Lecture 33: Simplified EM Update Derivations

Charles A. Bouman

Purdue University

Wednesday, November 11, 2026

Outline

1. The pattern in the clustering updates
2. Sufficient statistics: $p(y|\theta) = h(y)g(T(y), \theta)$
3. Example 12.1: i.i.d. Gaussian, b and S
4. Exponential families: $\langle \eta(\theta), T(y) \rangle + d(\theta) + s(y)$
5. Examples 12.2–12.4: multivariate Gaussian, discrete labels, the joint (Y, X) mixture
6. EM for exponential distributions: the Q function collapses
7. Example 12.5: EM for multivariate Gaussian mixtures

Core idea

For an exponential family, the **EM update has the same form as the ML estimate**, with the expected value of the sufficient statistic replacing the actual statistic. With this, you can write down the EM algorithm for most common distributions without computing Q.

Today's Big Ideas

- A sufficient statistic is a set of numbers computed from the data that carries all the information needed to estimate the parameter, and the ML estimate depends on the data only through it.
- An exponential family is a family of densities whose logarithm is an inner product between a function of the parameter and a statistic of the data, plus one term that depends on each alone.
- Most of the distributions we use, including the multivariate Gaussian and the distribution of discrete labels, are exponential families whose natural statistics are simple sums over the data.
- When the complete data comes from an exponential family, the Q function keeps the same shape, with the sufficient statistic replaced by its conditional expectation.
- The EM update is then the ML estimate computed from the expected statistic, so the updates for a Gaussian mixture can be written down without ever calculating Q.

There is a pattern in the EM updates

While directly calculating the Q function can be complicated, it is clear that there is a general pattern to the result.

- In the clustering example of Sections 12.2 and 12.5, the final EM update equations appear much like the conventional ML estimates, except that the means are weighted by the probability that a sample is from the particular class.
- This pattern turns out to have an underlying explanation which can be used to make the derivation of the EM updates much simpler.
- In fact, it turns out that for all exponential distributions, the form of the EM update is quite simple.

Plan for today

Derive an easy method for determining the EM update equations for any exponential distribution. Armed with this technique, you will be able to write down the EM algorithm for most common distributions, without any need to calculate the dreaded Q function.

Sufficient statistics

Two fundamental tools of statistical estimation make the EM derivations simpler: **sufficient statistics** and **exponential families**.

Sufficient statistic

We say that $T(Y) \in \mathbb{R}^k$ is a **sufficient statistic** for the **family of distributions** $p(y|\theta)$ for $\theta \in \Omega$, if the density functions can be written in the form

$$p(y|\theta) = h(y)\, g(T(y), \theta) \ ,$$

where $g(\cdot, \cdot)$ and $h(\cdot)$ are any two functions.

Intuitively, $T(Y)$ distills into a k-dimensional vector all the information required to estimate the parameter θ from the data Y. In particular, the ML estimator of θ must be a function of $T(Y)$. To see this, notice that

$$\begin{aligned} \hat{\theta}_{ML} &= \arg\max_{\theta \in \Omega} \log p(y|\theta) \\ &= \arg\max_{\theta \in \Omega} \{\log h(y) + \log g(T(y), \theta)\} \\ &= \arg\max_{\theta \in \Omega} \log g(T(y), \theta) \ = \ f(T(y)) \end{aligned}$$

for some function $f(\cdot)$.

Example 12.1: i.i.d. Gaussian random variables

Consider a set of i.i.d. Gaussian random variables $\{Y_i\}_{i=1}^N$ with distribution $N(\mu, \sigma^2)$. For this family parameterized by $\theta = \left[\mu, \sigma^2\right]^t$, the joint density can be written as

$$\begin{aligned} p(y|\theta) &= p(y_1|\theta)p(y_2|\theta)\cdots p(y_N|\theta) \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-S\frac{1}{2\sigma^2} + b\frac{\mu}{\sigma^2} - \frac{N\mu^2}{2\sigma^2}\right\} \ , \end{aligned}$$

where b and S are defined by

$$b = \sum_{i=1}^N y_i \ , \qquad S = \sum_{i=1}^N y_i^2 \ .$$

Example 12.1: i.i.d. Gaussian random variables

To see that $T(Y) = [b, S]^t$ is a sufficient statistic, we select the two functions $g(\cdot, \cdot)$ and $h(\cdot)$ to be

$$\begin{aligned} h(y) &= 1 \,, \\ g\left(\begin{bmatrix} b \\ S \end{bmatrix}, \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}\right) &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-S\frac{1}{2\sigma^2} + b\frac{\mu}{\sigma^2} - \frac{N\mu^2}{2\sigma^2}\right\} \,. \end{aligned}$$

In practice, we know that we do not need all the data in Y to estimate μ and σ. For example, we can compute the ML estimator of these parameters from the sufficient statistics b and S as

$$\begin{aligned} \hat{\mu} &= \frac{b}{N} \,, \\ \hat{\sigma}^2 &= \frac{S}{N} - \left(\frac{b}{N}\right)^2 \,. \end{aligned}$$

Exponential families

Many commonly used distributions, such as Gaussian, exponential, Poisson, Bernoulli, and binomial, have a structure which makes them particularly useful. These distributions are known as exponential families because their density can be written in a particularly useful exponential form.

Definition: Exponential Family

A family of density functions $p(y|\theta)$ for $y \in \mathbb{R}^N$ and $\theta \in \Omega$ is said to be a k-**parameter exponential family** if there exist functions $\eta(\theta) \in \mathbb{R}^k$, $s(y)$, and $d(\theta)$ and statistic $T(y) \in \mathbb{R}^k$ such that

$$p(y|\theta) = \exp\{\langle \eta(\theta), T(y) \rangle + d(\theta) + s(y)\}$$

for all $y \in \mathbb{R}^N$ and $\theta \in \Omega$, where $\langle \cdot, \cdot \rangle$ denotes an inner product. We refer to $T(y)$ as the **natural sufficient statistic** or **natural statistic** for the exponential distribution.

ML estimates for exponential families

Exponential families are extremely valuable because the log of their associated density functions forms an inner product that is easily manipulated when computing ML parameter estimates.

In general, ML estimates for exponential families have the form

$$\begin{aligned}\hat{\theta}_{ML} &= \arg\max_{\theta\in\Omega} \log p(y|\theta) \\ &= \arg\max_{\theta\in\Omega} \{\langle \eta(\theta), T(y)\rangle + d(\theta)\} \\ &= f(T(y))\end{aligned}$$

for some function $f(\cdot)$.

- The term $s(y)$ does not depend on θ and drops out.
- The remaining function of θ is an inner product plus $d(\theta)$.

The i.i.d. Gaussian is an exponential family

Referring back to Example 12.1, the i.i.d. Gaussian distribution is an example of an exponential family with natural sufficient statistic $T(y) = [b, S]^t$. More specifically, if we define

$$\begin{aligned} \eta(\theta) &= \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}, \\ d(\theta) &= -\frac{N}{2}\left\{\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2)\right\}, \end{aligned}$$

then the probability density has the form

$$p(y|\theta) = \exp\left\{\eta^t(\theta)\, T(Y) + d(\theta)\right\},$$

which means that it is an exponential family of densities parameterized by $\theta = \left[\mu, \sigma^2\right]^t$ with natural sufficient statistic $T(Y) = [b, S]^t$.

Next: three more exponential families, and for each one the natural sufficient statistic $T(Y)$.

Example 12.2: multivariate Gaussian

Let $\{Y_n\}_{n=1}^N$ be i.i.d. random vectors of dimension p with multivariate Gaussian distribution $N(\mu, R)$ and parameter $\theta = [\mu, R]$. Then $p(y|\theta)$ is an exponential distribution with natural sufficient statistics

$$b = \sum_{n=1}^{N} y_n \,, \qquad S = \sum_{n=1}^{N} y_n y_n^t \,.$$

To see why this is true, we use the trace form of the multivariate Gaussian density to write the density function for Y as

$$p(y|\theta) = \frac{1}{(2\pi)^{Np/2}} |R|^{-N/2} \exp\left\{ -\frac{1}{2} \left(\mathrm{tr}\left\{ SR^{-1} \right\} - 2b^t R^{-1}\mu + N\mu^t R^{-1}\mu \right) \right\},$$

which has the form required by a natural sufficient statistic of the distribution. From these sufficient statistics, we can also compute the ML estimate $\hat{\theta} = [\hat{\mu}, \hat{R}]$ as

$$\hat{\mu} = b/N \,, \qquad \hat{R} = (S/N) - \hat{\mu}\hat{\mu}^t \,.$$

I.i.d. Gaussian random variables are just a special case of this example when the vector dimension is $p = 1$.

Example 12.3: discrete labels

Let $\{X_n\}_{n=0}^{N-1}$ be i.i.d. random variables which take on the discrete values in the set $\{0, \cdots, M-1\}$. The distribution of X_n is parameterized by $\theta = [\pi_0, \cdots, \pi_{M-1}]$, where $P\{X_n = i\} = \pi_i$. Then $p(x|\theta)$ is an exponential distribution with natural sufficient statistics

$$N_m = \sum_{n=1}^{N} \delta(X_n - m) \ .$$

To see why this is true, we can write the density function as

$$p(x|\theta) = \prod_{m=0}^{M-1} \pi_m^{N_m}$$

because N_m counts the number of times X_n takes on the value m. This can then be rewritten as

$$p(x|\theta) = \exp\left\{ \sum_{m=0}^{M-1} N_m \log \pi_m \right\} \ ,$$

which has the required form. From these sufficient statistics we can compute the ML estimate $\hat{\theta} = [\hat{\pi}_0, \cdots, \hat{\pi}_{M-1}]$ as $\hat{\pi}_m = N_m/N$.

Example 12.4: the joint distribution of (Y, X)

Consider two random processes Y_n and X_n with a structure similar to the clustering problem of Sections 12.2 and 12.5, except that each Y_n is a p-dimensional multivariate Gaussian random vector when conditioned on each X_n. Assume the conditional distribution of each Y_n given each X_n is Gaussian with conditional mean μ_m and conditional covariance R_m, where $m = X_n$.

More specifically, we have that

$$p(y|x) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{p/2}} |R_{x_n}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(y_n - \mu_{x_n})^t R_{x_n}^{-1}(y_n - \mu_{x_n}) \right\} ,$$

where y and x represent the entire sequence of N random vectors and N random variables.

Example 12.4: the joint distribution of (Y, X)

Then in this case, $p(y, x|\theta)$ is an exponential distribution with natural sufficient statistics given by

$$N_m = \sum_{n=1}^{N} \delta(x_n - m) , \qquad b_m = \sum_{n=1}^{N} y_n \delta(x_n - m) ,$$

$$S_m = \sum_{n=1}^{N} y_n y_n^t \delta(x_n - m) .$$

The ML parameter estimate $\hat{\theta}$ can then be computed from these sufficient statistics as

$$\hat{\pi}_m = \frac{N_m}{N} , \qquad \hat{\mu}_m = \frac{b_m}{N_m} , \qquad \hat{R}_m = \frac{S_m}{N_m} - \frac{b_m b_m^t}{N_m^2} .$$

The proof of these facts is left as an exercise (Problem 4 at the end of the chapter).

Back to the incomplete data problem

Of course the problem with these ML estimates is that we may not know the labels X_n. This is the incomplete data problem that the EM algorithm addresses. In the next section, we will see how the EM algorithm can be easily applied for any such exponential distribution.

EM for exponential distributions: the setup

Assume that the joint distribution of (X, Y) comes from an exponential family, where Y is the observed or incomplete data, X is the unobserved data, and θ is the parameter to be estimated by the EM algorithm. Then we know that

$$p(y, x|\theta) = \exp\{\langle g(\theta), T(y, x)\rangle + d(\theta) + s(y, x)\}$$

for some natural sufficient statistic $T(y, x)$.

Assuming that the ML estimate of θ exists, it is given by

$$\begin{aligned} \theta_{ML} &= \arg\max_{\theta\in\Omega} \{\langle g(\theta), T(y, x)\rangle + d(\theta)\} \\ &= f(T(y, x)) \ , \end{aligned}$$

where $f(\cdot)$ is some function of $T(y, x)$.

The Q function collapses

Recalling the form of the Q function, we have

$$Q(\theta;\theta') = \mathbb{E}[\log p(y, X|\theta)|Y = y, \theta'] \ ,$$

where Y is the observed data and X is the unknown data. Using the assumed structure of the exponential distribution, we have that

$$\begin{aligned} Q(\theta;\theta') &= \mathbb{E}[\log p(y, X|\theta)|Y = y, \theta'] \\ &= \mathbb{E}[\langle g(\theta), T(y, X)\rangle + d(\theta) + s(y, X)|Y = y, \theta'] \\ &= \left\langle g(\theta), \bar{T}(y)\right\rangle + d(\theta) + \text{const.} \ , \end{aligned}$$

where

$$\bar{T}(y) = \mathbb{E}[T(y, X)|Y = y, \theta']$$

is the conditional expectation of the sufficient statistic $T(y, X)$, and const. is a constant which does not depend on θ. Since our objective is to maximize Q with respect to θ, this constant can be dropped.

The EM update, and the punchline

A single update of the EM algorithm is then given by the recursion

$$\begin{aligned} \theta'' &= \arg\max_{\theta \in \Omega} Q(\theta; \theta') \\ &= \arg\max_{\theta \in \Omega} \left\{ \left\langle g(\theta), \bar{T}(y) \right\rangle + d(\theta) \right\} \\ &= f(\bar{T}(y)) \ . \end{aligned}$$

Intuitively, we see that the EM update $\theta'' = f(\bar{T}(y))$ has the same form as the computation of the ML estimate $\theta_{ML} = f(T(y, x))$, but with the expected value of the statistic, $\bar{T}$, replacing the actual statistic, T.

Example 12.5: EM for multivariate Gaussian mixtures

Let Y_n and X_n be as in Example 12.4. So $\{Y_n\}_{n=1}^N$ are p-dimensional conditionally independent Gaussian random vectors given the class labels $\{X_n\}_{n=1}^N$, with conditional mean and covariance given by μ_m and R_m when $X_n = m$, and X_n are assumed i.i.d. with $P\{X_n = m\} = \pi_m$ for $0 \leq m < M$.

Then we know from Example 12.4 that $p(y, x|\theta)$ is an exponential distribution with parameter

$$\theta = (\pi_0, \mu_0, R_0, \cdots, \pi_{M-1}, \mu_{M-1}, R_{M-1})$$

and natural sufficient statistics N_m, b_m, and S_m.

In order to derive the EM update, we only need to replace the sufficient statistics in the ML estimate by their expected values.

Example 12.5: EM for multivariate Gaussian mixtures

So to compute the EM update of the parameter θ, we first must compute the conditional expectation of the sufficient statistics. We can do this for the statistic N_m as follows:

$$\begin{aligned}
\bar{N}_m &= \mathbb{E}\left[N_m | Y = y, \theta^{(k)}\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{N} \delta(x_n - m) \middle| Y = y, \theta^{(k)}\right] \\
&= \sum_{n=1}^{N} \mathbb{E}\left[\delta(x_n - m) | \, Y = y, \theta^{(k)}\right] \\
&= \sum_{n=1}^{N} P\{X_n = m | Y = y, \theta^{(k)}\} \, .
\end{aligned}$$

Using a similar approach for all three sufficient statistics yields the E-step, on the next slide.

Example 12.5: EM for multivariate Gaussian mixtures

Using a similar approach for all three sufficient statistics yields the E-step

$$\begin{aligned}
\bar{N}_m &= \sum_{n=1}^{N} P\{X_n = m | Y = y, \theta^{(k)}\} , \\
\bar{b}_m &= \sum_{n=1}^{N} y_n P\{X_n = m | Y = y, \theta^{(k)}\} , \\
\bar{S}_m &= \sum_{n=1}^{N} y_n y_n^t P\{X_n = m | Y = y, \theta^{(k)}\} .
\end{aligned}$$

Then in order to calculate the M-step, we simply use these expected statistics in place of the conventional statistics in the equations for the ML estimator:

$$\pi_m^{(k+1)} = \frac{\bar{N}_m}{N} , \qquad \mu_m^{(k+1)} = \frac{\bar{b}_m}{\bar{N}_m} , \qquad R_m^{(k+1)} = \frac{\bar{S}_m}{\bar{N}_m} - \frac{\bar{b}_m \bar{b}_m^t}{\bar{N}_m^2} .$$

With each repetition of this process, we increase the likelihood of the observations. Assuming that the likelihood has a maximum[1] and that one is not trapped in a local maximum, repeatedly applying the EM iterations will converge to the ML estimate of θ.

[1] For the case of a Gaussian mixture with no lower bound on the eigenvalues of $R_m^{(k)}$, the likelihood is not bounded. However, in practice a local maximum of the likelihood usually provides a good estimate of the parameters.

Summary

- $T(Y)$ is a sufficient statistic when $p(y|\theta) = h(y)g(T(y), \theta)$; then $\hat{\theta}_{ML} = f(T(y))$.
- A k-parameter exponential family has $p(y|\theta) = \exp\{\langle \eta(\theta), T(y) \rangle + d(\theta) + s(y)\}$, with $T(y)$ the natural sufficient statistic.
- Natural statistics: i.i.d. and multivariate Gaussian (b, S); discrete labels (N_m); the joint (Y, X) mixture (N_m, b_m, S_m).
- If $p(y, x|\theta)$ is an exponential family, then $Q(\theta; \theta') = \langle g(\theta), \bar{T}(y) \rangle + d(\theta) + \text{const.}$, with $\bar{T}(y) = \mathbb{E}[T(y, X)|Y = y, \theta']$.
- EM update: $\theta'' = f(\bar{T}(y))$ — the ML estimate with $\bar{T}$ in place of T.
- Example 12.5: the multivariate Gaussian mixture E-step ($\bar{N}_m$, $\bar{b}_m$, $\bar{S}_m$) and M-step ($\pi_m^{(k+1)}$, $\mu_m^{(k+1)}$, $R_m^{(k+1)}$).

Reading: Section 12.7. Section 12.8 (order estimation for Gaussian mixtures, the MDL criterion) is not covered in lecture; you will work with it directly in Lab 7. Next lecture: Chapter 13, Markov chains.