SUFFICIENT CONDITIONS FOR NORM CONVERGENCE OF THE EM ALGORITHM

Alfred O. Hero

Department of EECS
University of Michigan
Ann Arbor, MI 48109

Jeffrey A. Fessler

Division of Nuclear Medicine
University of Michigan
Ann Arbor, MI 48109

I. INTRODUCTION

In this paper we provide sufficient conditions for convergence of a general class of alternating estimation-maximization (EM) type continuous-parameter estimation algorithms with respect to a given norm. This class includes EM, penalized EM, Peter Green's OSL-EM, and other approximate EM algorithms. The convergence analysis can be extended to include alternating coordinate-maximization EM algorithms such as Meng and Rubin's ECM and Fessler and Hero's SAGE. For illustration, we apply our results to estimation of Poisson rate parameters in emission tomography and establish that in the final iterations the logarithm of the EM iterates converge monotonically in a weighted Euclidean norm.

Let $\theta = [\theta_1, \ldots, \theta_p]^T$ be a real parameter residing in an open subset Θ of the p-dimensional space $I\!R^p$. Given a general function $Q : \Theta \times \Theta \rightarrow I\!R$ and an initial point $\theta^0 \in \Theta$, consider the following recursive algorithm, called the A-algorithm:

$$\textbf{A-algorithm:} \qquad \theta^{i+1} = \text{argmax}_{\theta \in \Theta} Q(\theta, \theta^i). \quad (1)$$

If there are multiple maxima, then θ^{i+1} can be taken to be any one of them. Let $\theta^* \in \Theta$ be a fixed point of (1), i.e. θ^* satisfies: $\theta^* = \text{argmax}_{\theta \in \Theta} Q(\theta, \theta^*)$

The A-algorithm contains a large number of popular iterative estimation algorithms such as: the maximum-likelihood EM algorithm (ML-EM) (e.g. Dempster, Laird, and Rubin (1977), Shepp and Vardi (1982), Lange and Carson (1984), Miller and Snyder (1987), the penalized EM algorithm (e.g. Hebert and Leahy (1989)) , and EM-type algorithms implemented with E-step or M-step approximations (e.g., Antoniadis and Hero (1994), Green (1990), DePierro (1994)).

A general property of the ML-EM algorithm is that successive iterates monotonically increase the likelihood. This property guarantees global convergence when the likelihood function satisfies conditions such as boundedness and unimodality (Wu (1983), Csiszar and Tusnady (1984)). While increasing the likelihood is an attractive property, it does not guarantee monotone convergence of the parameter estimates in norm, i.e. $\|\theta^i - \theta^*\|$ may not decrease monotonically to zero as i tends to infinity. When global convergence to a given fixed point cannot be established using the methods of Wu, investigation of regions of monotone convergence can often yield a radius of convergence for the algorithm. The radius of convergence provides useful information about the proper choice of initialization θ^0 for the algorithm. Furthermore, the property of monotone convergence provides a practical verification tool for testing for errors in algorithm implementation. This tool can be used as a complement to the popular procedure of checking an implementation for the increasing-likelihood property.

II. CONVERGENCE THEOREM

A *region of monotone convergence* relative to the vector norm $\|\cdot\|$ of the A-algorithm (1) is defined as any open ball $B(\theta^*, \delta) = \{\theta : \|\theta - \theta^*\| < \delta\}$ centered at $\theta = \theta^*$ with radius $\delta > 0$ such that if the initial point θ^0 is in this region then $\|\theta^i - \theta^*\|$, $i = 1, 2, \ldots$, converges monotonically to zero. Note that as defined, the shape in $I\!R^p$ of the region of monotone convergence depends on the norm used. For the Euclidean norm $\|u\|^2 = u^T u$ the region of monotone convergence is a spherically shaped region in Θ. For a general positive definite matrix $\mathbf{B}$ the induced norm $\|u\|^2 = u^T \mathbf{B} u$ makes this region an ellipsoid in Θ. Since all norms are equivalent for the case of a finite dimensional parameter space, monotone convergence in a given norm implies convergence, however possibly non-monotone, in any other norm.

Define the $p \times p$ matrices obtained by averaging $\nabla^{20} Q(u, \overline{u})$ and $\nabla^{11} Q(u, \overline{u})$ over the line segments $u \in \overrightarrow{\theta\theta^*}$ and $\overline{u} \in \overrightarrow{\overline{\theta}\theta^*}$:

$$A_1(\theta, \overline{\theta}) = -\int_0^1 \nabla^{20} Q(t\theta + (1-t)\theta^*, t\overline{\theta} + (1-t)\theta^*) dt$$
$$A_2(\theta, \overline{\theta}) = \int_0^1 \nabla^{11} Q(t\theta + (1-t)\theta^*, t\overline{\theta} + (1-t)\theta^*) dt. \quad (2)$$

[1]This research was supported in part by the National Science Foundation under grant BCS-9024370, a DOE Alexander Hollaender Postdoctoral Fellowship, DOE Grant DE-FG02-87ER60561, and NIH grant CA-60711.

Also, define the following set:

$$\mathcal{S}(\overline{\theta}) = \{\theta \in \Theta \ : \ Q(\theta, \overline{\theta}) \geq Q(\overline{\theta}, \overline{\theta})\}.$$

By the construction of the A-algorithm (1), we have $\theta^{i+1} \in \mathcal{S}(\theta^i)$.

Definition 1 *For a given vector norm $\|\cdot\|$ and induced matrix norm $\|\!| \cdot |\!\|$ define $\mathcal{R}_+ \subset \Theta$ as the largest open ball $B(\theta^*, \delta) = \{\theta : \|\theta - \theta^*\| < \delta\}$ such that for each $\overline{\theta} \in B(\theta^*, \delta)$:*

$$A_1(\theta, \overline{\theta}) > 0, \qquad \textit{for all } \theta \in \mathcal{S}(\overline{\theta}) \tag{3}$$

and for some $0 \leq \alpha < 1$

$$\left|\!\left|\!\left| \left[A_1(\theta, \overline{\theta})\right]^{-1} \cdot A_2(\theta, \overline{\theta}) \right|\!\right|\!\right| \leq \alpha, \qquad \textit{for all } \theta \in \mathcal{S}(\overline{\theta}). \tag{4}$$

The following convergence theorem establishes that, if $\mathcal{R}_+$ is not empty, the region in Definition 1 is a region of monotone convergence in the norm $\|\cdot\|$ for an algorithm of the form (1). It can be shown that $\mathcal{R}_+$ is non-empty for sufficiently regular problems (Hero and Fessler (1995)).

Theorem 1 *Let $\theta^* \in \Theta$ be a fixed point of the A algorithm (1), where $\theta^{i+1} = argmax_{\theta \in \Theta} Q(\theta, \theta^i)$, $i = 0, 1, \ldots$. Assume: i) for all $\overline{\theta} \in \Theta$, the maximum $\max_\theta Q(\theta, \overline{\theta})$ is achieved on the interior of the set Θ; ii) $Q(\theta, \overline{\theta})$ is twice continuously differentiable in $\theta \in \Theta$ and $\overline{\theta} \in \Theta$, and iii) the A-algorithm (1) is initialized at a point $\theta^0 \in \mathcal{R}_+$ for a norm $\|\cdot\|$.*

1. *The iterates θ^i, $i = 0, 1, \ldots$ all lie in $\mathcal{R}_+$,*
2. *the successive differences $\Delta\theta^i = \theta^i - \theta^*$ of the A algorithm obey the recursion:*

$$\Delta\theta^{i+1} = [A_1(\theta^{i+1}, \theta^i)]^{-1} A_2(\theta^{i+1}, \theta^i) \cdot \Delta\theta^i. \tag{5}$$

3. *the norm $\|\Delta\theta^i\|$ converges monotonically to zero with at least linear rate, and*
4. *$\Delta\theta^i$ asymptotically converges to zero with root convergence factor*

$$\rho\left(\left[-\nabla^{20} Q(\theta^*, \theta^*)\right]^{-1} \nabla^{11} Q(\theta^*, \theta^*)\right) < 1.$$

If the iterates are initialized within a region $\mathcal{R}_+$, or for that matter if any iterate θ^i lies in $\mathcal{R}_+$, then all subsequent iterates will also lie within $\mathcal{R}_+$. Within that region, Theorem 1 provides a functional relationship (5) between successive iterates, which in turn ensures that the iterates converge monotonically in norm to θ^* with an asymptotic linear rate governed by the spectral radius of a matrix depending on the partial derivatives of Q. When specialized to the EM algorithm, the root convergence factor is equivalent to the expression obtained by Dempster, Laird and Rubin (1977) and used by Meng and Rubin (1991) to estimate the asymptotic estimator covariance matrix.

III. APPLICATION

In the ECT problem the objective is to estimate the intensity vector $\theta = [\theta_1, \ldots, \theta_p]^T$, $\theta_b \geq 0$, governing the number of gamma-ray emissions $\mathbf{N} = [\mathbf{N}_1, \ldots, \mathbf{N}_p]^T$ over an imaging volume of p pixels. The estimate of θ must be based on the projection data $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_m]^T$. The elements $\mathbf{N}_b$ of $\mathbf{N}$ are independent Poisson distributed with rate parameters θ_b, and the elements $\mathbf{Y}_d$ of $\mathbf{Y}$ are independent Poisson distributed with rate parameters $\mu_d(\theta) = \sum_{b=1}^p P_{d|b}\theta_b$, where $P_{d|b}$ is the transition probability corresponding to emissions from pixel b being detected at detector module d.

The Shepp-Vardi implementation of the ML-EM algorithm for estimating the intensity θ has the form:

$$\theta_b^{i+1} = \frac{\theta_b^i}{P_b} \sum_{d=1}^m \frac{\mathbf{Y}_d P_{d|b}}{\mu_d(\theta^i)}, \qquad b = 1, \ldots, p, \tag{6}$$

where $P_b \stackrel{def}{=} \sum_{b=1}^p P_{d|b}$ is positive under the assumption that $P_{d|b}$ has full column rank.

Using Theorem 1 we can obtain the following result:

Theorem 2 *Assume that the unpenalized ECT EM algorithm specified by (6) converges to the strictly positive limit θ^*. Then, for some sufficiently large positive integer M:*

$$\|\ln\theta^{i+1} - \ln\theta^*\| \leq \alpha \|\ln\theta^i - \ln\theta^*\|, \qquad i \geq M,$$

where $\alpha = \rho([\boldsymbol{B} + \boldsymbol{C}]^{-1}\boldsymbol{C})$, $\boldsymbol{B} = \boldsymbol{B}(\theta^*)$, $\boldsymbol{C} = \boldsymbol{C}(\theta^*)$, the norm $\|\bullet\|$ is defined as:

$$\|u\|^2 \stackrel{def}{=} \sum_{b=1}^p P_b \theta_b^* \ u_b^2, \tag{7}$$

and $P_b \stackrel{def}{=} \sum_{d=1}^m P_{d|b}$.

Lange and Carson (1984) showed that the ECT EM algorithm converges to the maximum likelihood estimate. As long as θ^* is strictly positive, the theorem asserts that in the final iterations of the algorithm the logarithmic differences $\ln\theta^i - \ln\theta^*$ converge monotonically to zero relative to the norm (7).

1. REFERENCES

[] Antoniadis, N. and Hero, A. O. (1994). Time delay estimation for filtered Poisson processes using an EM-type algorithm. *IEEE Trans. on Signal Processing* **42**, 2112-2123.

[] Csiszar, I. and Tusnady, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplement Issue No. **1**, 205–237.

[] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Ser. B* **39**, 1–38.

[] De Pierro, A. R. (1994). A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Trans. on Medical Imaging*, to appear.

[] Fessler, J. A. and Hero, A. O. (1994). Space-alternating generalized EM algorithm. *IEEE Trans. on Signal Processing*, **42**, 2664–2677.

[] Green, P. J. (1990). On the use of the EM algorithm for penalized likelihood estimation. *J. Royal Statistical Society, Ser. B* **52**, 443–452.

[] Hebert, T. and Leahy, R. (1989). A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans. on Medical Imaging* **8**, 194–203.

[] Hero, A. O. and Fessler, J. A. (1995). Convergence in norm for alternating expectation-minimization (EM) type algorithms. *Statistica Sinica* (Special issue on the EM algorithm), to appear Jan. 1995.

[] Lange, K. and Carson, R. (1984). EM reconstruction algorithms for emission and transmission tomography. *J. Comp. Assisted Tomography* **8**, 306–316.

[] Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Assoc.*, **86**, 899–909.

[] Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 267–728.

[] Miller, M. I. and Snyder, D. L. (1987). The role of likelihood and entropy in incomplete-data problems: applications to estimating point-process intensities and Toeplitz constrained covariances. *IEEE Proceedings* **75**, 892–907.

[] Shepp, L. A. and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. on Medical Imaging* **1**, 113–122.

[] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.