The Expectation Maximization Algorithm
A short tutorial

1 Introduction

This tutorial discusses the Expectation Maximization (EM) algorithm of Dempster, Laird and Rubin [1]. The approach taken follows that of an unpublished note by Stuart Russel, but fleshes out some of the gory details. In order to ensure that the presentation is reasonably self-contained, some of the results on which the derivation of the algorithm is based are presented prior to the main results. The EM algorithm has become a popular tool in statistical estimation problems involving incomplete data, or in problems which can be posed in a similar form, such as mixture estimation [3, 4]. The EM algorithm has also been used in various motion estimation frameworks [5] and variants of it have been used in multiframe superresolution restoration methods which combine motion estimation along the lines of [2].

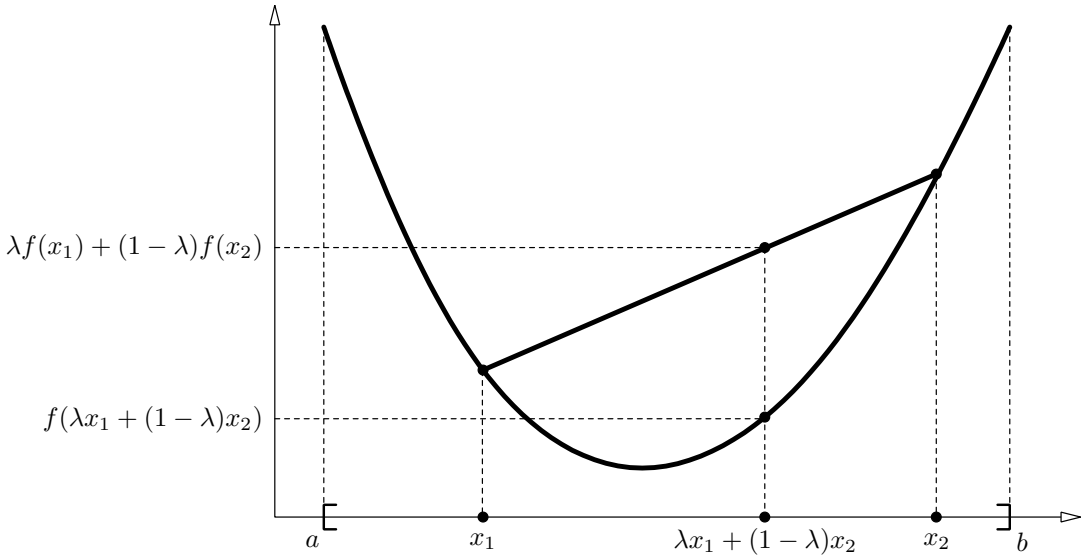

Figure 1: f is *convex* on $[a, b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a, b], \quad \lambda \in [0, 1]$.

2 Convex Functions

Definition 1 *Let f be a real valued function defined on an interval $I = [a, b]$. f is said to be* convex *on I if $\forall x_1, x_2 \in I, \lambda \in [0, 1]$,*

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

f is said to be strictly convex *if the inequality is strict. Intuitively, this definition states that the function falls below (strictly convex) or is never above (convex) the straight line (the secant) from points $(x_1, f(x_1))$ to $(x_2, f(x_2))$. See Figure (1).*

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

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: For $x \leq y \in [a, b]$ and $\lambda \in [0, 1]$ let $z = \lambda y + (1-\lambda)x$. By definition, f is convex iff $f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda) f(x)$. Writing $z = \lambda y + (1-\lambda)x$, and noting that $f(z) = \lambda f(z) + (1-\lambda) f(z)$ we have that $f(z) = \lambda f(z) + (1-\lambda) f(z) \leq \lambda f(y) + (1-\lambda) f(x)$. By rearranging terms, an equivalent definition for convexity can be obtained: f is convex if

$$\lambda \left[f(y) - f(z) \right] \geq (1-\lambda) \left[f(z) - f(x) \right] \tag{1}$$

By the mean value theorem, $\exists s, x \leq s \leq z$ s.t.

$$f(z) - f(x) = f'(s)(z - x) \tag{2}$$

Similarly, applying the mean value theorem to $f(y) - f(z)$, $\exists t, z \leq t \leq y$ s.t.

$$f(y) - f(z) = f'(t)(y - z) \tag{3}$$

Thus we have the situation, $x \leq s \leq z \leq t \leq y$. By assumption, $f''(x) \geq 0$ on $[a, b]$ so

$$f'(s) \leq f'(t) \text{ since } s \leq t. \tag{4}$$

Also note that we may rewrite $z = \lambda y + (1 - \lambda)x$ in the form

$$(1 - \lambda)(z - x) = \lambda(y - z). \tag{5}$$

Finally, combining the above we have,

$$\begin{array}{rcll} (1-\lambda)[f(z) - f(x)] & = & (1-\lambda)f'(s)(z-x) & \text{by Equation (2)} \\ & \leq & f'(t)(1-\lambda)(z-x) & \text{by Equation (4)} \\ & = & \lambda f'(t)(y-z) & \text{by Equation (5)} \\ & = & \lambda[f(y) - f(z)] & \text{by Equation (3).} \end{array}$$

■

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

Proof: With $f(x) = -ln(x)$, we have $f''(x) = \frac{1}{x^2} > 0$ for $x \in (0, \infty)$. By Theorem (1), $-\ln(x)$ is strictly convex on $(0, \infty)$. Also, by Definition (2) $\ln(x)$ is strictly concave on $(0, \infty)$.

■

The notion of convexity can be extended to apply to n points. This result is known as Jensen's inequality.

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

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

Proof: For $n = 1$ this is trivial. The case $n = 2$ corresponds to the definition of convexity. To show that this is true for all natural numbers, we proceed by induction. Assume the theorem is true for some n then,

$$\begin{array}{rcl} f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) & = & f\left(\lambda_{n+1} x_{n+1} + \sum_{i=1}^{n} \lambda_i x_i\right) \\ & = & f\left(\lambda_{n+1} x_{n+1} + (1 - \lambda_{n+1}) \frac{1}{1 - \lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i\right) \end{array}$$

$$
\begin{aligned}
&\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\left(\frac{1}{1-\lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i\right) \\
&= \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\left(\sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i\right) \\
&\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} f(x_i) \\
&= \lambda_{n+1} f(x_{n+1}) + \sum_{i=1}^{n} \lambda_i f(x_i) \\
&= \sum_{i=1}^{n+1} \lambda_i f(x_i)
\end{aligned}
$$

■

Since $\ln(x)$ is concave, we may apply Jensen's inequality to obtain the useful result,

$$\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i). \tag{6}$$

This allows us to lower-bound a logarithm of a sum, a result that is used in the derivation of the EM algorithm.

Jensen's inequality provides a simple proof that the arithmetic mean is greater than or equal to the geometric mean.

Proposition 2

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

Proof: If $x_1, x_2, \ldots, x_n \geq 0$ then, since $\ln(x)$ is concave we have

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

Thus, we have

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

■

3 The Expectation-Maximization Algorithm

The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. In ML estimation, we wish to estimate the model parameter(s) for which the observed data are the most likely.

Each iteration of the EM algorithm consists of two processes: The E-step, and the M-step. In the expectation, or E-step, the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology. In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step are used in lieu of the actual missing data.

Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

3.1 Derivation of the EM-algorithm

Let $\mathbf{X}$ be random vector which results from a parameterized family. We wish to find θ such that $\mathcal{P}(\mathbf{X}|\theta)$ is a maximum. This is known as the Maximum Likelihood (ML) estimate for θ. In order to estimate θ, it is typical to introduce the *log likelihood function* defined as,

$$L(\theta) = \ln \mathcal{P}(\mathbf{X}|\theta). \tag{7}$$

The likelihood function is considered to be a function of the parameter θ given the data $\mathbf{X}$. Since $\ln(x)$ is a strictly increasing function, the value of θ which maximizes $\mathcal{P}(\mathbf{X}|\theta)$ also maximizes $L(\theta)$.

The EM algorithm is an iterative procedure for maximizing $L(\theta)$. Assume that after the n^{th} iteration the current estimate for θ is given by θ_n. Since the objective is to maximize $L(\theta)$, we wish to compute an updated estimate θ such that,

$$L(\theta) > L(\theta_n) \tag{8}$$

Equivalently we want to maximize the difference,

$$L(\theta) - L(\theta_n) = \ln \mathcal{P}(\mathbf{X}|\theta) - \ln \mathcal{P}(\mathbf{X}|\theta_n). \tag{9}$$

So far, we have not considered any unobserved or missing variables. In problems where such data exist, the EM algorithm provides a natural framework for their inclusion. Alternately, hidden variables may be introduced purely as an artifice for making the maximum likelihood estimation of θ tractable. In this case, it is assumed that knowledge of the hidden variables will make the maximization of the likelihood function easier. Either way, denote the hidden random vector by $\mathbf{Z}$ and a given realization by $\mathbf{z}$. The total probability $\mathcal{P}(\mathbf{X}|\theta)$ may be written in terms of the hidden variables $\mathbf{z}$ as,

$$\mathcal{P}(\mathbf{X}|\theta) = \sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta). \tag{10}$$

We may then rewrite Equation (9) as,

$$L(\theta) - L(\theta_n) = \ln\left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n). \tag{11}$$

Notice that this expression involves the logarithm of a sum. In Section (2) using Jensen's inequality, it was shown that,

$$\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i)$$

for constants $\lambda_i \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$. This result may be applied to Equation (11) provided that the constants λ_i can be identified. Consider letting the constants be of the form $\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)$. Since $\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)$ is a probability measure, we have that $\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \geq 0$ and that $\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) = 1$ as required.

Then starting with Equation (11) the constants $\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)$ are introduced as,

$$\begin{aligned}
L(\theta) - L(\theta_n) &= \ln\left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \ln\left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \cdot \frac{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \ln\left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&\geq \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln\left(\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}\right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \quad (12) \\
&= \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln\left(\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\mathcal{P}(\mathbf{X}|\theta_n)}\right) \quad (13) \\
&\triangleq \Delta(\theta|\theta_n). \quad (14)
\end{aligned}$$

In going from Equation (12) to Equation (13) we made use of the fact that $\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) = 1$ so that $\ln \mathcal{P}(\mathbf{X}|\theta_n) = \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X}|\theta_n)$ which allows the term $\ln \mathcal{P}(\mathbf{X}|\theta_n)$ to be brought into the summation.

We continue by writing

$$L(\theta) \geq L(\theta_n) + \Delta(\theta|\theta_n) \tag{15}$$

and for convenience define,

$$l(\theta|\theta_n) \triangleq L(\theta_n) + \Delta(\theta|\theta_n)$$

so that the relationship in Equation (15) can be made explicit as,

$$L(\theta) \geq l(\theta|\theta_n).$$

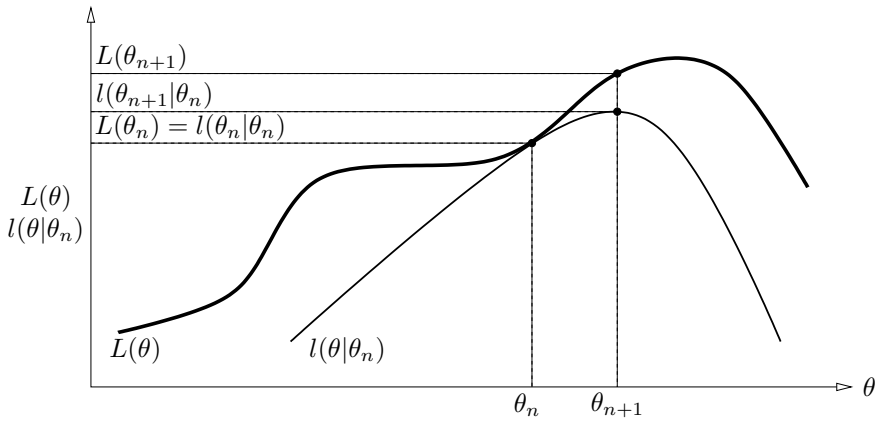

Figure 2: Graphical interpretation of a single iteration of the EM algorithm: The function $l(\theta|\theta_n)$ is bounded above by the likelihood function $L(\theta)$. The functions are equal at $\theta = \theta_n$. The EM algorithm chooses θ_{n+1} as the value of θ for which $l(\theta|\theta_n)$ is a maximum. Since $L(\theta) \geq l(\theta|\theta_n)$ increasing $l(\theta|\theta_n)$ ensures that the value of the likelihood function $L(\theta)$ is increased at each step.

We have now a function, $l(\theta|\theta_n)$ which is bounded above by the likelihood function $L(\theta)$. Additionally, observe that,

$$\begin{aligned} l(\theta_n|\theta_n) &= L(\theta_n) + \Delta(\theta_n|\theta_n) \\ &= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta_n)\mathcal{P}(\mathbf{z}|\theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\mathcal{P}(\mathbf{X}|\theta_n)} \\ &= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X},\mathbf{z}|\theta_n)}{\mathcal{P}(\mathbf{X},\mathbf{z}|\theta_n)} \\ &= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln 1 \\ &= L(\theta_n), \end{aligned} \tag{16}$$

so for $\theta = \theta_n$ the functions $l(\theta|\theta_n)$ and $L(\theta)$ are equal.

Our objective is to choose a values of θ so that $L(\theta)$ is maximized. We have shown that the function $l(\theta|\theta_n)$ is bounded above by the likelihood function $L(\theta)$ and that the value of the functions $l(\theta|\theta_n)$ and $L(\theta)$ are equal at the current estimate for $\theta = \theta_n$. Therefore, any θ which increases $l(\theta|\theta_n)$ will also increase $L(\theta)$. In order to achieve the greatest possible increase in the value of $L(\theta)$, the EM algorithm calls for selecting θ such that $l(\theta|\theta_n)$ is maximized. We denote this updated value as θ_{n+1}. This process is illustrated in Figure (2).

Formally we have,

$$\begin{aligned}
\theta_{n+1} &= \arg\max_{\theta} \{l(\theta|\theta_n)\} \\
&= \arg\max_{\theta} \left\{ L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{X}|\theta_n)\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right\} \\
&\quad \text{Now drop terms which are constant w.r.t. } \theta \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X},\mathbf{z},\theta)}{\mathcal{P}(\mathbf{z},\theta)} \frac{\mathcal{P}(\mathbf{z},\theta)}{\mathcal{P}(\theta)} \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta) \right\} \\
&= \arg\max_{\theta} \left\{ \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_n} \left\{ \ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta) \right\} \right\} \qquad (17)
\end{aligned}$$

In Equation (17) the expectation and maximization steps are apparent. The EM algorithm thus consists of iterating the:

1. *E-step*: Determine the conditional expectation $\mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_n}\{\ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta)\}$
2. *M-step*: Maximize this expression with respect to θ.

At this point it is fair to ask what has been gained given that we have simply traded the maximization of $L(\theta)$ for the maximization of $l(\theta|\theta_n)$. The answer lies in the fact that $l(\theta|\theta_n)$ takes into account the unobserved or missing data $\mathbf{Z}$. In the case where we wish to estimate these variables the EM algorithms provides a framework for doing so. Also, as alluded to earlier, it may be convenient to introduce such hidden variables so that the maximization of $L(\theta|\theta_n)$ is simplified given knowledge of the hidden variables. (as compared with a direct maximization of $L(\theta)$)

3.2 Convergence of the EM Algorithm

The convergence properties of the EM algorithm are discussed in detail by McLachlan and Krishnan [3]. In this section we discuss the general convergence of the algorithm. Recall that θ_{n+1} is the estimate for θ which maximizes the difference $\Delta(\theta|\theta_n)$. Starting with the current estimate for θ, that is, θ_n we had that $\Delta(\theta_n|\theta_n) = 0$. Since θ_{n+1} is chosen to maximize $\Delta(\theta|\theta_n)$, we then have that $\Delta(\theta_{n+1}|\theta_n) \geq \Delta(\theta_n|\theta_n) = 0$, so for each iteration the likelihood $L(\theta)$ is nondecreasing.

When the algorithm reaches a fixed point for some θ_n the value θ_n maximizes $l(\theta|\theta_n)$. Since L and l are equal at θ_n if L and l are differentiable at θ_n, then θ_n must be a stationary point of L. The stationary point need not, however, be a local maximum. In [3] it is shown that it is possible for the algorithm to converge to local minima or saddle points in unusual cases.

3.3 The Generalized EM Algorithm

In the formulation of the EM algorithm described above, θ_{n+1} was chosen as the value of θ for which $\Delta(\theta|\theta_n)$ was maximized. While this ensures the greatest increase in $L(\theta)$, it is however possible to relax the requirement of maximization to one of simply increasing $\Delta(\theta|\theta_n)$ so that $\Delta(\theta_{n+1}|\theta_n) \geq \Delta(\theta_n|\theta_n)$. This approach, to simply increase and not necessarily maximize $\Delta(\theta_{n+1}|\theta_n)$ is known as the Generalized Expectation Maximization (GEM) algorithm and is often useful in cases where the maximization is difficult. The convergence of the GEM algorithm can be argued as above.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, November 1977.

[2] R. C. Hardie, K. J. Barnard, and E. E. Armstrong. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6(12):1621–1633, December 1997.

[3] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions*. John Wiley & Sons, New York, 1996.

[4] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[5] Yair Weiss. *Bayesian motion estimation and segmentation*. PhD thesis, Massachusetts Institute of Technology, May 1998.