Convergence of EM Variants

William Byrne and Asela Gunawardana
CLSP/ECE, Johns Hopkins University, Baltimore, MD 21218
Email: zilla@jhu.edu

I. Introduction

Csiszár and Tusnády [1] present an alternative framework to that of Dempster, Laird and Rubin (DLR) [2] for describing the EM algorithm. Instead of describing an EM iteration as maximization of the EM auxiliary function, they describe it as a forward projection under the I-divergence from a model family to a desired family consisting of distributions satisfying the linear constraint that their marginal agrees with the training data, followed by a projection back to the model family.

The GEM algorithm of DLR corresponds to replacing this backward projection (the M step) by a map that decreases the divergence rather than minimizes it. We investigate the properties of algorithms in which the forward projection (the E step) is also replaced by a more general map. We also investigate the use of generalizations of the divergence when carrying out the E and M steps. We here summarize our investigation of such generalizations of the EM algorithm, and their convergence properties [3].

II. Convergence Under Extended E Steps

Wu [4] uses the convergence result of Zangwill [5] to show convergence of the (G)EM algorithm to stationary points in likelihood. Unlike Wu, who used this result on sequences of parameter values, we use it on sequences of pairs consisting of a desired distribution and a model parameter. This requires the compactness of the space of such pairs, which derives from the linear constraints that define the desired family and the compactness of the parameter family. We show that extended E and M steps which are closed and decrease the divergence between the desired distribution and the model whenever possible give algorithms which converge to EM fixed points.

To our knowledge there are no convergence results on such extensions that do not satisfy the GEM condition. For example, the AECM algorithms that are extensively studied by Meng and Van Dyk [6] are GEM procedures. Our results hold for procedures which are not GEM, such as those of Neal and Hinton [7]. They show that their extended auxiliary function is increasing, and that a stationary point of this function is also a stationary point of the likelihood, but they do not show that their extensions of EM necessarily converge to such points. In fact, there are cases where convergence to such points is not obtained. Our result mentioned above gives conditions under which such convergence is obtained.

Although these algorithms converge to EM fixed points, they may have different rates of convergence and their fixed points may have different basins of attraction. Two such algorithms are moment interpolation [8] and incremental EM [7]. The former slows down convergence, which has been found in some applications to provide a guard against overtraining [8, 9]), while the latter speeds up convergence, is not monotonic in likelihood, has different basins of attraction, but converges to the ML estimate.

III. Convergence Under Extended Divergences

We show that the convergence results obtained hold under extensions of the divergence such as the f-divergences and the Bregman distances. The EM variants obtained under the f-divergences are shown to converge to the same interior points as the standard EM algorithm. However convergence behavior and the behavior on the boundary of the model family may differ. These results hold for the α-EM algorithm of Matsuyama [10], since the α-divergence is an f-divergence. Under the Bregman distances, convergence to local maxima in likelihood is lost. However, for the class of Bregman distances discussed in [11], the convergence points lie close to the local maxima in likelihood though the convergence behavior is very different.

References

[1] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Stat. & Dec., Supp. Iss. No. 1*, pp. 205–237, 1984.

[2] A. P. Dempster, A. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data," *J. Roy. Stat. Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1997.

[3] A. Gunawardana and W. Byrne, "Convergence of EM variants," tech. rep., CLSP, Johns Hopkins University, 1998. CLSP Research Note No. 32.

[4] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, no. 1, pp. 95–103, 1983.

[5] W. I. Zangwill, *Nonlinear Programming: A Unified Approach.* Prentice-Hall, 1969.

[6] X.-L. Meng and D. van Dyk, "The EM algorithm – an old folk-song sung to a fast new tune," *J. Roy. Stat. Soc., Ser. B*, vol. 59, no. 3, pp. 511–567, 1997.

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

[8] W. J. Byrne, "Generalization and maximum likelihood from small data sets," in *IEEE-SP Wkshp. Neur. Net. Sig. Proc.*, 1993.

[9] V. Digalakis *et al.*, "Rapid speech recognizer adaptation to new speakers," in *ICASSP*, 1999. Submitted.

[10] Y. Matsuyama, "Non-logarithmic information measures, α-weighted EM algorithms and speedup of learning," in *IEEE Int. Symp. Inf. Thy.*, 1998.

[11] J. D. Lafferty *et al.*, "Statistical learning algorithms based on Bregman distances," in *Can. Wkshp. Inf. Thy.*, 1997.