Optimization with EM and Expectation-Conjugate-Gradient

Ruslan Salakhutdinov RSALAKHU@CS.TORONTO.EDU
Sam Roweis ROWEIS@CS.TORONTO.EDU
Department of Computer Science, University of Toronto
6 King's College Rd, M5S 3G4, Canada

Zoubin Ghahramani ZOUBIN@GATSBY.UCL.AC.UK
Gatsby Computational Neuroscience Unit, University College London
17 Queen Square, London WC1N 3AR, UK

Abstract

We show a close relationship between the Expectation - Maximization (EM) algorithm and direct optimization algorithms such as gradient-based methods for parameter learning. We identify analytic conditions under which EM exhibits Quasi-Newton behavior, and conditions under which it possesses poor, first-order convergence. Based on this analysis, we propose two novel algorithms for maximum likelihood estimation of latent variable models, and report empirical results showing that, as predicted by the theory, the proposed new algorithms can substantially outperform standard EM in terms of speed of convergence in certain cases.

1. Introduction

The problem of Maximum Likelihood (ML) parameter estimation for latent variable models is an important problem in the area of machine learning and pattern recognition. ML learning with unobserved quantities arises in many probabilistic models such as density estimation, dimensionality reduction, or classification, and generally reduces to a relatively hard optimization problem in terms of the model parameters after the hidden quantities have been integrated out.

A common technique for ML estimation of model parameters in the presence of latent variables is Expectation-Maximization (EM) algorithm [3]. The EM algorithm alternates between estimating the unobserved variables given the current model and refitting the model given the estimated, complete data. As such it takes discrete steps in parameter space similar to to first order method operating on the gradient of a locally reshaped likelihood function.

In spite of tremendous success of the EM algorithm in practice due to its simplicity and fast initial progress, some authors [8] have argued that the speed of EM convergence can be extremely slow, and that more complicated second-order methods should generally be favored to EM. Many methods have been proposed to enhance the convergence speed of the EM algorithm, mostly based on conventional optimization theory [6, 7]. Several authors [8, 1] have also proposed hybrid approaches for ML learning, advocating switching to a Newton or Quasi-Newton method after performing several EM iterations. All of these approaches, 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.

Our goal in this paper is to contrast the EM algorithm with a direct gradient-based optimization approach. As a concrete alternative, we present an Expectation-Conjugate-Gradient (ECG) algorithm for maximum likelihood estimation in latent variable models, and show that it can outperform EM in terms of convergence in certain cases. However, in other cases the performance of EM is superior. To understand these behaviours, we study the convergence properties of the EM algorithm and identify analytic conditions under which EM algorithm exhibits Quasi-Newton convergence behavior, and conditions under which it possesses extremely poor, first-order convergence. Based on this analysis, we introduce a simple hybrid EM-ECG algorithm that switches between EM and ECG based on estimated quantities suggested by our analysis. We report empirical results on synthetic as well as real-world data sets, showing that, as predicted by the theory, this simple algorithm almost never performs worse than standard EM and can substantially outperform EM's convergence.

2. Linear and Newton Convergence of Expectation Maximization

We first focus on the analysis of the convergence properties of the Expectation-Maximization (EM) algorithm. Consider a probabilistic model of observed data $\boldsymbol{x}$ which uses latent variables $\boldsymbol{y}$. The log-likelihood (objective function) can be written as a difference between expected complete log-likelihood and negative entropy terms:

$$L(\Theta) = \ln p(\boldsymbol{x}|\Theta) = \int p(\boldsymbol{y}|\boldsymbol{x},\Psi)\ln p(\boldsymbol{x}|\Theta)d\boldsymbol{y} = \int p(\boldsymbol{y}|\boldsymbol{x},\Psi)\ln p(\boldsymbol{x},\boldsymbol{y}|\Theta)d\boldsymbol{y} - \int p(\boldsymbol{y}|\boldsymbol{x},\Psi)\ln p(\boldsymbol{y}|\boldsymbol{x},\Theta)d\boldsymbol{y} = Q(\Theta,\Psi) - H(\Theta,\Psi)$$

The EM algorithm implicitly defines a mapping: $M : \Theta \to \Theta'$ from parameter space to itself, such that $\Theta^{t+1} = M(\Theta^t)$. If iterates of Θ^t converge to Θ^* (and $M(\Theta)$ is continuous), then $\Theta^* = M(\Theta^*)$, and in the neighbourhood of Θ^*, by a Taylor series expansion:

$$\Theta^{t+1} - \Theta^* = M'(\Theta^*)(\Theta^t - \Theta^*) \quad (1)$$

where $M'(\Theta^*) = \frac{\partial M}{\partial \Theta}|_{\Theta=\Theta^*}$. Therefore EM is essentially a linear iteration algorithm with a convergence rate matrix $M'(\Theta^*)$ (which is typically nonzero).

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

$$\Theta^{(t+1)} - \Theta^{(t)} = P(\Theta^t)\nabla_L(\Theta^t) \quad (2)$$

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

C1: *Expected complete log-likelihood* $Q(\Theta,\Theta^t)$ *is well-defined, and differentiable everywhere in* Θ. **and**

C2: *For any fixed* $\Theta^t \neq \Theta^{(t+1)}$, $Q(\Theta,\Theta^t)$ *has only a single critical point along any direction in* Θ, *located at the maximum* $\Theta = \Theta^{t+1}$; **then**

$$\nabla_L^\top(\Theta^t)P(\Theta^t)\nabla_L(\Theta^t) > 0 \quad \forall\Theta^t \quad (3)$$

The second condition **C2** may seem very strong. However, for the EM algorithm C2 is satisfied whenever the M-step has a single unique solution.[2]

The important consequence of the above analysis is that (when the expected complete log-likelihood function has a unique optimum), EM 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 EM similar to a first order method 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 transformation matrix $P(\Theta^t)$ was first described by Xu and Jordan[15]. We extended their results by deriving the explicit form for the transformation matrix for several other latent variables models such as Factor Analysis (FA), Probabilistic Principal Component Analysis (PPCA), mixture of PPCAs, mixture of FAs, and Hidden Markov Models [10].[3]

We can further study the structure of the transformation matrix $P(\Theta^t)$ and relate it to the convergence rate matrix M'. Taking the negative derivatives of both sides of (2) with respect to Θ^t, we have:

$$I - M'(\Theta^t) = -P'(\Theta^t)\nabla_L(\Theta^t) - P(\Theta^t)S(\Theta^t) \quad (4)$$

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 EM mapping and $P'(\Theta^t) = \frac{\partial P(\Theta)}{\partial \Theta}|_{\Theta=\Theta^t}$ is the tensor derivative of $P(\Theta^t)$ with respect to Θ^t. 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 (4) becomes much smaller than the second term, and the transformation matrix becomes a rescaled version of the negative inverse Hessian:

$$P(\Theta^t) \approx \Big[I - M'(\Theta^t)\Big]\Big[-S(\Theta^t)\Big]^{-1} \quad (5)$$

In particular, if the EM algorithm iterates converge to a local optima at Θ^*, then near this point (i.e. for sufficiently large t) EM may exhibit 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 EM becomes a true Newton

[1] Note that $\nabla_Q^\top(\Theta^t)(\Theta^{(t+1)} - \Theta^t)$, where $\nabla_Q^\top(\Theta^t) = \frac{\partial Q(\Theta,\Theta^t)}{\partial \Theta}|_{\Theta=\Theta^t}$ is the directional derivative of function $Q(\Theta,\Theta^t)$ 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) $Q(\Theta,\Theta^t)$ would have a critical point along some direction, located at a point other than Θ^{t+1} (C2). By using the identity $\nabla_L(\Theta^t) = \frac{\partial Q(\Theta,\Theta^t)}{\partial \Theta}|_{\Theta=\Theta^t}$, we have $\nabla_L^\top(\Theta^t)P(\Theta^t)\nabla_L(\Theta^t) = \nabla_Q^\top(\Theta^t)(\Theta^{(t+1)} - \Theta^t) > 0$.

[2] In particular C2 holds for any exponential family model, due to the well-known convexity property of $Q(\Theta,\Theta^t)$ for these models.

[3] We also derived the general form of the transformation matrix for the exponential family models in term of their natural parameters.

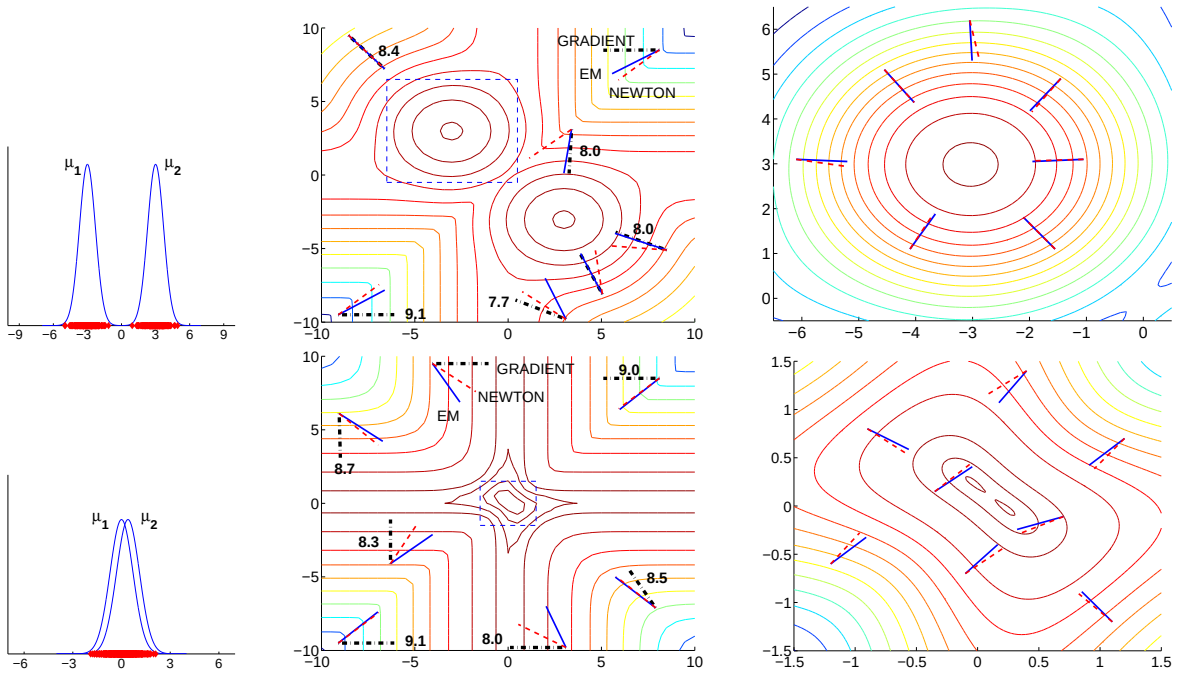

Figure 1. Contour plots of the likelihood function $L(\Theta)$ for MoG examples using well-separated (upper panels) and not-well-separated (lower panels) one-dimensional data sets. Axes correspond to the two means. The dashdot line shows the direction of the true gradient $\nabla_L(\Theta)$, the solid 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.

method, rescaling the gradient by exactly the negative inverse Hessian. As the eigenvalues tend to unity, EM takes smaller and smaller stepsizes, giving poor, first-order, convergence.

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

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

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):

$$P(\Theta^t) \approx \left[I - \left(\frac{\partial^2 H}{\partial \Theta^2}\right)\left(\frac{\partial^2 Q}{\partial \Theta^2}\right)^{-1}|_{\Theta=\Theta^t}\right]\left[-S(\Theta^t)\right]^{-1} \quad (6)$$

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 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 (6) approach zero and EM will exhibit extremely slow convergence.

Figure 1 illustrates the above results in the simple case of fitting a mixture of Gaussians model to well-clustered data – for which EM exhibits Quasi-Newton convergence – and not-well-clustered data, for which EM is slow. As we will see from the empirical results of the later sections, many other models also show this same effect. For example, when Hidden Markov Models or Aggregate Markov Models [11] are trained on very structured sequences, EM exhibits Quasi-Newton behavior, in particular when the state transition matrix is sparse and the output distributions are almost deterministic at each state.

The above analysis and experiments 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 analyze exactly such an alternative, the Expectation-Conjugate Gradient (ECG) algorithm, a simple direct optimization method for learning the parameters of latent variables models.

3. 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. (Note that if one can derive exact EM for a latent variable model, then one can always derive ECG by computing the above integral over hidden variables.) As an example, we describe a conjugate gradient algorithm:

Expectation-Conjugate-Gradient 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:** $\nabla_L(\Theta^t) = \int p(\boldsymbol{z}|\boldsymbol{x},\Theta^t)\frac{\partial}{\partial\Theta}\log p(\boldsymbol{x},\boldsymbol{z}|\Theta)d\boldsymbol{z}$

When certain parameters must obey positivity or linear constraints, we can either modify our optimizer to respect the constraints, or we can reparameterize to allow unconstrained optimization. In our experiments, we use simple reparameterizations of model parameters that allow our optimizers to work with arbitrary values. For example, in the MoG model we use a "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 (or log variances for diagonal covariance matrices). In HMMs, we reparameterize probabilities via softmax functions as well.

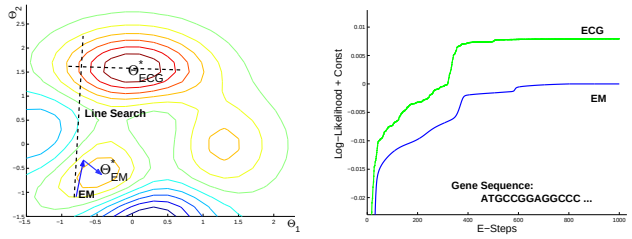

Figure 2. Left panel illustrates why ECG may converge to a better local optimum. Right panel displays the learning curves for EM and ECG algorithms of training fully connected 7-state HMM to model human DNA sequences. Both algorithms started from the same initial parameter values, but converged to the different local optimum.

Of course, the choice of initial conditions is very important for the EM algorithm or for ECG. Since EM is based on optimizing a convex lower bound on the likelihood, once EM is trapped in a poor basin of attraction, it can never find a better local optimum. Algorithms such as split and merge EM [13] were developed to escape from such configurations. It turns out that direct optimization methods such as ECG may also avoid this problem because of the nonlocal nature of the line search. In many of our experiments, ECG actually converges to a better local optimum than EM; figure 2 illustrates exactly such case.

4. Hybrid EM-ECG Algorithm

As we have seen, the relative performance of EM versus direct optimization depends on the missing information ratio for the given model and data set. The key to practical speedups is the ability to design a hybrid algorithm that can estimate the local missing information ratio $M'(\Theta^t)$ to detect whether to use EM or a direct approach such as ECG. Some authors have attacked this problem by finding the top eigenvector of $\frac{\partial M(\Theta)}{\partial\Theta}|_{\Theta=\Theta^t}$ as Θ^t approaches Θ^* using conventional numerical methods, such as finite-difference approximations, or power methods [4]. These approaches are computationally intensive and difficult to implement, and thus they have not been popular in practice.

Here, we propose using the *entropy of the posterior over hidden variables*, which can be computed after performing an E-step, as a crude estimate of the local missing information ratio. This entropy has a natural interpretation as the uncertainty about missing information, and thus can serve as a guide between switching regimes of EM and ECG. For many models with discrete hidden variables this quantity is quite easy to compute. In particular, we define the Normalized Entropy term:
$\bar{H}_t = \frac{-1}{N\ln M}\sum_n^N\sum_i^M p(\boldsymbol{z}=i|\boldsymbol{x}_n,\Theta^t)\ln p(\boldsymbol{z}=i|\boldsymbol{x}_n,\Theta^t)$
with $\boldsymbol{z}$ being discrete hidden variable taking on M values, and N observed data vectors $\boldsymbol{x}_n$. We now simply switch between EM and ECG based on thresholding this quantity:

Hybrid EM-ECG algorithm:

- Perform EM iterations, evaluating $\bar{H}_t$ after each E-step
- If $\bar{H}_t \geq \tau$ Then[a] Switch to ECG
- Perform ECG, evaluating $\bar{H}_t$ at the end of each line search
- If $\bar{H}_t < \tau$ Then Switch back to EM
- Exit at either phase **IF**:
  1. $(L(\Theta^t) - L(\Theta^{t-1}))/\text{abs}(L(\Theta^t)) < \text{tol}$ **OR**
  2. t > Tmax

[a] We are near the optimum or in plateau region with high entropy

As we will see from the experimental results, this simple hybrid EM-ECG algorithm performs no worse, and often far better than either EM or ECG.

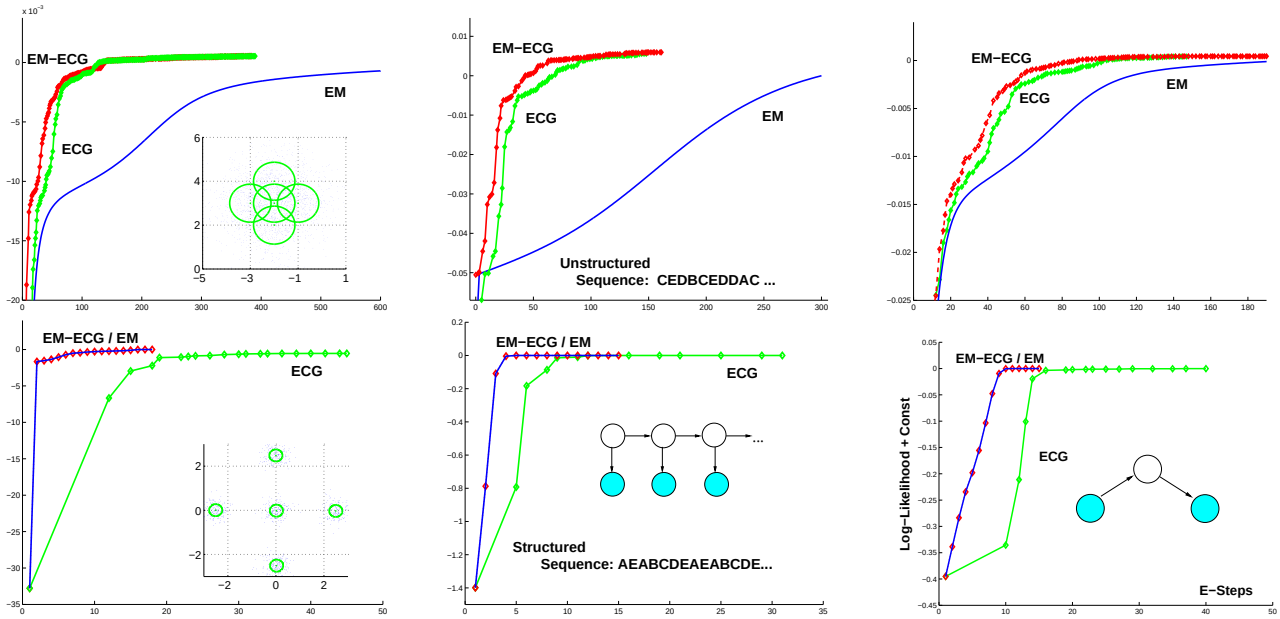

Figure 3. Learning curves for ECG, EM-ECG, and EM algorithms, showing superior (upper panels) and inferior (lower panels) performance of ECG under different conditions for three models: MoG (left), HMM (middle), and Aggregate Markov Models (right). The number of E-steps taken by either algorithm is shown on the horizontal axis, and log likelihood is shown on the vertical axis. For ECG and EM-ECG, diamonds indicate the maximum of each line search, and the zero-level likelihood corresponds to the converging point of the EM algorithm. The bottom panels use "well-separated", or "structured" data for which EM possesses Quasi-Newton convergence behavior. All models in this case converge in 10-15 iterations with stopping criterion: $[L(\Theta^{t+1}) - L(\Theta^t)]/\text{abs}(L(\Theta^{t+1})) < 10^{-15}$. The upper panels use "overlapping", "aliased", or "unstructured" data for which proposed algorithms performs much better.

5. Experimental Results

We now present empirical results comparing the performance of EM, ECG, and hybrid EM-ECG for learning the parameters of three latent variable models: Mixtures of Gaussians (MoG), Hidden Markov Models (HMM), and Aggregate Markov Models. In many latent variable models, performing inference (E-step) is significantly more expensive compared to either the parameter updates (M-step) or the line search overhead in the CG step of ECG. To compare the performance of the algorithms, we therefore simply compare the number of E-steps each algorithm executes until its convergence. We first show results on synthetic data sets, whose properties we can control to verify certain aspects of our theoretical analysis. We also report empirical results on several real world data sets, showing that our algorithms do work well in practice. Though we show examples of single runs, we have confirmed that the convergence results presented in all our experiments do not vary significantly for different initial parameter conditions. For all of the reported experiments, we used tol $= 10^{-8}$ and $\tau = 0.5$.

5.1. Synthetic Data Sets

First, consider a mixture of Gaussians (MoG) model. We considered two types of data sets, 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 3 shows that ECG and Hybrid EM-ECG outperform standard EM in the poorly separated cases. For the well-separated case, the hybrid EM-ECG algorithm never switches to ECG due to the small normalized entropy term, and EM converges very quickly. This is predicted by our analysis: 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 that EM will have Quasi-Newton convergence behavior.

We then applied our algorithms to the training of Hidden Markov Models (HMMs). Missing information in this model is high when the observed data do not well determine the underlying state sequence (given the parameters). We therefore generated two data sets from a 5-state HMM, with an alphabet size of 5 characters. The first data set ("aliased" sequences) was generated from a HMM where output parameters were set to uniform values plus some small noise. The second data set ("very structured sequences") was generated from a HMM with sparse transition and output matrices. For the ambiguous or aliased data, ECG and hybrid EM-ECG outperform EM substantially. For the very structured data, EM performs well and exhibits second

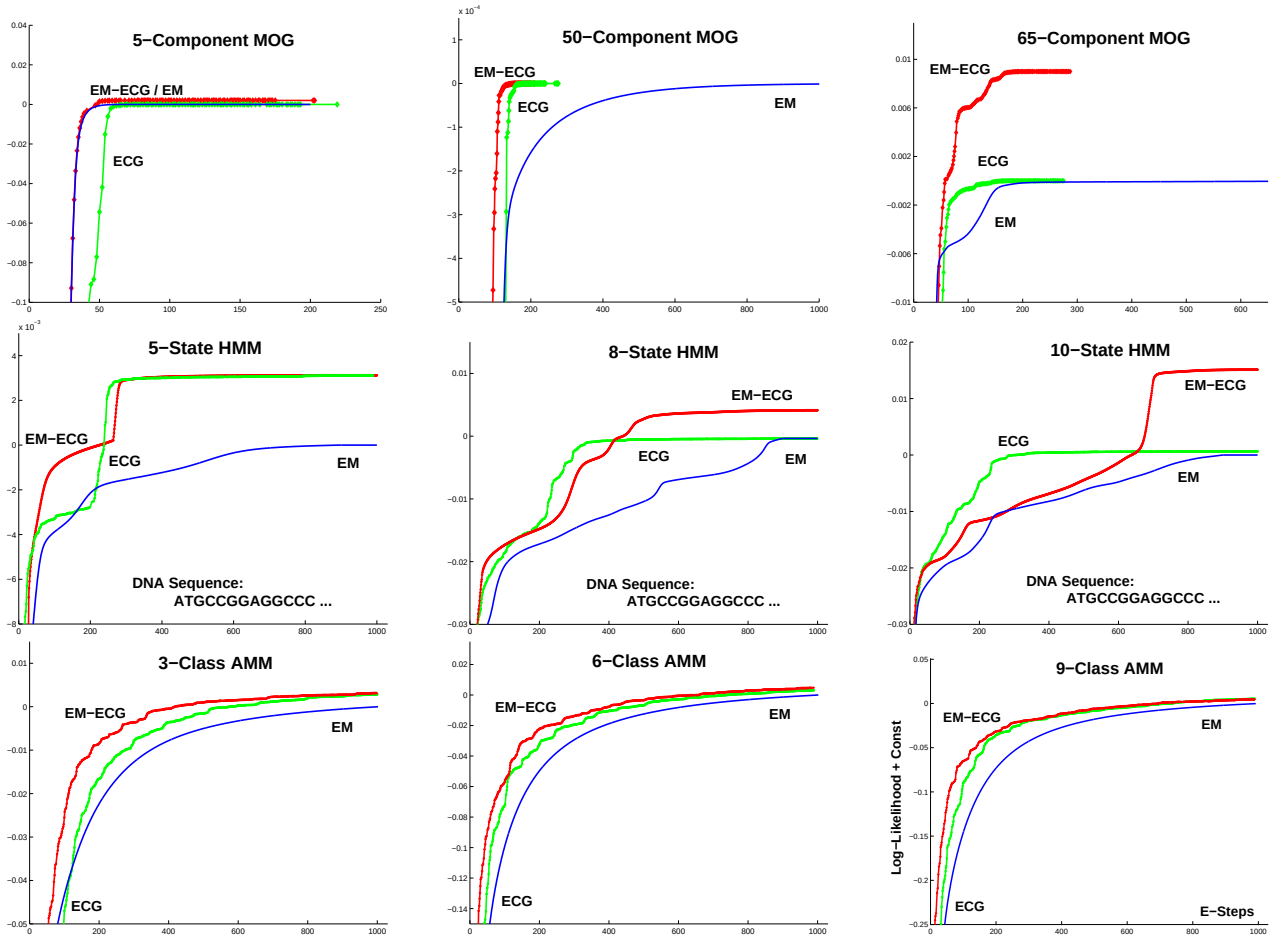

Figure 4. Learning curves for ECG, EM-ECG, and EM algorithms, displaying convergence performance under different conditions for three models: MoG (upper), HMM (middle), and Aggregate Markov Models (bottom). The number of E-steps taken by either algorithm is shown on the horizontal axis, and log likelihood is shown on the vertical axis. For ECG and EM-ECG, diamonds indicate the maximum of each line search, and the zero-level likelihood corresponds to the converging point of the EM algorithm. The number of learned clusters for MoG model were 5 (left), 50 (middle), and 65 (right). For HMM model, the number of states were 5 (left), 8 (middle), and 10 (right). The number of learned themes for the AMM model were 3 (left), 6 (middle), and 9 (right).

order convergence in the vicinity of the local optimum.

Finally, we experimented with Aggregate Markov Models (AMMs) [11]. AMMs model define a discrete conditional probability table $p_{ij} = p(y = j|x = i)$ using a low rank approximation. In the context of n-gram models for word sequences, AMMs are class-based bigram models in which the mapping from words to classes is probabilistic. In particular, the class-based bigram model predicts that word w_1 is followed by word w_2 with probability: $P(w_2|w_1) = \sum_{c=1}^{C} P(w_2|c)P(c|w_1)$ with C being the total number of classes. Here, the concept of missing information corresponds to how well or poor a set of words determine the class labels C based on the observation words that follow them. The right panels of figure 3 show training of a 2-class 50-state AMM model on ambiguous (aliased) data, in which words do not well determine class labels, and on more structured data, in which the proportion of missing information is very small. ECG and hybrid EM-ECG are superior to EM by at least a factor of two for ambiguous data; for structured data EM shows the expected Quasi-Newton convergence behavior.

5.2. Real World Data Sets

In our first experiment, we cluster a set of 50,000 8×8 greyscale pixel image patches[4] using a mixture of Gaussians model. The patches were extracted from 768×512 natural images, described in [14] (see fig 5 for an example of a natural image, and sample patches). To speed-up the experiments, the patch data was projected with PCA down to a 10-dimensional linear sub-

[4]The data set used was the **imlog** data set publicly available at ftp://hlab.phys.rug.nl/pub/samples/imlog

space and the mixing proportions and covariances of the model were held fixed. The means were initialized by performing K-means. We experimented with mixtures having M=2 up to M=65 clusters.

Figure 5. An example of a natural image and some samples of 8×8 grey pixel image patches, used in the clustering experiment.

Figure 4 displays the convergence of EM, ECG, and Hybrid EM-EC algorithms for M=5, M=50 and M=65. The experimental results show that with fewer mixture components EM outperforms ECG, since the components generally model the data with fairly distinct, non-contiguous clusters. As the number of mixtures components increases, clusters overlap in contiguous regions and the normalized entropy term grows, suggesting a relatively high proportion of the missing information. In this case ECG outperform EM by several orders of magnitude. Hybrid EM-ECG algorithm is never inferior to either EM or ECG (using our untuned setting of switching threshold $\tau = 0.5$).

Our second experiment consisted of training a fully connected HMM to model DNA sequences. For the training, we used publicly available "GENIE gene finding data set", provided by UCSC and LBNL [5], that contains 793 unrelated human genomic DNA sequences. We applied our different algorithms on 66 DNA sequences with length varying anywhere between 200 to 3000 genes per sequence. The number of states ranged from M=5 to M=10 and all the parameter values were randomly initialized. Figure 4 shows the convergence of EM, ECG, and Hybrid EM-ECG algorithms for M=5,8,10. This data set contains very complex structure which is not easily modeled by HMMs, resulting in a very high proportion of missing information. As a result, hybrid EM-ECG and ECG substantially outperform EM in terms of convergence.

In our last experiment, we applied Aggregate Markov Models to the data set consisting of 2,037 NIPS authors and corresponding counts of the top 1,000 most frequently used words of the NIPS conference proceedings, volumes 1 to 12.[5] The goal was to model the probability that an author A will use word W using a small number of "soft" classes (t): $P(W|A) = \sum_{t=1}^{T} P(W|t)P(t|A)$. Once again, we observe that for this simple model, this data set has a large fraction of missing information. Figure 4 displays the convergence of EM, ECG, and EM-ECG algorithms for T=3,6,9. with hybrid EM-ECG and ECG having superior convergence over EM.

[5]NIPS corpus used in the experiments is publicly available at http://www.cs.toronto.edu/∼roweis/data.html

6. Discussion and Conclusions

Although we have focused here on discrete latent variables, the ECG and hybrid algorithms can also be derived for latent variable models with continuous hidden variables. As an example figure 6 illustrates convergence behaviour of the Probabilistic Principal Component Analysis (PPCA) latent variable model[9, 12], 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 EM performs well; for nearly circular data ECG converges faster.[6]

In some degenerate cases, where the proportion of missing information is very high, i.e. $M'(\Theta^*)$ approaches identity, EM convergence can be exponentially slow. Figure 6 illustrates such example for the case of HMM training using almost random sequences. It takes about 7,000 iterations for ECG and EM-ECG to converge to the ML estimate, whereas even after 250,000 iterations EM is still only approaching the local optimum.

In this paper we have presented comparative analysis of EM and direct optimization algorithms for latent variable models, and developed a theoretical connection between these two approaches. We have also analyzed and determined conditions under which EM algorithm can demonstrate local-gradient and Quasi-Newton convergence behaviors. Our results extend those of Xu and Jordan[15] who analyzed the convergence properties of the EM algorithm in the special case of Gaussian mixtures, and apply to any exponential family model.

Motivated by these analyses, we have proposed an alternative hybrid optimization method that can significantly outperform EM in many cases and is almost never inferior. We tested the proposed algorithms by

[6]The slow convergence of EM in PPCA is also true for factor analysis 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.

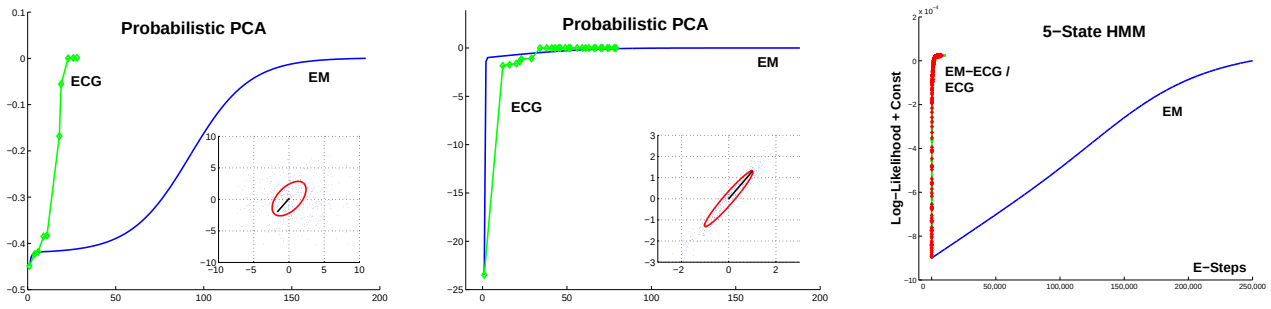

Figure 6. Learning curves for ECG (dots) and EM (solid lines) algorithms, showing superior (left) and inferior (middle) performance of ECG. The left panel uses "ill-conditioned" data for which ECG converges quickly; the middle panel uses "low-rank" data for which EM performs better. Right panel displays "non-converging" case of the EM. Very unstructured data (30 sequences, each of length 50) was generated from a full 5-state HMM with alphabet size of 5. Parameter values were set to be uniform plus some small uniform noise. ECG and EM-ECG converge in about 7,000 iterations, whereas after even 250,000 iterations, EM is only approaching to the ML estimate.

training several basic latent variable models on several synthetic as well as real world data sets, reporting convergence behavior and explaining the results with reference to our analysis.

Our convergence analysis can also be extended to a broader class of bound optimization techniques, such as 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[16]. These analyses allow us to gain a deeper understanding of the nature of these algorithms and the conditions under which certain optimization techniques can be expected to outperform others. Based on these extended analyses we are designing accelerated fitting algorithms for these models as well.

Acknowledgments

We would like to thank Yoshua Bengio and Drew Bagnell 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. *J. of the RS Society series B*, 39:1–38, 1977.

[4] Chris Fraley. On computing the largest fraction of missing information for the EM algorithm and the worst linear function for data augmentation. Technical report, University of Washington.

[5] GENIE gene data set. LBNL and UC Santa Cruz, `http://www.fruitfly.org/sequence`.

[6] Mortaza Jamshidian and Robert I. Jennrich. Acceleration of the EM algorithm by using quasi-newton methods. *J. of the RS Society series B*, 49, 1997.

[7] Meng X. L. and van Dyk D. Fast EM-type implementations for mixed effects models. *J. of the Royal Statistical Society series B*, 60:559–578, 1998.

[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.

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

[11] Lawrence Saul and Fernando Pereira. Aggregate and mixed-order Markov models for statistical language processing. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 81–89. 1997.

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

[13] Naonori Ueda, Ryohei Nakano, Zoubin Ghabramani, and Geoffrey E. Hinton. SMEM algorithm for mixture models. *Neural Computation*, 12(9):2109–2128, 2000.

[14] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. In *Proceedings of the Royal Society of London*, pages 359–366, 1998.

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

[16] Alan Yuille and Anand Rangarajan. The convex-concave computational procedure (CCCP). In *Advances in NIPS*, volume 13, 2001.