2006

TR-2006001: The EM Algorithm As a Lower Bound Optimization Technique

Rave Harpaz

Robert Haralick

The EM Algorithm as a Lower Bound Optimization Technique

Rave Harpaz, Robert Haralick
Pattern Recognition Laboratory
The Graduate Center, City University of New York,
365 Fifth Avenue New York, NY 10016, USA

January 18, 2006

Abstract

The Expectation-Maximization (EM) algorithm is an iterative optimization technique that seeks to find the maximum likelihood parameter estimates in problems where some of the data is missing or hidden, or in problems that can be posed in a similar form, such as mixture model parameter estimation. The EM algorithm can be viewed in many different ways, one of the most insightful being in terms of lower bound maximization which better illustrates its underlying principles. There are several very good references discussing the EM algorithm in greater generality. The purpose of report is to present the EM algorithm in a more self-contained way from a lower bounding viewpoint, and show how it can be used used to find the parameters of a mixture of densities.

1 Maximum Likelihood Estimation

Suppose we have a known parametric density function $p(\mathbf{x}|\theta)$ governed by a set of parameters θ which are unknown but which need to be estimated. For example for a multivariate Gaussian we would like to estimate $\theta = (\mu, \Sigma)$. Assuming we have a data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of d-dimensional random vectors/samples independently drawn according to $p(\mathbf{x}|\theta)$, then the joint pdf $p(X|\theta)$ is given by

$$p(X|\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i|\theta) = L(\theta|X) \tag{1}$$

where $L(\theta|X)$ is known as the *likelihood* function of θ with respect to, or given X. The maximum likelihood estimate of θ is by definition the value $\hat{\theta}$ that maximizes $L(\theta|X)$ i.e.,

$$\hat{\theta} = \arg\max_{\theta} L(\theta|X) \tag{2}$$

Intuitively this estimate corresponds the set of parameters θ that best agrees or supports the observed data. For analytical purposes, it is usually easier to work with the logarithm of the likelihood. Because the logarithm is a monotonically increasing, the value $\hat{\theta}$ that maximizes the *log-likelihood* also maximizes the likelihood, in which case we define

$$L(\theta|X) = \log \prod_{i=1}^{n} p(\mathbf{x}_i|\theta) = \sum_{i=1}^{n} \log p(\mathbf{x}_i|\theta) \tag{3}$$

If $L(\theta|X)$ is a well-behaved, differentiable function of θ then $\hat{\theta}$ can be found analytically by standard methods of differential calculus. Specifically, the set of necessary conditions for the maximum likelihood estimate are given by

$$\nabla_\theta = \frac{\partial L(\theta|X)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log p(\mathbf{x}_i|\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{p(\mathbf{x}_i|\theta)} \frac{\partial p(\mathbf{x}_i|\theta)}{\partial \theta} = \mathbf{0} \tag{4}$$

That is, the maximum likelihood estimate must satisfy the condition that the gradient of the log-likelihood with respect to θ must equal zero, and if θ is a k-component vector then we will need to solve a system of k equations. In cases where a solution to this set of equations cannot be obtained in closed form we must resort to more elaborate techniques such as iterative optimization methods.

A simple case where a closed form solution does exist is when $\mathbf{x}$ follows a multivariate Gaussian distribution i.e.,

$$p(\mathbf{x}|\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

where neither μ nor Σ are known, but need to be estimated given the data using the principle of maximum likelihood. The log-likelihood in this case is given by

$$L(\theta|X) = \sum_{i=1}^{n} -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{x}_i-\mu)'\Sigma^{-1}(\mathbf{x}_i-\mu)$$

Taking partial derivatives of the log-likelihood with respect to μ and Σ and equating them to zero, gives the following two equations

$$\frac{\partial L(\theta|X)}{\partial \mu} = \sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x}_i-\mu) = \mathbf{0}$$

$$\frac{\partial L(\theta|X)}{\partial \Sigma} = \sum_{i=1}^{n} -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-2}(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)' = \mathbf{0}$$

Rearranging each of the two equations gives us the following maximum likelihood estimates of μ and Σ

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i-\hat{\mu})(\mathbf{x}_i-\hat{\mu})'$$

Note however that the maximum likelihood estimate of the covariance matrix is biased.

2 Mixtures of Parametric Models

A finite mixture model is a model for $p(\mathbf{x}|\theta)$ which has the form of a weighted sum of component densities as follows

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} p(\mathbf{x}|k,\theta)P(k|\theta) \tag{5}$$

where

$$\sum_{k=1}^{K} P(k|\theta) = 1 \qquad \int_{\mathbf{x}} p(\mathbf{x}|k,\theta)d\mathbf{x} = 1$$

In other words, it is assumed that there are K densities contributing to the formation of the overall density $p(\mathbf{x}|\theta)$, and the model decomposes the overall density into a weighted linear combination of K *component* or *class* densities $p(\mathbf{x}|k,\theta)$ where $P(k|\theta)$ represents the probability of the the k-th class, or the probability that a randomly chosen data point was generated by the k-th component.

In many applications of pattern recognition where K classes or categories are involved, we are usually interested in determining the membership of a data point in a given class. Using this model, and assuming we know θ and $P(k|\theta)$ for each k we can compute the membership of a data point in class k given θ by a direct application of the Bayes rule,

$$p(k|\mathbf{x},\theta) = \frac{p(\mathbf{x}|k,\theta)P(k|\theta)}{p(\mathbf{x}|\theta)} = \frac{p(\mathbf{x}|k,\theta)P(k|\theta)}{\sum_{k=1}^{K} p(\mathbf{x}|k,\theta)P(k|\theta)}$$

However, in most applications neither θ nor $P(k|\theta)$ are known in advance and must be estimated. An immediate thought would be to use maximum likelihood, where the goal would be to maximize

$$\sum_{i=1}^{n} \log p(\mathbf{x}_i|\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} p(\mathbf{x}_i|k,\theta)P(k|\theta)$$

with respect to θ and each $P(k|\theta)$. In contrast to regular maximum likelihood estimation, the difficulty here is two-fold. First, the unknown parameters enter the maximization task in a nonlinear fashion which calls for nonlinear optimization techniques to be employed. But more so is the fact that the labels of the data points are unknown, that is, the specific class or component from which each data point arises is unknown. This makes the problem analytically intractable. If the class labels were known then we could collect the data from each class and then carry out K separate maximum likelihood estimation tasks. This missing or hidden label information makes the current problem a typical problem with incomplete data for which the EM algorithm was designed for.

3 The EM Algorithm

The Expectation-Maximization (EM) algorithm is a general method of finding the maximum likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values. Its wide spread applicability was first discussed by Dempster et al. [DLR77]. There are two main applications of the EM Algorithm. The first occurs when the data has missing values, due to problems or limitations with the observation process. The other, which applies to the case of mixture models and which is more common in pattern recognition problems, is when the optimization of the likelihood function is analytically intractable but can be simplified by assuming the existence of additional but *missing* or *hidden* variables, such as class labels.

The EM algorithm, similar to other optimization schemes, is an iterative optimization technique which gradually improves the parameter estimates. However, unlike gradient ascent which makes a local linear approximation to the objective function or Newton methods which make quadratic approximations at each iteration and then take some *uphill* step, EM makes a local approximation which is a lower bound approximation of the objective function. Choosing a new guess to maximize the lower bound will always be an improvement over the previous guess, unless the gradient there was zero. An illustration of this concept is presented in Fig. 1. The main difficulty with gradient ascent or Newton methods which is not present in the EM scheme is that we do not know in advance how good the linear or quadratic approximations are, neither do we know in advance how big of an uphill step must be taken.

The underlying idea of the EM algorithm is as follows. Starting with an initial guess of the parameters that need to be estimated. Each iteration consists of two steps. The first an Expectation (E) step for

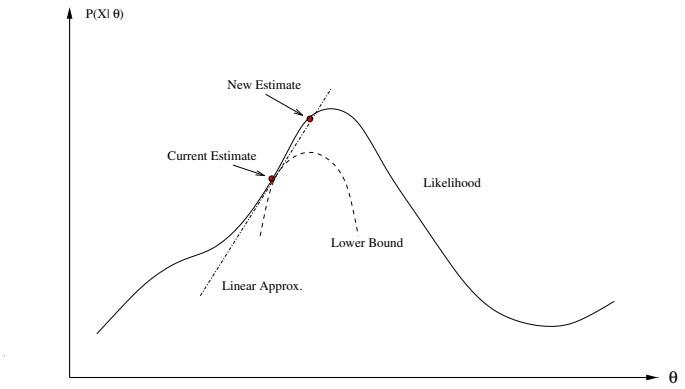

Figure 1: linear vs. lower bound approximation for maximum likelihood

computing a local lower bound approximation to the objective function (log-likelihood), and maximizing it with respect to the distribution of the unobserved data. This step will be latter shown to be equivalent to finding the distribution of the unobserved or missing variables given the observed data and the current parameter estimates. The second step is a Maximization (M) step which maximizes the lower bound with respect to the parameters of the underlying distribution assuming that the distribution of the missing data found in the E-step is correct. These two steps are repeated until convergence of the parameter estimates is reached, or until a local maximum is found.

As before let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be our set of d-dimensional observations. We refer to X as our *incomplete* data. We will assume that the complete data is $Z = (X, Y)$ where $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ represents the n unobserved or hidden data vectors which are in one-to-one correspondence with X, i.e., $\mathbf{y}_i$ is associated with $\mathbf{x}_i$. Although not necessary we will assume Y to be discrete and in the case of mixture models to represent the class labels, i.e., $\mathbf{y}_i = k$ where $k \in \{1, 2, \ldots, K\}$. Furthermore we assume a joint density function $p(\mathbf{x}, \mathbf{y}|\theta)$ between the observed and missing values. With this density function we can define the complete-data likelihood function $L(\theta|Z) = L(\theta|X, Y) = p(X, Y|\theta)$. Note that this function is a random variable since the missing information Y is unknown, assumed to be random, and presumably governed by some underlying distribution $q(Y)$. Given this joint density function we can also define the incomplete-data log-likelihood as

$$L(\theta|X) = \log p(X|\theta) = \log \sum_Y p(X, Y|\theta) \tag{6}$$

That is, the incomplete-data log-likelihood can be expressed as the complete-data log-likelihood summed over or marginalized over the unobserved data values. As mentioned earlier the problem with maximizing (6) is that it involves the log of a sum, and that both θ and the hidden data Y are unknown.

The main idea of EM's optimization scheme is to construct a tractable lower bound $G(\theta, q(Y))$ for $L(\theta|X)$, i.e., $L(\theta|X) \geq G(\theta, q(Y))$, which is parameterized by θ and the distribution of Y, and that instead contains a sum of logarithms. To derive this bound we can first trivially rewrite the log-likelihood as

$$L(\theta|X) = \log p(X|\theta) = \log \sum_Y p(X, Y|\theta) = \log \sum_Y q(Y) \frac{p(X, Y|\theta)}{q(Y)}$$

where $q(Y)$ is at the moment some arbitrary distribution of the hidden data Y. Because of the concavity of

the log function we may use Jensen's inequality (Appendix A) to get the bound as follows

$$\begin{aligned} L(\theta|X) &= \log \sum_Y q(Y) \frac{p(X,Y|\theta)}{q(Y)} \\ &\geq \sum_Y q(Y) \log \frac{p(X,Y|\theta)}{q(Y)} \\ &= \sum_Y q(Y) \log p(X,Y|\theta) - q(Y) \log q(Y) \\ &= G(\theta, q(Y)) \end{aligned} \tag{7}$$

Inequality (7) is true for any distribution q, however what we would like is to obtain a distribution q that will yield an optimal or tight bound, and not just any bound. In other words we require the bound to touch the objective function $L(\theta|X)$ at our current estimate of θ as depicted in Fig. 1. This will also guarantee that we obtain an improved estimate of θ when we locally maximize the bound with respect to θ in the M-step step. Finding the optimal bound is done by maximizing $G(\theta, q(Y))$ with respect to the distribution $q(Y)$. This maximization can be done by introducing a Lagrange multiplier to enforce the the constraint $\sum_Y q(Y) = 1$, and then using differential calculus to find the maxima of a Lagrangian function(Appendix C). However a simpler and more insightful way is to rewrite $G(\theta, q(Y))$ as follows [Min98]:

$$\begin{aligned} G(\theta, q(Y)) &= \sum_Y q(Y) \log \frac{p(X,Y|\theta)}{q(Y)} \\ &= E_Y \left[\log \frac{p(X,Y|\theta)}{q(Y)} \right] \\ &= E_Y \left[\log \frac{p(Y|X,\theta)p(X|\theta)}{q(Y)} \right] \\ &= E_Y \left[\log \frac{p(Y|X,\theta)}{q(Y)} + \log p(X|\theta) \right] \\ &= E_Y \left[\log \frac{p(Y|X,\theta)}{q(Y)} \right] + E_Y \left[\log p(X|\theta) \right] \\ &= -E_Y \left[\log \frac{q(Y)}{p(Y|X,\theta)} \right] + \log p(X|\theta) \\ &= -D(q(Y) \| p(Y|X,\theta) + L(\theta|X) \end{aligned} \tag{8}$$

Assuming $p(Y)$ is a valid probability distribution, the Kullback-Leibler Distance (Relative Entropy) $D(q(Y)\|p(Y|X,\theta)$ is a measure of the distance between the distributions $q(Y)$ and $p(Y|X,\theta)$. Since this measure is always nonnegative and equal to zero when the two distributions are the same (Appendix B), $G(\theta, q(Y))$ is maximized with respect to $q(Y)$ when $D(q(Y)\|p(Y|X,\theta) = 0$, i.e., when $q(Y) = p(Y|X,\theta)$. From (8) it is easy to see that indeed when this is the case, the lower bound $G(\theta, q(Y))$ is tight and equals to the log-likelihood $L(\theta|X)$.

Finding the distribution q that yields the optimal bound given the current estimate of θ is the **E-step**. To get the next estimate of θ which is the **M-step**, we maximize the bound with respect to θ given q that was found in the E-step. This step is problem dependent and in many cases has a closed form solution. From (7) we can see that the relevant term to maximize the bound $G(\theta, q(Y))$ with respect to θ is

$$\sum_Y q(Y) \log p(X,Y|\theta) = \sum_Y p(Y|X,\theta) \log p(X,Y|\theta) = E_{Y|X,\theta} \left[\log p(X,Y|\theta) \right] \tag{9}$$

Letting θ_i, θ_{i+1}, q_i and q_{i+1} denote our current and next best estimates of θ and q respectively, we can now state the EM algorithm as - "repeat until convergence":

$$\begin{aligned} \textbf{E-step} &: q_{i+1}(Y) = \arg\max_q \; G(\theta_i, q(Y)) = p(Y|X,\theta_i) \\ \textbf{M-step} &: \theta_{i+1} = \arg\max_\theta \; G(\theta, q_{i+1}(Y)) = \arg\max_\theta \; E_{Y|X,\theta_i} \left[\log p(X,Y|\theta) \right] \end{aligned} \tag{10}$$

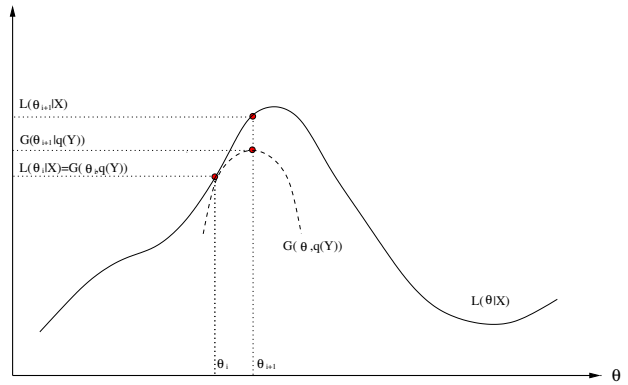

Figure 2: Graphical depiction of a single EM iteration. The function $G(\theta, q(Y))$ is a lower-bound to the likelihood $L(\theta|X)$. The functions are equal at θ_i. In the M-step θ_{i+1} is chosen as the value of θ that maximizes G. Since $L(\theta_{i+1}|X) \geq G(\theta_{i+1}, q(Y))$, increasing G insures that the likelihood is also increased at each iteration.

A graphical depiction of a single EM iteration is illustrated in Fig. 2. Note that because the bound can be expressed as an expectation, the first step is called the "expectation-step" or the E-step.

4 Convergence of the EM Algorithm

The convergence properties of the EM algorithm are discussed in detail in [MK97]. In this section we discuss the general convergence of the algorithm. Suppose that θ_{i+1} and θ_i are the parameter estimates from two successive iterations. Since θ_{i+1} is chosen to maximize G we have $G(\theta_{i+1}, q(Y)) \geq G(\theta_i, q(Y))$, because G is a tight lower bound for L we have $L(\theta_{i+1}|X) \geq G(\theta_{i+1}, q(Y))$ and $G(\theta_i, q(Y)) = L(\theta_i|X)$. Thus in summary we have

$$L(\theta_{i+1}|X) \geq G(\theta_{i+1}, q(Y)) \geq G(\theta_i, q(Y)) = L(\theta_i|X)$$

so that $L(\theta_{i+1}|X) \geq L(\theta_i|X)$, which shows that the likelihood is monotonically increasing. If in every iteration L is increased or at least does not decrease, and L has a local maximum, then at some point we are bound to reach that maximum.

5 Estimating the Parameters of a Mixture of Gaussians

Up until now the EM algorithm was presented in its most general form, obscuring implementation details which are very application dependent. In this section we discuss the mixture model parameter estimation problem which is the most widely studied application of the EM algorithm.

Recall that in the mixture model parameter estimation problem we assume the following probabilistic model:

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} p(\mathbf{x}|k, \theta) P(k|\theta)$$

where for a mixture of Gaussians $p(\mathbf{x}|k, \theta)$ represents the pdf of the k-th Gaussian, $P(k|\theta)$ its prior, and where both the parameters and priors of each Gaussian must be estimated.

The E-step

According to (10) in the E-step we need to compute $p(Y|X,\theta)$, i.e. compute the distribution of the unobserved data given the observed data and the current parameter estimates. This is equivalent to computing probability of each data point arising from each component and then each possible joint, i.e., computing for each i and k, where $i = 1,\ldots,n$ and $k = 1,\ldots,K$ the following probability

$$p(\mathbf{y}_i = k|\mathbf{x}_i,\theta) = \frac{p(\mathbf{x}_i|\mathbf{y}_i = k,\theta)p(\mathbf{y}_i = k|\theta)}{p(\mathbf{x}_i|\theta)} = \frac{p(\mathbf{x}_i|\mathbf{y}_i = k,\theta)p(\mathbf{y}_i = k|\theta)}{\sum_{j=1}^{K} p(\mathbf{x}_i|j,\theta)P(j|\theta)} \tag{11}$$

and then computing all K^n joints, each of the form $\prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{x}_i,\theta)$. Nonetheless, because the joint will not be used directly in the M-step, the E-step reduces to computing for each i and k $p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)$ as given in (11). To avoid confusion note that $p(\mathbf{y}_i = k|\theta)$ and $p(\mathbf{x}_i|\mathbf{y}_i = k,\theta)$ which are used in (11) have the same meaning as $P(k|\theta)$ and $p(\mathbf{x}_i|k,\theta)$ respectively, which were used to define a mixture model. These two forms can be used interchangeably, however to stress the idea of hidden variables, we will use the first form, in which case the denominator in (11) can be rewritten as $\sum_{j=1}^{K} p(\mathbf{x}_i|\mathbf{y}_i = j,\theta)p(\mathbf{y}_i = j|\theta)$.

The M-step

According to (10) the M-step requires us to maximize $E_{Y|X,\theta}[\log p(X,Y|\theta)]$ with respect to θ. For the case of a mixture of Gaussians this can be done in closed form using standard differential calculus, substituting the pdf of a Gaussian wherever necessary. Furthermore the objective function for the maximization can be rewritten as (proof in Appendix C)

$$\begin{aligned} E_{p(Y|X,\theta)}[\log p(X,Y|\theta)] &= \sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)\log\left[p(\mathbf{x}_i|\mathbf{y}_i = k,\theta)p(\mathbf{y}_i = k|\theta)\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)\log p(\mathbf{x}_i|\mathbf{y}_i = k,\theta) \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)\log p(\mathbf{y}_i = k|\theta) \end{aligned} \tag{12}$$

Since $p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)$ was already computed in the E-step, the objective function is now expressed as the addition of two unrelated terms, the first involving the pdf of the k-th component and the second its prior. Therefore to compute an update for the parameters of the k-th pdf we only need to maximize the first term, likewise to compute an update for its prior we only need to maximize the second term. Upon preforming this step the updated estimates in terms of the old ones for a mixture of Gaussians are (derivation in Appendix C):

$$p(\mathbf{y}_i = k|\theta)^{new} = \frac{1}{n}\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i,\theta) \tag{13}$$

$$\begin{aligned} \mu_k^{new} &= \frac{\sum_{i=1}^{n}\mathbf{x}_i p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)}{\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)} \\ &= \frac{1}{n\; p(\mathbf{y}_i = k|\theta)^{new}}\sum_{i=1}^{n}\mathbf{x}_i p(\mathbf{y}_i = k|\mathbf{x}_i,\theta) \end{aligned} \tag{14}$$

$$\begin{aligned} \Sigma_k &= \frac{\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)(\mathbf{x}_i - \mu_k^{new})(\mathbf{x}_i - \mu_k^{new})'}{\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)} \\ &= \frac{1}{n\; p(\mathbf{y}_i = k|\theta)^{new}}\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)(\mathbf{x}_i - \mu_k^{new})(\mathbf{x}_i - \mu_k^{new})' \end{aligned} \tag{15}$$

Note that each of the new estimates in (14) and (15) is similar to those involved when estimating the parameters of a single Gaussian, except that each point is now weighted by the probability that it belongs to the k-th component.

6 Limitations of the EM Algorithm

The EM algorithm like many other iterative optimization techniques is a local technique and as so is just as susceptible as other techniques to local optima. The rate of convergence is typically good during the first few iterations, but can become very slow as it approaches the local optima. Generally the EM algorithm works best when the fraction of missing information is small and the dimensionality of the data is not too large. In general, there is no consensus on whether EM performs better than other iterative optimization methods, and it is widely accepted that the performance depends on the shape of the lower bound G in the EM case, and the shape of the objective function in the case of other methods such as Newton's method.

7 Generalized EM Algorithm (GEM)

In the M-step, θ_{i+1} was chosen as the value of θ for which the lower bound $G(\theta, q_{i+1}(Y))$ was maximized. While this ensures the greatest increase in G and subsequently the log-likelihood $L(\theta|X)$, it is possible to relax the requirement of maximization to one of simply increasing $G(\theta, q_{i+1}(Y))$. This approach to simply increase and not necessarily maximize is known as the Generalized Expectation Maximization (GEM) algorithm. GEM is useful when the maximization is difficult or does not have a closed form solution, in which case gradient ascent methods can be used to obtain an increase in G at each iteration. In the same spirit we can also perform partial E-steps, i.e., we do not need to maximize G over q during the E-step. Any local maximum of G in both q and θ is also a local of $L(\theta|X)$. Thus any way of maximizing G will suffice, we can take a small step along q and a small step along θ. The convergence of the GEM algorithm can be argued along the same lines as that of the EM algorithm.

Appendix A

A.1 Convexity and Concavity

Definition 1. *A function $f(x)$ is said to be **convex** over an interval (a, b) if $\forall x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

*f is said to be **strictly convex** if equality holds only when $\lambda = 0$ or $\lambda = 1$.*

Intuitively the definition states that a function is convex (strictly) if the function is always below (strictly convex) or never above (convex) the secant (line) connecting the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$.

Definition 2. *A function f is **concave** (strictly) if $-f$ is convex (strictly).*

Theorem 1. *If $f(x)$ is twice differentiable on (a, b) and $f''(x) \geq 0$ on (a, b) then $f(x)$ is convex on (a, b).*

Proof: the second order Taylor expansion of f about the point x_0 is

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + 1/2 f''(x_0)(x - x_0)^2$$

if $f''(x) \geq 0$ then the last term is non-negative and therefore

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0)$$

Let $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and $x = x_1$, then since $(1 - \lambda)(x_1 - x_2) = x_1 - x_0$ we have

$$f(x_1) \geq f(x_0) + f'(x_0)(x_1 - x_0) = f(\lambda x_1 + (1 - \lambda)x_2) + f'(\lambda x_1 + (1 - \lambda)x_2)((1 - \lambda)(x_1 - x_2)) \quad \text{(A-1)}$$

Similarly, now let $x = x_2$, then since $\lambda(x_2 - x_1) = x_2 - x_0$ we have

$$f(x_2) \geq f(x_0) + f'(x_0)(x_2 - x_0) = f(\lambda x_1 + (1 - \lambda)x_2) + f'(\lambda x_1 + (1 - \lambda)x_2)(\lambda(x_2 - x_1)) \quad \text{(A-2)}$$

Now multiplying (A-1) by λ and (A-2) by $(1 - \lambda)$, and adding then together gives the convexity inequality.

Proposition 1. *$-\ln(x)$ is strictly convex on $(0, \infty)$.*

Proof: with $f(x) = -\ln(x)$, we have $f''(x) = 1/x^2 > 0$ for $x \in (0, \infty)$. By definition 2 $\ln(x)$ is strictly concave. Other convex functions are $x^2,\ e^x,\ |x|,\ x \log x$. Other concave functions are $\log(x),\ \sqrt{x}$.

A.2 Jensen's Inequality

The idea of convexity can be extended to more than two points. This result is known as Jensen's inequality.

Theorem 2 (Jensen's inequality)**.** *Let $f(x)$ be a convex function defined on interval (a, b). If $x_1, x_2, \ldots, x_n \in (a, b)$ and $0 \leq \lambda_1, \lambda_2, \ldots, \lambda_n \leq 1$ with $\sum_{i=1}^n \lambda_i =$ then*

$$\sum_{i=1}^{n} \lambda_i f(x_i) \geq f\left(\sum_{i=1}^{n} \lambda_i x_i\right)$$

Proof: the proof is by induction. For $n = 1, n = 2$ the inequality is trivially true ($n = 2$ corresponds to the definition of convexity). Suppose the inequality is true for some k then,

$$\begin{aligned}
\sum_{i=1}^{k+1} \lambda_i f(x_i) &= \lambda_{k+1} f(x_{k+1}) + (1 - \lambda_{k+1}) \sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} f(x_i) \\
&\geq \lambda_{k+1} f(x_{k+1}) + (1 - \lambda_{k+1}) f\left(\sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} x_i\right) \qquad \text{(inductive hypothesis)} \\
&\geq f\left(\lambda_{k+1} x_{k+1} + (1 - \lambda_{k+1}) \sum_{i=1}^{k} \frac{\lambda_i}{1 - \lambda_{k+1}} x_i\right) \qquad \text{(convexity)} \\
&= f\left(\sum_{i=1}^{k+1} \lambda_i x_i\right)
\end{aligned}$$

If we think of X as a discrete random variable and λ_i's as probabilities then Jensen's inequality can be restated as

Theorem 3 (Jensen's inequality)**.** *If $f(x)$ is a convex function and X is a random variable then*

$$E[f(X)] \geq f(E[X])$$

and if $f(x)$ is strictly convex then equality implies that $X = E[X]$.

Similarly, one can show that if a function is concave then Jensen's inequality is reversed. Jensen's inequality can be used to obtain a useful result. Since $\log(x)$ is concave we have

$$E[\log(X)] \leq \log(E[X])$$

This enables us to lower-bound the log-likelihood (logarithm of sum), a result that is used in the derivation of the EM algorithm.

Finally, Jensen's inequality can also be used to prove that the arithmetic mean is greater than or equal to the geometric mean.

Theorem 4.

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \sqrt[n]{x_1 x_2 \cdots x_n}$$

Proof: if $x_1 x_2 \cdots x_n \geq 0$ then since $\ln(x)$ is concave we have

$$\begin{aligned} \ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) &\geq \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) \\ &= \frac{1}{n}\ln(x_1 x_2 \cdots x_n) \\ &= \ln(x_1 x_2 \cdots x_n)^{1/n} \end{aligned}$$

therefore we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \sqrt[n]{x_1 x_2 \cdots x_n}$$

Appendix B

B.1 Kullback-Leibler Distance (Relative Entropy)

The Kullback-Leibler (KL) measure of two distributions with pdf's $f(\mathbf{x})$ and $g(\mathbf{x})$ denoted $D(f\|g)$ can be thought of as an information theoretic distance measure between the two distributions. For the discrete case it is defined as

$$D(f\|g) = \sum_{\mathbf{x}} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} = E_f\left[\log \frac{f(\mathbf{x})}{g(\mathbf{x})}\right]$$

Note that

$$D(f\|g) = \sum_{\mathbf{x}} f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} = -\sum_{\mathbf{x}} f(\mathbf{x}) \log \frac{g(\mathbf{x})}{f(\mathbf{x})} = -E_f\left[\log \frac{g(\mathbf{x})}{f(\mathbf{x})}\right]$$

The KL distance can be shown to be always nonnegative and equal to zero when the two distributions are the same, however is is not symmetric. The proof that KL is nonnegative is as follows

$$\begin{aligned} -D(f\|g) &= \sum_{\mathbf{x}} f(\mathbf{x}) \log \frac{g(\mathbf{x})}{f(\mathbf{x})} \\ &\leq \log\left(\sum_{\mathbf{x}} f(\mathbf{x}) \frac{g(\mathbf{x})}{f(\mathbf{x})}\right) \qquad \text{(Jensen's inequality)} \\ &= \log\left(\sum_{\mathbf{x}} g(\mathbf{x})\right) = \log 1 = 0 \end{aligned}$$

so $D(f\|g) \geq 0$. Since log is strictly concave we have equality ($D(f\|g) = 0$) if and only if $f(\mathbf{x}) = g(\mathbf{x})$.

Another somewhat simpler proof uses the inequity $\log x \leq x - 1$.

$$\begin{aligned} -D(f\|g) &= \sum_{\mathbf{x}} f(\mathbf{x}) \log \frac{g(\mathbf{x})}{f(\mathbf{x})} \\ &\leq \sum_{\mathbf{x}} f(\mathbf{x}) \left(\frac{g(\mathbf{x})}{f(\mathbf{x})} - 1\right) \qquad (\log x \leq x - 1) \\ &= \sum_{\mathbf{x}} (g(\mathbf{x}) - f(\mathbf{x})) = \sum_{\mathbf{x}} g(\mathbf{x}) - \sum_{\mathbf{x}} f(\mathbf{x}) d\mathbf{x} = 1 - 1 = 0 \end{aligned}$$

Appendix C

C.1 Deriving the Optimal Lower Bound

Deriving the optimal lower bound can be done by maximizing $G(\theta, q(Y))$ with respect to the distribution $q(Y)$. This will elevate the bound to touch the objective $L(\theta|X)$ at θ. Introducing a Lagrange multiplier to enforce the constraint that $\sum_Y q(Y) = 1$ gives us the following Lagrangian function that needs to be maximized

$$F(\theta, q(Y)) = \sum_Y q(Y) \log p(X, Y|\theta) - q(Y) \log q(Y) - \lambda \left(\sum_Y q(Y) - 1\right)$$

Taking partial derivatives with respect to $q(Y)$ and λ give the following two equations

$$\begin{aligned} \frac{\partial F}{\partial q(Y)} &= \log p(X, Y|\theta) - (1 + \log q(Y)) - \lambda = 0 \qquad \text{(C-1)} \\ \frac{\partial F}{\partial \lambda} &= \sum_Y q(Y) - 1 = 0 \end{aligned}$$

where after some algebraic manipulation and taking the natural logarithm we obtain

$$\sum_Y q(Y) = e^{-(1+\lambda)} \sum_Y p(X, Y|\theta) = e^{-(1+\lambda)} p(X|\theta) = 1 \qquad \text{(C-2)}$$

Solving (C-2) for λ and substituting back into (C-1) we get

$$\begin{aligned} \log p(X, Y|\theta) - 1 - \log q(Y) - \lambda &= \log p(X, Y|\theta) - 1 - \log q(Y) - \log p(X|\theta) + 1 \\ &= \log p(X, Y|\theta) - \log q(Y) - \log p(X|\theta) \\ &= 0 \end{aligned}$$

From which we find that $G(\theta, q(Y))$ is maximized when

$$q(Y) = \frac{p(X,Y|\theta)}{p(X|\theta)} = \frac{p(Y|X,\theta)p(X|\theta)}{p(X|\theta)} = p(Y|X,\theta) \tag{C-3}$$

It is then easy to see that when $q(Y)$ is substituted by $p(Y|X,\theta)$,

$$G(\theta, q(Y)) = \log p(X|\theta) = L(\theta|X)$$

C.2 Simplifying the Bound for the M-step

In the M-step we need to maximize the bound $G(\theta, q(Y))$ with respect to θ. Having found $p(Y|X,\theta)$ in the E-step, this is equivalent to maximizing $\sum_Y p(Y|X,\theta)\log p(X,Y|\theta)$, which includes the relevant term for the maximization. Assuming $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is an instance of the unobserved data we have

$$\log p(X,Y|\theta) = \sum_{i=1}^{n} \log p(\mathbf{x}_i, \mathbf{y}_i|\theta) \qquad \text{and} \qquad p(Y|X,\theta) = \prod_{j=1}^{n} p(\mathbf{y}_j|\mathbf{x}_j,\theta)$$

Given the above and following the lines of the simplification presented in [Bil97], the objective function for maximization can be rewritten as

$$\begin{aligned}
E_{Y|X,\theta}\left[\log p(X,Y|\theta)\right] &= \sum_{Y} p(Y|X,\theta)\log p(X,Y|\theta) \\
&= \sum_{Y}\sum_{i=1}^{n} \log[p(\mathbf{x}_i,\mathbf{y}_i|\theta)] \prod_{j=1}^{n} p(\mathbf{y}_j|\mathbf{x}_j,\theta) \\
&= \sum_{\mathbf{y}_1=1}^{K} \cdots \sum_{\mathbf{y}_n=1}^{K}\sum_{i=1}^{n} \log[p(\mathbf{x}_i,\mathbf{y}_i|\theta)] \prod_{j=1}^{n} p(\mathbf{y}_j|\mathbf{x}_j,\theta) \\
&= \sum_{\mathbf{y}_1=1}^{K} \cdots \sum_{\mathbf{y}_n=1}^{K}\sum_{i=1}^{n}\sum_{k=1}^{K} \delta_{k,\mathbf{y}_i} \log[p(\mathbf{x}_i,\mathbf{y}_i=k|\theta)] \prod_{j=1}^{n} p(\mathbf{y}_j|\mathbf{x}_j,\theta)
\end{aligned} \tag{C-4}$$

where $\delta_{k,\mathbf{y}_i}$ equals one when the value of $\mathbf{y}_i$ equals k and zero otherwise.

Since the value of $\mathbf{y}_i$ in $\log p(\mathbf{x}_i,\mathbf{y}_i=k|\theta)$ now only depends on k and i we can rewrite (C-4) as

$$E_{Y|X,\theta}\left[\log p(X,Y|\theta)\right] = \sum_{i=1}^{n}\sum_{k=1}^{K} \log[p(\mathbf{x}_i,\mathbf{y}_i=k|\theta)] \sum_{\mathbf{y}_1=1}^{K} \cdots \sum_{\mathbf{y}_n=1}^{K} \delta_{k,\mathbf{y}_i} \prod_{j=1}^{n} p(\mathbf{y}_j|\mathbf{x}_j,\theta) \tag{C-5}$$

Having i and k fixed

$$\begin{aligned}
\sum_{\mathbf{y}_1=1}^{K} \cdots \sum_{\mathbf{y}_n=1}^{K} \delta_{k,\mathbf{y}_i} \prod_{j=1}^{n} p(\mathbf{y}_i|\mathbf{x}_i,\theta) &= \sum_{\mathbf{y}_1=1}^{K} \cdots \left(\sum_{\mathbf{y}_i=1}^{K} \delta_{k,\mathbf{y}_i} p(\mathbf{y}_i|\mathbf{x}_i,\theta)\right) \cdots \sum_{\mathbf{y}_n=1}^{K} \prod_{j=1,j\neq i}^{n} p(\mathbf{y}_j|\mathbf{x}_j,\theta) \\
&= \sum_{\mathbf{y}_1=1}^{K} \cdots \sum_{\mathbf{y}_i-1=1}^{K} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta) \sum_{\mathbf{y}_i+1=1}^{K} \cdots \sum_{\mathbf{y}_n=1}^{K} \prod_{j=1,j\neq i}^{n} p(\mathbf{y}_j|\mathbf{x}_j,\theta) \\
&= p(\mathbf{y}_i=k|\mathbf{x}_i,\theta) \sum_{\mathbf{y}_1=1}^{K} \cdots \sum_{\mathbf{y}_i-1=1}^{K} \sum_{\mathbf{y}_i+1=1}^{K} \cdots \sum_{\mathbf{y}_n=1}^{K} \prod_{j=1,j\neq i}^{n} p(\mathbf{y}_j|\mathbf{x}_j,\theta) \\
&= p(\mathbf{y}_i=k|\mathbf{x}_i,\theta) \prod_{j=1,j\neq i}^{n} \sum_{\mathbf{y}_j=1}^{K} p(\mathbf{y}_j|\mathbf{x}_j,\theta)
\end{aligned} \tag{C-6}$$

Since $\sum_{\mathbf{y}_j=1}^{K} p(\mathbf{y}_j|\mathbf{x}_j,\theta) = 1$, using (C-6) we can rewrite (C-5) as

$$\begin{aligned} E_{Y|X,\theta}\left[\log p(X,Y|\theta)\right] &= \sum_{i=1}^{n}\sum_{k=1}^{K} \log[p(\mathbf{x}_i,\mathbf{y}_i=k|\theta)]p(\mathbf{y}_i=k|\mathbf{x}_i,\theta) \prod_{j=1,j\neq i}^{n} 1 \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta)\log p(\mathbf{x}_i,\mathbf{y}_i=k|\theta) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta)\log[p(\mathbf{x}_i|\mathbf{y}_i=k|\theta)p(\mathbf{y}_i=k|\theta)] \end{aligned} \tag{C-7}$$

C.3 Deriving the Updated Parameter Estimates for a Mixture of Gaussians

To derive an update for the k-th prior we need to maximize

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta)\log p(\mathbf{y}_i=k|\theta)$$

with respect to $p(\mathbf{y}_i = k|\theta)$, where the value for $p(\mathbf{y}_i = k|\mathbf{x}_i,\theta)$ was already computed in the E-step. Introducing a Lagrange multiplier to enforce the constraint that $\sum_{k=1}^{K} p(\mathbf{y}_i = k|\theta) = 1$ gives us the following Lagrangian function

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta)\log p(\mathbf{y}_i=k|\theta) - \lambda\left(\sum_{k=1}^{K} p(\mathbf{y}_i=k|\theta) - 1\right)$$

Taking partial derivatives with respect to $p(\mathbf{y}_i = k|\theta)$ and λ, and using the natural logarithm gives the following two equations

$$\frac{1}{p(\mathbf{y}_i=k|\theta)}\sum_{i=1}^{n} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta) - \lambda = 0 \tag{C-8}$$

$$\sum_{k=1}^{K} p(\mathbf{y}_i=k|\theta) - 1 = 0$$

from which we get

$$\lambda = \sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta)$$

Summing the above over k and then over i gives $\lambda = n$. Substituting n for λ in (C-8) then gives the updated estimate of the k-th prior

$$p(\mathbf{y}_i=k|\theta) = \frac{1}{n}\sum_{i=1}^{n} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta)$$

To derive an update for the parameters $\theta_k = (\mu_k, \Sigma_k)$ of the k-th Gaussian we need to maximize

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i=k|\mathbf{x}_i,\theta)\log p(\mathbf{x}_i|\mathbf{y}_i=k,\theta)$$

with respect to θ_k. Since

$$p(\mathbf{x}_i|\mathbf{y}_i=k,\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}\ \exp\left(-\frac{1}{2}(\mathbf{x}_i-\mu_k)'\Sigma_k^{-1}(\mathbf{x}_i-\mu_k)\right)$$

this is equivalent to maximizing

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p(\mathbf{y}_i = k|\mathbf{x}_i, \theta)\left(-\frac{1}{2}\ln(2\pi)^d|\Sigma_k| - \frac{1}{2}(\mathbf{x}_i - \mu_k)'\Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\right) \tag{C-9}$$

Taking the partial derivative of (C-9) with respect to μ_k we get

$$\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i, \theta)\Sigma_k^{-1}(\mathbf{x}_i - \mu_k) = \mathbf{0}$$

from which we can easily solve for μ_k to obtain the updated estimate

$$\mu_k = \frac{\sum_{i=1}^{n} \mathbf{x}_i p(\mathbf{y}_i = k|\mathbf{x}_i, \theta)}{\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i, \theta)}$$

Given that

$$\frac{\partial(x'Ay)}{\partial A} = xy' \qquad \frac{\partial(\ln|A|)}{\partial A} = (A')^{-1} \qquad |A^{-1}| = 1/|A|$$

taking the derivative of (C-9) with respect to Σ_k^{-1} we get

$$\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i, \theta)\left(\frac{1}{2}\Sigma_k - \frac{1}{2}(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)'\right) = 0$$

from which we can solve for Σ_k to obtain the updated estimate

$$\Sigma_k = \frac{\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i, \theta)(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)'}{\sum_{i=1}^{n} p(\mathbf{y}_i = k|\mathbf{x}_i, \theta)}$$

References

[Bil97] Jeff Bilmes, *A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models*, 1997, Technical Report, University of Berkeley, ICSI-TR-97-021,.

[DHS00] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification, second edition*, Wiley, 2000.

[DLR77] A. Dempster, N. Laird, and D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Series B **39** (1977), no. 1, 1–38.

[Min98] Thomas P. Minka, *Expectation-maximization as lower bound maximization*, 1998, Tutorial published on the web at http://www-white.media.mit.edu/tpminka/papers/em.html.

[MK97] Geoffrey J. McLachlan and Thriyamakam Krishnan, *The EM Algorithm and Extensions*, John Wiley and Sons, 1997.

[NH98] R. Neal and G. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, Learning in Graphical Models (M. I. Jordan, ed.), Kluwer, 1998.

[TK03] S. Theodoridis and K. Koutroumbas, *Pattern Recognition 2nd ed*, Elsevier, 2003.