{Y}\sim p}\left[\log\left(\frac{p(\boldsymbol{Y})}{q(\boldsymbol{Y})}\right).\right]$$

   Show that $D_{KL}(p\|q) \geq 0$ for all p and q, with equality iff and only if p and q correspond to the same distribution.

2. (★★) Given an alternate proof that J_t in (1) is a minorizer by showing that

$$\ell(\boldsymbol{\theta}; \underline{\boldsymbol{x}}) - \ell(\boldsymbol{\theta}_t; \underline{\boldsymbol{x}}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}) - Q(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t) + D_{KL}(p \| q)$$

for certain pdfs/pmfs p and q, and applying the previous problem.