Expectation-Conjugate Gradient: An Alternative to EM

Ruslan Salakhutdinov, Sam Roweis
CS Department, University of Toronto
rsalakhu,roweis@cs.toronto.edu

Zoubin Ghahramani
Gatsby Unit, Univ College London
zoubin@gatsby.ucl.ac.uk

Abstract

We show a close relationship between bound optimization (BO) algorithms such as Expectation-Maximization and direct optimization (DO) algorithms such as gradient-based methods for parameter learning. We identify analytic conditions under which BO algorithms exhibit Quasi-Newton convergence behavior, and conditions under which these algorithms possess poor, first-order convergence. In particular, for the EM algorithm we show that if a certain measure of the proportion of missing information is small, then EM exhibits Quasi-Newton behavior; when it is large, EM converges slowly. Based on this analysis, we present a new Expectation-Conjugate-Gradient (ECG) algorithm for maximum likelihood estimation, and report empirical results, showing that, as predicted by the theory, ECG outperforms EM in certain cases.

1 Introduction

Many problems in machine learning and pattern recognition ultimately reduce to the optimization of a scalar valued function $L(\Theta)$ of a free parameter vector Θ. For example, in (supervised or) unsupervised probabilistic modeling the objective function may be the (conditional) data likelihood or the posterior over parameters. In discriminative learning we may use a classification or regression score; in reinforcement learning we may use average discounted reward. Optimization may also arise during inference; for example we may want to reduce the cross entropy between two distributions or minimize a function such as the Bethe free energy.

A variety of general techniques exist for optimizing such objective functions. Broadly, they can be placed into one of two categories: direct optimization (DO) algorithms and what we will refer to as bound optimization (BO) algorithms. Direct optimization works directly with the objective and its derivatives (or estimates thereof), trying to maximize or minimize it by adjusting the free parameters in a local search. This category of algorithms includes random search, standard gradient-based algorithms, line search methods such as conjugate gradient (CG), and more computationally intensive second-order methods, such as Newton-Raphson. They can be applied, in principle, to any deterministic function of the parameters. Bound optimization, on the other hand, takes advantage of the fact that many objective functions arising in practice have a special structure. We can often exploit this structure to obtain a bound on the objective function and proceed by optimizing this bound. Ideally, we seek a bound that is valid everywhere in parameter space, easily optimized, and equal to the true objective function at one (or more) point(s). A general form of a bound maximizer which iteratively lower bounds the objective function is given below:

General form of Bound Optimization for maximizing $L(\Theta)$:

- **Assume**: $\exists$ functions $Q(\Theta|z)$ and $z(\Theta)$ such that:
    1. $Q(\Theta|z(\Theta)) = L(\Theta) \geq Q(\Theta|\bar{z})$ for any Θ, and any $\bar{z} \neq z(\Theta)$.
    2. $\arg\max_{\Theta} Q(\Theta|z(\bar{\Theta}))$ can be found easily for any $\bar{\Theta}$
- **Iterate**: $\Theta^{t+1} = \arg\max_{\Theta} Q(\Theta|z(\Theta^t))$
- **Guarantee**: $L(\Theta^{t+1}) = Q(\Theta^{t+1}|z(\Theta^{t+1})) \geq Q(\Theta^{t+1}|z(\Theta^t)) \geq Q(\Theta^t|z(\Theta^t)) = L(\Theta^t)$

Many popular iterative algorithms are bound optimizers, including the EM algorithm for maximum likelihood learning in latent variable models[3], iterative scaling (IS) algorithms for parameter estimation in maximum entropy models[2] and the recent CCCP algorithm for minimizing the Bethe free energy in approximate inference problems[13]. Bound optimization algorithms enjoy a strong guarantee; they never worsen the objective function.

In this paper we study the relationship between direct and bound optimizers and determine conditions under which one technique can be expected to outperform another. Our general results apply to any model for which a bound optimizer can be constructed, although in later sections we focus on the case of probabilistic models with latent variables.

2 Gradient and Newton behaviors of bound optimization

For most objective functions, the BO step $\Theta^{(t+1)} - \Theta^{(t)}$ in parameter space and true gradient can be trivially related by *transformation matrix* $P(\Theta^t)$, that changes at each iteration:

$$\Theta^{(t+1)} - \Theta^{(t)} = P(\Theta^t)\nabla_L(\Theta^t) \tag{1}$$

(We define $\nabla_L(\Theta^t) = \frac{\partial L(\Theta)}{\partial \Theta}|_{\Theta=\Theta^t}$.) Furthermore, under certain conditions, $P(\Theta^t)$ is guaranteed to be positive definite with respect to the gradient.[1] In particular, **if**

C1: *$Q(\Theta|z(\Theta^t))$ is well-defined, continuous and differentiable everywhere in Θ and $\boldsymbol{z}$.* **and**
C2: *For any fixed $\Theta^t \neq \Theta^{(t+1)}$, $Q(\Theta|z(\Theta^t))$ has only a single critical point along any direction, located at the maximum Θ^{t+1}* ; **then**

$$\nabla_L^\top(\Theta^t)P(\Theta^t)\nabla_L(\Theta^t) > 0 \quad \forall\Theta^t \tag{2}$$

The second condition **C2** may seem very strong. However, this condition is satisfied in many cases, for example when considering the EM-algorithm as a bound optimizer, C2 is satisfied whenever the M-step has a single unique solution.

The important consequence of the above analysis is that when the bound function has a unique optimum, BO has the appealing quality of always taking a step $\Theta^{(t+1)} - \Theta^t$ having positive projection onto the true gradient of the objective function $L(\Theta^t)$. This makes BO similar to first order methods operating on the gradient of a locally reshaped likelihood function.

For maximum likelihood learning of a mixture of Gaussians model using the EM-algorithm, this positive definite matrix was first described by Xu and Jordan[12]. In the appendix, we extend their results by deriving the explicit form for the transformation matrix in mixture of FAs[4]; we have also derived the equivalent results for several other latent variables models such as Factor Analysis (FA), Probabilistic Principal Component Analysis (PPCA), mixture of PPCAs, and Hidden Markov Models (HMM)[10].

[1] Note that $\nabla_Q^\top(\Theta^t)(\Theta^{(t+1)} - \Theta^t)$, where $\nabla_Q^\top(\Theta^t) = \frac{\partial Q(\Theta|z(\Theta^t))}{\partial \Theta}|_{\Theta=\Theta^t}$ is the directional derivative of function $Q(\Theta|z(\Theta^t))$ in the direction of $\Theta^{(t+1)} - \Theta^t$. C1 and C2 together imply that this quantity is positive, otherwise by the Mean Value Theorem (C1) $Q(\Theta|z(\Theta^t))$ would have a critical point along some direction, located at a point other than Θ^{t+1} (C2). By using the the identity $\nabla_L(\Theta^t) = \frac{\partial Q(\Theta|z(\Theta^t))}{\partial \Theta}|_{\Theta=\Theta^t}$, we have $\nabla_L^\top(\Theta^t)P(\Theta^t)\nabla_L(\Theta^t) = \nabla_Q^\top(\Theta^t)(\Theta^{(t+1)} - \Theta^t) > 0$.

We can further study the structure of the transformation matrix $P(\Theta^t)$ by considering the mapping defined by one step of BO: $\Theta^{(t+1)} = M(\Theta^t)$. Taking derivatives of both sides of (1) with respect to Θ, we have

$$I - M'(\Theta^t) = -P'(\Theta^t)\nabla_L(\Theta^t) - P(\Theta^t)S(\Theta^t) \tag{3}$$

where $S(\Theta^t) = \frac{\partial^2 L(\Theta)}{\partial \Theta^2}|_{\Theta=\Theta^t}$ is the Hessian of the objective function, $M'_{ij}(\Theta^t) = \frac{\partial \Theta_i^{t+1}}{\partial \Theta_j^t}$ is the input-output derivative matrix for the BO mapping and $P'(\Theta^t) = \frac{\partial P(\Theta^t)}{\partial \Theta}|_{\Theta=\Theta^t}$ is the tensor derivative of $P(\Theta^t)$ with respect to Θ. In "flat" regions of $L(\Theta)$, where $\nabla_L(\Theta)$ approaches zero (and $P'(\Theta^t)$ does not become infinite), the first term on the RHS of equation (3) becomes much smaller than the second term, and the transformation matrix becomes a rescaled version of the negative inverse Hessian:

$$P(\Theta^t) = \left[I - M'(\Theta^t)\right]\left[-S(\Theta^t)\right]^{-1} \tag{4}$$

In particular, if the bound optimization algorithm iterates converge to a local optima at Θ^*, then near this point (i.e. for sufficiently large t) BO exhibits Quasi-Newton convergence behavior. This is also true in “plateau” regions where the gradient is very small even if they are not near a local optimum.

The nature of the Quasi-Newton behavior is controlled by the eigenvalues of the matrix $M'(\Theta^t)$. If all eigenvalues tend to zero, then BO becomes a true Newton method, rescaling the gradient by exactly the negative inverse Hessian. As the eigenvalues tend to unity, BO takes smaller and smaller stepsizes, giving poor, first-order, convergence.

3 The Expectation-Maximization (EM) Algorithm

We now consider a particular bound optimizer, the popular Expectation-Maximization (EM) algorithm and derive specific cases of the results above for models which use EM to adjust their parameters. To begin, consider a probabilistic model of observed data $\boldsymbol{x}$ which uses latent variables $\boldsymbol{y}$. The log likelihood function can be written as a difference between two terms:

$$\begin{aligned} L(\Theta) &= \ln p(\boldsymbol{x}|\Theta) = \int_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x},\psi)\ln p(\boldsymbol{x}|\Theta)d\boldsymbol{y} \qquad (5) \\ &= \int_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x},\psi)\ln p(\boldsymbol{x},\boldsymbol{y}|\Theta)d\boldsymbol{y} - \int_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x},\psi)\ln p(\boldsymbol{y}|\boldsymbol{x},\Theta)d\boldsymbol{y} \\ &= Q(\Theta|z(\psi)) - H(\Theta|z(\psi)) \end{aligned}$$

Dempster, Laird, and Rubin [3] showed that if EM iterates converge to Θ^*, then

$$\frac{\partial M(\Theta)}{\partial \Theta}|_{\Theta=\Theta^*} = \left[\frac{\partial^2 H(\Theta|\Theta^*)}{\partial \Theta^2}|_{\Theta=\Theta^*}\right]\left[\frac{\partial^2 Q(\Theta|\Theta^*)}{\partial \Theta^2}|_{\Theta=\Theta^*}\right]^{-1} \tag{6}$$

which can be interpreted as the ratio of missing information to the complete information near the local optimum. Thus, in the neighbourhood of a solution (for sufficiently large t),

$$P(\Theta^t) = \left[I - \left(\frac{\partial^2 H}{\partial \Theta^2}\right)\left(\frac{\partial^2 Q}{\partial \Theta^2}\right)^{-1}|_{\Theta=\Theta^t}\right]\left[-S(\Theta^t)\right]^{-1} \tag{7}$$

This formulation of the EM algorithm has a very interesting interpretation which is applicable to any latent variable model: *When the missing information is small compared to the complete information, EM exhibits a Quasi-Newton behavior and enjoys fast, typically superlinear convergence in the neighborhood of* Θ^*. If fraction of missing information approaches unity, the eigenvalues of the first term above approach zero and EM will exhibit

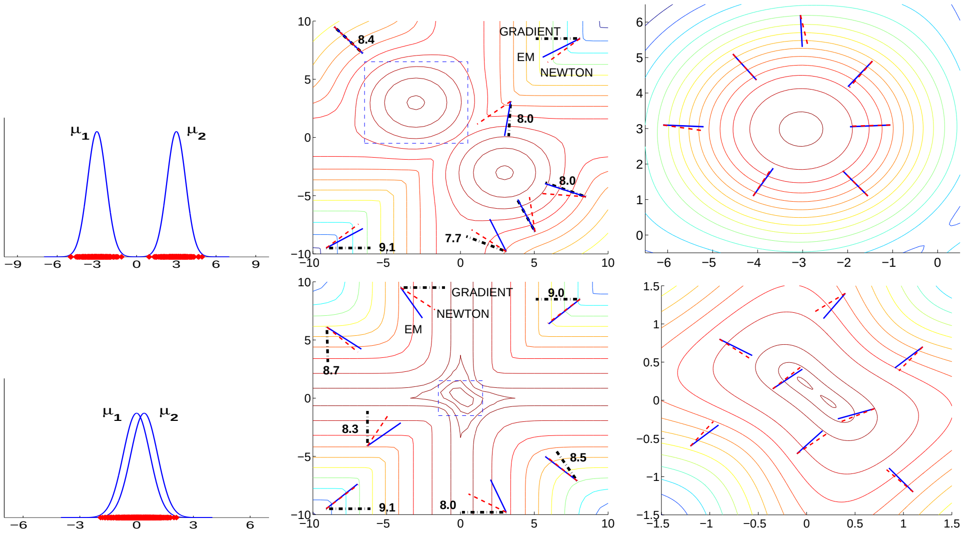

Figure 1: Contour plots of the likelihood function $L(\Theta)$ for MoG examples using well-separated (upper panels) and not-well-separated (lower panels) one-dimensional datasets. Axes correspond to the two means. The dashdot line shows the direction of the true gradient $\nabla_L(\Theta)$, the solid line shows the direction of $P(\Theta)\nabla_L(\Theta)$ and the dashed line shows the direction of $(-S)^{-1}\nabla_L(\Theta)$. Right panels are blowups of dashed regions on the left. The numbers indicate the log of the l_2 norm of the gradient. Note that for the "well-separated" case, in the vicinity of the maximum, vectors $P(\Theta)\nabla_L(\Theta)$ and $(-S)^{-1}\nabla_L(\Theta)$ become identical.

extremely slow convergence. Figure 1 illustrates these results in the simple case of fitting a mixture of Gaussians model to well-clustered and not-well-clustered data.

This analysis motivates the use of alternative optimization techniques in the regime where missing information is high and EM is likely to perform poorly. In the following section, we present exactly such an alternative, the Expectation-Conjugate Gradient (ECG) algorithm, a novel and simple direct optimization method for optimizing the parameters of latent variables models. We go on to show experimentally that ECG can in fact outperform EM under the conditions described above.

4 Expectation Conjugate Gradient (ECG) Algorithm

The key idea of the ECG algorithm is to note that if we can easily compute the derivative $\frac{\partial}{\partial\Theta}\ln p(\boldsymbol{x},\boldsymbol{z}|\Theta)$ of the complete log likelihood, then knowing the posterior $p(\boldsymbol{z}|\boldsymbol{x},\Theta)$ we can compute the exact gradient $\nabla_L(\Theta)$. In particular: $\nabla_L(\Theta) = \int_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x},\Theta)\frac{\partial}{\partial\Theta}\log p(\boldsymbol{x},\boldsymbol{z}|\Theta)d\boldsymbol{z}$. This exact gradient can then be utilized in any standard manner, for example to do gradient (as)descent or to control a line search technique. As an example, we describe a conjugate gradient algorithm:

Expectation-Conjugate-Gradient (ECG) algorithm:

- Apply a conjugate gradient optimizer to $L(\Theta)$, performing an "EG" step whenever the value or gradient of $L(\Theta)$ is requested (e.g. during a line search).
- The gradient computation is given by
  E-Step: Compute posterior $p(\boldsymbol{z}|\boldsymbol{x},\Theta^t)$ and log-likelihood $L(\Theta)$ as normal.
  G-Step: $\nabla_L(\Theta^t) = \int_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x},\Theta^t)\frac{\partial}{\partial\Theta}\log p(\boldsymbol{x},\boldsymbol{z}|\Theta)d\boldsymbol{z}$

5 Experimental Results

We now present empirical results comparing the performance of EM and ECG for learning the parameters[2] of three well know latent variable models: Mixtures of Gaussians (MoG), Probabilistic PCA (PPCA), and Hidden Markov Models (HMM). The models were trained on different data sets and with different initial conditions to illustrate both the regime in which ECG is superior to EM and in which it is inferior. Figure 2 summarizes our results: for "well-separated", "low-rank", or "structured" data in which the fraction of missing information is small, EM converges quickly; for "overlapping", "ill-conditioned" or "aliased" data, where the latent variables are poorly determined, ECG significantly outperforms EM.

First, consider a mixture of Gaussians (MoG) model. For visualization purposes, we have plotted and learned only the values of the means μ_i, fixing the mixing proportions π_i, and variances σ_i^2. We considered two types of datasets, one in which the data is "well-separated" into distinct clusters and another "not-well-separated" case in which the data overlaps in one contiguous region. Figure 2 shows that ECG outperforms EM in the poorly separated cases. For the well-separated cases, in the vicinity of the local optima Θ^* the directions of the vectors $P(\Theta)\nabla_L(\Theta)$ and $(-S)^{-1}\nabla_L(\Theta)$ become identical (fig. 1), suggesting EM will have a Quasi-Newton type convergence behavior. For the not well-separated case, this is generally not true.

[2] To conduct unconstrained optimization, we use simple reparameterizations of the model parameters. In MoG model we use softmax parameterization of the mixing coefficients $\pi_i = \frac{\exp(\gamma_i)}{\sum_{j=1}^{M} \exp(\gamma_j)}$, for covariance matrices to be symmetric positive definite, we use the Choleski decomposition. To keep diagonal entries of the noise models in FA/PPCA positive, we set $\epsilon = \exp\gamma$, and in HMMs, we reparameterize probabilities via softmax functions as well.

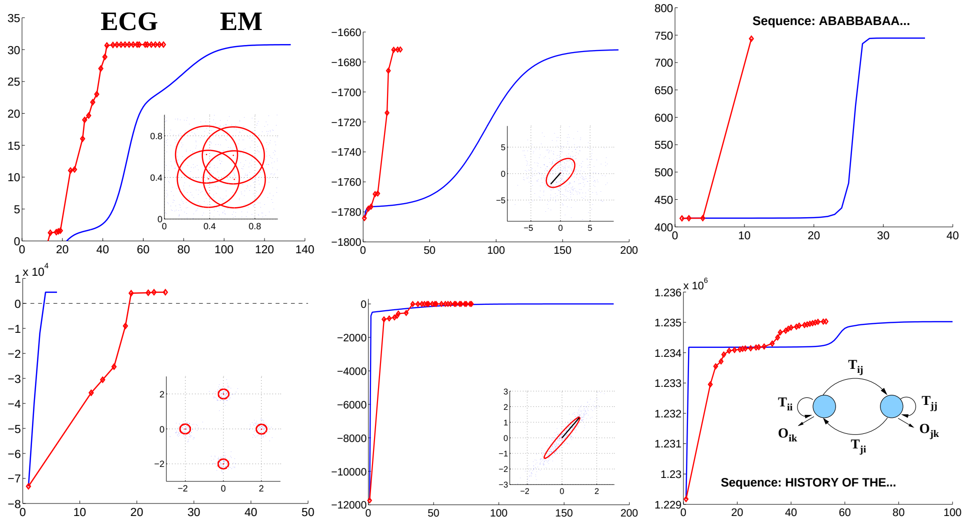

Figure 2: Learning curves for ECG (dots) and EM (solid lines) algorithms, showing superior (upper panels) and inferior (lower panels) performance of ECG under different conditions for three models: MoG (left), PPCA (middle), and HMM (right). The number of E-steps taken by either algorithm is shown on the horizontal axis, and log likelihood is shown on the vertical axis. For ECG, diamonds indicate the maximum of each line search. The zero-level for likelihood corresponds to fitting a single Gaussian density for MoG and PPCA, and to fitting a histogram using empirical symbol counts for HMM. The bottom panels use "well-separated", "low-rank", or "structured" data for which EM converges quickly; the upper panels use "overlapping", "ill-conditioned" or "aliased" data for which ECG performs much better.

We also experimented with the Probabilistic Principal Component Analysis (PPCA) latent variable model[9, 11], which has continuous rather than discrete hidden variables. Here the concept of missing information is related to the ratios of the leading eigenvalues of the sample covariance, which corresponds to the ellipticity of the distribution. For "low-rank" data with a large ratio, our experiments show that EM performs well; for nearly circular data ECG converges faster.

As a confirmation that this behavior is in accordance with our analysis, in figure 3 we show the evolution of the eigenvalues of the matrix $\left(\frac{\partial^2 H}{\partial \Theta^2}\right)\left(\frac{\partial^2 Q}{\partial \Theta^2}\right)^{-1}$ during learning of the same datasets, generated from known parameters for which we can compute this missing information matrix exactly. For the well-separated MoG case the eigenvalues of the matrix approach zero, and the ratio of missing information to the complete information becomes very small, driving $P(\Theta)$ toward the negative of the inverse Hessian. Interestingly, in the case of PPCA, even though the rank of the matrix approaches zero, one of its eigenvalues remains nonzero even in the low-rank data case (fig. 3). This suggests that the convergence of the EM algorithm for PPCA can still be slow very close to the optimum in certain directions in parameter space, even for "nice" data.[3] Hence, direct optimization methods may be preferred for the final stages of learning, even in these cases.

Finally, we applied our algorithm to the training of Hidden Markov Models (HMMs). A simple 2-state HMM (see inset fig. 2) was trained to model sequences of discrete symbols. Missing information in this model is high when the observed data do not well determine the underlying state sequence (given the parameters). In one case ("aliased" sequences), we used sequences from a two-symbol alphabet consisting of alternating "AB..." of length 600 (with probability of alternation 95% and probablity of repeating 5%). In the other case ("structured sequences"), the training data consisted of 41 character sequences from the book "Decline and Fall of the Roman Empire" by Gibbon, with an alphabet size of 30 characters. (Parameters were initialized to uniform values plus small noise.) Once again,

[3] The slow convergence of EM in PPCA is also true for FA and especially for linear dynamic systems. In these models, there is large amount of missing information due to the fact that latent variables are continuous and they can be rotated without affecting the likelihood as long as the parameters are rotated accordingly.

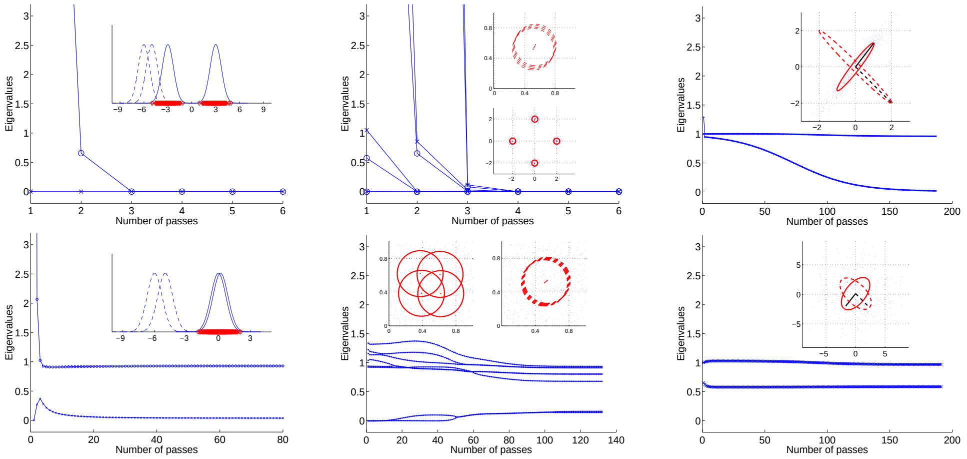

Figure 3: Eigenvalues of matrix in eq.(6) for MoG (left, centre) and PPCA (right) using well-separated or low-rank (top) and not-well-separated or circular (bottom) datasets of 3000 points. Insets show data and initial/final conditions: initial conditions are shown as dashed circles at the 1-σ^2 contour of each model; final converged model is shown by solid circles.

we observe that for the ambiguous or aliased data, ECG outperforms EM substantially. For real, structured data, ECG slightly outperforms EM as well.

6 Discussion

In this paper we have presented comparative analysis of the bound and direct optimization algorithms, and built up the connection between those two classes of optimizers. We have also analyzed and determined conditions under which BO algorithms can demonstrate local-gradient and Quasi-Newton convergence behaviors. In particular, we gave a new analysis of the EM algorithm by showing that if the fraction of missing information is small, EM is expected to have Quasi-Newton behavior near local optima.

Motivated by these analyses, we have proposed a novel direct optimization method (ECG) that can significantly outperform EM in some cases. We tested this algorithm on several basic latent variable models, showing regimes in which it is both superior and inferior to EM and explaining these behaviors with reference to our analysis.

Previous studies have considered the convergence properties of the EM algorithm in specific cases. Xu and Jordan[12], Ma, Xu and Jordan[7] studied a relationship between EM and gradient-based methods for ML learning of finite Gaussian mixture models. These authors state conditions under which EM can approximate a superlinear method (but only in the MoG setting), and give general preference to EM over gradient-based methods. Redner and Walker[8], on the other hand, argued that the speed of EM convergence can be extremely slow, and that second-order methods should generally be favored to EM.

Many methods have also been proposed to enhance the convergence speed of the EM algorithm, mostly based on the conventional optimization theory. Louis[6] proposed an approximate Newton's method, known as *Turbo EM*, that makes use of Aitken's acceleration method to yield the next iterate. Jamshidian and Jennrich[5] proposed accelerating the EM algorithm by applying generalized conjugate gradient algorithm. Other authors (Redner and Walker[8], Atkinson[1]) have proposed hybrid approaches for learning, advocating switching to a Newton or Quasi-Newton method after performing several EM iterations. All of the methods, although sometimes successful in terms of convergence, are much more complex than EM, and difficult to analyze; thus they have not been popular in practice.

While BO algorithms have played a dominating role in learning with hidden variables and in some approximate inference procedures, our results suggest that it is important not to underestimate the power of DO methods. Our analysis has indicated when one strategy may outperform another; however it is limited by being valid only in the neighbourhood of optima or plateaus and also by requiring the computation of quantities not readily available at runtime. The key to practical speedups will be the ability to design hybrid algorithms which can detect on the fly when to use bound optimizers like EM and when to switch to direct optimizers like ECG via efficiently estimating the local missing information ratio.

Acknowledgments

We would like to thank Yoshua Bengio, Drew Bagnell, and Max Welling for many useful comments and Carl Rasmussen for providing an initial version of conjugate gradient code.

References

[1] S.E. Atkinson. The performance of standard and hybrid EM algorithms for ML estimates of the normal mixture model with censoring. *J. of Stat. Computation and Simulation*, 44, 1992.

[2] Stephen Della Pietra, Vincent J. Della Pietra, and John D. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

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

[4] Zoubin Ghahramani and Geoffrey Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Dept. of Computer Science, University of Toronto, May 1996.

[5] Mortaza Jamshidian and Robert I. Jennrich. Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association*, 88(421):221–228, March 1993.

[6] T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society series B*, 44:226–233, 1982.

[7] Jinwen Ma, Lei Xu, and Michael Jordan. Asymptotic convergence rate of the EM algorithm for gaussian mixtures. *Neural Computation*, 12(12):2881–2907, 2000.

[8] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[9] S. T. Roweis. EM algorithms for PCA and SPCA. In *Advances in neural information processing systems*, volume 10, pages 626–632, Cambridge, MA, 1998. MIT Press.

[10] Ruslan Salakhutdinov. Relationship between gradient and EM steps for several latent variable models. `http://www.cs.toronto.edu/∼rsalakhu/ecg`.

[11] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[12] L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1):129–151, 1996.

[13] Alan Yuille and Anand Rangarajan. The convex-concave computational procedure (CCCP). In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.

Appendix: Explicit relationships between EM step and gradient

In this section, we derive the exact relationship between gradient of log-likelihood and the step EM performs in the parameter space for the Mixture of Factor Analyzers (MFA) model, extending the results of Xu and Jordan[12]. The derivation can be easily modified to yield identical result for PPCA, FA, Mixture of PPCA, and HMM models. The log-likelihood function for MFA model with parameters $\{\pi_i, \mu_i, \Lambda_i, \Psi_i\}_{i=1}^{M}$ is $L(\Theta) = \sum_n \ln \sum_{i=1}^{M} \pi_i \mathcal{N}(x_n|\mu_i, \Lambda_i\Lambda_i^T + \Psi_i)$.[4]

At each iteration of EM algorithm we have

$$P_{\Pi}^{(t)} = \frac{1}{N}\left[\text{diag}[\pi_1^{(t)}, ..., \pi_M^{(t)}] - \Pi^{(t)}(\Pi^{(t)})^T\right]; \quad P_{\mu_i}^{(t)} = \frac{\Psi_i^{(t)}}{\sum_n h_i^{(t)}(x_n)}$$

$$P_{\Lambda_i}^{(t)} = \left(\sum_n h_i^{(t)}(x_n) E_i^{(t)}(x_n)\right)^{-1} \otimes \Psi_i^{(t)}$$

$$P_{\Psi_i}^{(t)} = \frac{2}{\sum_n h_i^{(t)}(x_n)}\text{diag}^*\left[\left(\Lambda_i^{(t)}(\Lambda_i^{(t)})^T + \Psi_i^{(t)}\right) \otimes \Psi_i^{(t)}\right]$$

where Π denotes mixing coefficients, $\Pi = [\pi_1, ..., \pi_M]^T$, $E_i(x_n) \equiv I - \beta_i\Lambda_i + \beta_i(x_n - \mu_i)(x_n - \mu_i)^T\beta_i^T$ with $\beta_i \equiv \Lambda_i^T(\Lambda_i\Lambda_i^T + \Psi_i)$, $h_i(x_n)$ are the responsibilities, diag*(A) sets appropriate rows of the matrix A to zero, and "$\otimes$" denotes the Kroneker product.

Using the notation $\Theta = \left[\Pi^T, \mu_1^T, ..., \mu_M^T, \text{vec}[\Lambda_1]^T, ..., \text{vec}[\Lambda_M]^T, \text{vec}[\Psi_1]^T, ..., \text{vec}[\Psi_M]^T\right]^T$, and $P(\Theta) = \text{diag}[P_{\Pi}, P_{\mu_1}, ..., P_{\mu_M}, P_{\Lambda_1}, ..., P_{\Lambda_M}, P_{\Psi_1}, ..., P_{\Psi_M}]$ we can write

$$\Theta^{(t+1)} = \Theta^{(t)} + P(\Theta^t)\frac{\partial L(\Theta)}{\partial \Theta}|_{\Theta=\Theta^{(t)}} \tag{8}$$

The reader can easily verify the validity of this symmetric positive definite transformation matrix by multiplying it by the gradient of the log likelihood function.

The general form of the transformation matrix P can also be easily derived for the regular exponential family in terms of its natural parameters[10]. The P matrix is positive definite with respect to the gradient (by C1 and C2) due to the well-known convexity property of $Q(\Theta|z(\Theta^t))$.

[4] In regular Mixture of Factor Analyzers model, the the isotropic noise covariance Ψ is fixed across all component densities[4]. In our derivation we have different noise models across different component densities.