Convergence Rate of Expectation-Maximization

Raunak Kumar raunak.kumar17@outlook.com
University of British Columbia
Mark Schmidt schmidtm@cs.ubc.ca
University of British Columbia

Abstract

Expectation-maximization (EM) is an iterative algorithm for finding the maximum likelihood or maximum a posteriori estimate of the parameters of a statistical model with latent variables or when we have missing data. In this work, we view EM in a generalized surrogate optimization framework and analyze its convergence rate under commonly-used assumptions. We show a lower bound on the decrease in the objective function value on each iteration, and use it to provide the first convergence rate for non-convex functions in the generalized surrogate optimization framework and, consequently, for the EM algorithm. We also discuss how to improve EM by using ideas from optimization.

1 Introduction

Expectation-maximization (EM) [2] is one of the most cited statistics papers of all time and is a popular tool in machine learning. It is widely used to fit datasets with missing values or to fit models with latent variables. Some of the cannonical applications include mixture models, hidden Markov models, semi-supervised learning, and fitting generative models when there are missing values in the data. EM only converges to a stationary point, which may not be a local maxima [6], but it tends to perform well in practice. However, not much is known about its convergence rate.

The original EM paper by Dempster et al. [2] contained an error in their proof of the convergence of EM, which was subsequently fixed by Wu [6] who showed that it is guaranteed to converge to a stationary point under suitable continuity assumptions. Wu [6] and Figueiredo et al. [7] also discuss convergence to a local or global maxima under stronger assumptions. However, they do not provide convergence rates. Salakhutdinov et al. [5] adopt an optimization perspective similar to ours. They show that if the ratio of missing information to observed data is small, then in a neighbourhood of an optima, EM typically displays superlinear convergence. Balakrishanan et al. [4] analyze EM in the limit of infinite data and in the case of a finite set of samples. If the initial estimate λ^0 is in a neighbourhood of an optima, and the gradient of the Q-function is bounded in this neighbourhood, then with infinte data they show that the sequence of iterates produced by EM converges linearly. For the case with finite samples, they provide similar results with some additional assumptions. The assumptions in these previous works are quite strong, and in this work we use milder assumptions and a simpler argument. Indeed, most of our assumptions are standard in optimization literature. For example, the first of our assumptions is that the function EM is minimizing is bounded below. The other assumption is that the surrogate functions (Section 2) are strongly-convex. Alternatively, we could assume that the surrogates have a Lipschitz-continuous gradient, and that the gradient of the surrogate and the function agree. The strong-convexity assumption leads to a convergence rate in terms of the iterates, while the latter leads to a rate in terms of the gradient norm. We discuss the assumptions used in this work and the previous work in the next section.

Our main contribution is that we provide the first convergence rate for non-convex functions in a generalized version of Mairal's surrogate optimization framework [1] and, consequently, a non-asymptotic convergence rate for many common variations of the EM algorithm. The rest of the paper is organized as follows. First, we generalize the definition of first-order surrogate functions [1] and show that EM is a surrogate optimization algorithm in this framework. Then, we proceed to show a lower bound on the progress made by EM on each iteration, and use this bound to derive a $O(\frac{1}{t})$ convergence rate in terms of the squared difference between successive iterates produced by the EM algorithm. We also propose a similar convergence rate in terms of the norm of the gradient of the objective under different assumptions. Finally, we propose some future research directions that can utilize ideas from optimization to improve EM.

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

2 Surrogate optimization

Mairal [1] defines first-order surrogate functions and presents a surrogate optimization framework for solving the following problem: suppose $\Lambda \subset \mathbb{R}^d$ is convex, and $f : \mathbb{R}^d \to \mathbb{R}$ is continuous and bounded below; solve for

$$\lambda^* \in \arg\min_{\lambda \in \Lambda} f(\lambda).$$

In this section, we generalize Mairal's definition of first-order surrogate functions and view EM in this generalized framework.

Definition (Surrogate functions)**.** Let f and g be functions from $\mathbb{R}^d \to \mathbb{R}$. We say that g is a surrogate of f near $\lambda^k \in \Lambda$ if it satisfies:

- **Majorization**: $\forall \lambda' \in \arg\min_{\lambda \in \Lambda} g(\lambda),\ f(\lambda') \leq g(\lambda')$. If $f(\lambda) \leq g(\lambda)$ for all $\lambda \in \Lambda$, then g is called a majorant function;
- **Smoothness**: Denote the approximation error as $h = g - f$. Then, the functions agree at λ^k so that $h(\lambda^k) = 0$.

We will use $S_\rho(f, \lambda^k)$ to denote the set of such surrogates that are ρ-strongly-convex. In this setting Mairal defines the following surrogate optimization framework:

Algorithm 1 Mairal's Surrogate Optimization Scheme

Input: $\lambda^0 \in \Lambda$, number of iterations t.
for k = 1 to t **do**
  Compute a surrogate function g_k of f near λ^{k-1}.
  Update solution $\lambda^k \in \arg\min_{\lambda \in \Lambda} g_k(\lambda)$.
end for
Output final estimate λ^t.

However, in contrast to the definition of surrogate functions in Mairal [1], we do not require h to be differentiable and $\nabla h(\lambda^k) = 0$. Thus, we are only requiring that the surrogate be a zero-order surrogate function rather than a first-order surrogate.

2.1 EM as a surrogate optimization algorithm

In EM we want to find parameters $\lambda \in \Lambda$ to maximize the likelihood, $P(X|\lambda) = \sum_z P(X, z|\lambda)$, of data X where we have written the likelihood in terms of missing data or latent variables z. We can equivalently minimize the negative log-likelihood (NLL), so our goal is to find

$$\lambda^* \in \arg\min_{\lambda \in \Lambda} -\log \sum_z P(X, z|\lambda).$$

Let λ^k denote the estimate of the parameters after the k^{th} iteration and define

$$Q(\lambda|\lambda^k) = \sum_z P(z|X, \lambda^k) \log P(X, z|\lambda).$$

Using Jensen's inequality, we get the following well-known upper bound on the NLL

$$-\log P(X|\lambda) \leq -Q(\lambda|\lambda^k) - \text{entropy}(z|X, \lambda^k), \tag{1}$$

and the iterations of EM are defined as

$$\begin{aligned} \lambda^{k+1} &\in \arg\min_{\lambda \in \Lambda} -Q(\lambda|\lambda^k) - \text{entropy}(z|X, \lambda^k) \\ \equiv \lambda^{k+1} &\in \arg\min_{\lambda \in \Lambda} -Q(\lambda|\lambda^k). \end{aligned}$$

We'll define

$$\begin{aligned} f(\lambda) &= -\log P(X|\lambda) = -\log \sum_z P(X, z|\lambda), \\ g_k(\lambda) &= -Q(\lambda|\lambda^{k-1}) - \text{entropy}(z|X, \lambda^{k-1}). \end{aligned}$$

To view EM as a surrogate optimization algorithm (Algorithm 1) in our generalized framework, we need to verify that g_k as defined above is indeed a surrogate of f. From Equation (1), we can see that g_k is a majorant of f, and thus it satisfies the majorization condition. It is a well known fact (Dempster et al. [2]) that $h_k(\lambda^{k-1}) = 0$, where $h_k = g_k - f$ is the approximation error. In addition, to derive our convergence results, we will also assume that for all iterations, g_k is ρ-strongly-convex.

Previous works that study the convergence rate of EM give fast rates but they make very strong assumptions [4], [5]. For example, they require the initial estimate of the parameters λ^0 to be within some small neighbourhood of the optimal solution. Additionally, they require the fraction of missing information to be small [5] or other regularity conditions [4].

In contrast, we make relatively mild assumptions. For example, we first simply assume that the NLL f is bounded below which is satisfied by most real world datasets (particularly if we include a proper prior on the parameters). Restricting the iterates λ^k to stay within a convex set is essentially a non-assumption, since this set could simply be all of $\mathbb{R}^d$. Beyond that f is bounded below, the only strong assumption that we make is that the surrogates g_k are strongly-convex. But this is satisfied in many important applications. For example, when the complete-data NLL ($-\log P(X, z|\lambda)$) is convex and we use a strongly-convex regularizer, then the surrogate is strongly-convex (even though the objective f itself is non-convex). Even without a regularizer, in the common case of EM for mixtures of exponential families the surrogate will be strongly-convex if the mixture probabilities are positive and the covariances of the distributions are positive-definite (both of these are automatically achieved under standard choices of the prior).

3 Results

We now lower bound the decrease in the objective function value on each iteration of EM. Informally, if the iterates of EM stay within a convex set and the surrogates are ρ-strongly-convex, then the further away successive iterates λ^{k-1} and λ^k are, the greater the decrease in the objective function value. Then, we use this bound to derive an $O(\frac{1}{t})$ convergence rate in terms of the squared difference between the successive iterates. We present these results formally in two theorems, the first of which is based on Mairal's Lemma 2.1 [1] and the second of which is based on Khan et al.'s Proposition 1 [3].

Theorem 1 (Lower bound)**.** *Let $g_k \in S_\rho(f, \lambda^{k-1})$, and $\lambda^k \in \arg\min_{\lambda\in\Lambda} g_k(\lambda)$. Then,*

$$f(\lambda^k) \leq f(\lambda^{k-1}) - \frac{\rho}{2}\|\lambda^k - \lambda^{k-1}\|_2^2.$$

Proof. Using that λ^k minimizes g_k and that g_k is ρ-strongly-convex, it follows that for all $\lambda \in \Lambda$,

$$g_k(\lambda^k) + \frac{\rho}{2}\|\lambda - \lambda^k\|_2^2 \leq g_k(\lambda).$$

Now using that g_k is a majorant, we get

$$\begin{aligned} f(\lambda^k) + \frac{\rho}{2}\|\lambda^k - \lambda\|_2^2 &\leq g_k(\lambda^k) + \frac{\rho}{2}\|\lambda^k - \lambda\|_2^2 \\ &\leq g_k(\lambda) \\ &= f(\lambda) + h_k(\lambda). \end{aligned}$$

Setting $\lambda = \lambda^{k-1}$ and using that $h_k(\lambda^{k-1}) = 0$ from the definition of surrogate functions gives

$$\begin{aligned} f(\lambda^k) + \frac{\rho}{2}\|\lambda^k - \lambda^{k-1}\|_2^2 &\leq f(\lambda^{k-1}) + h_k(\lambda^{k-1}) \\ \frac{\rho}{2}\|\lambda^k - \lambda^{k-1}\|_2^2 &\leq f(\lambda^{k-1}) - f(\lambda^k), \end{aligned} \tag{2}$$

which can be re-arranged to get the result. □

Theorem 2 (Convergence rate)**.** *Let $g_k \in S_\rho(f, \lambda^{k-1})$, and $\lambda^k \in \arg\min_{\lambda\in\Lambda} g_k(\lambda)$. Then,*

$$\min_{k\in\{1,2,\ldots,t\}} \|\lambda^k - \lambda^{k-1}\|_2^2 \leq \frac{2(f(\lambda^0) - f(\lambda^*))}{\rho t}.$$

Proof. Summing up (2) for all k and telescoping the sum we get

$$\begin{aligned}\sum_{k=1}^{t} \frac{\rho}{2}\|\lambda^k - \lambda^{k-1}\|_2^2 &\leq \sum_{k=1}^{t} f(\lambda^{k-1}) - f(\lambda^k) \\ &= f(\lambda^0) - f(\lambda^t) \\ &\leq f(\lambda^0) - f(\lambda^*).\end{aligned}$$

Taking the $\min$ over all iterations, we get

$$\min_{k\in 1,2,\ldots,t} \|\lambda^k - \lambda^{k-1}\|_2^2 \cdot \frac{\rho t}{2} \leq f(\lambda^0) - f(\lambda^*)$$

$$\min_{k\in\{1,2,\ldots,t\}} \|\lambda^k - \lambda^{k-1}\|_2^2 \leq \frac{2(f(\lambda^0) - f(\lambda^*))}{\rho t}.$$

□

Due to the non-convexity of f, the above rate does not necessarily hold for the last iteration. However, it holds for the average or the minimum value of $\|\lambda^k - \lambda^{k-1}\|_2^2$. Additionally, the above results do *not* rely on the differentiability of the original function or its surrogates.

4 Discussion

Although our analysis is quite general and relies on relatively mild assumptions, it would be interesting to see if some assumptions, like strong-convexity of the surrogates, can be relaxed, or if we can derive stronger convergence results using the same set of assumptions for "nice" scenarios like mixtures of exponential family distributions. Our convergence rate is in terms of the squared difference between the iterates. If we make a different assumption that g_k is a first-order surrogate function [1] so that the approximation error h_k is differentiable and ∇h_k is L-Lipschitz continuous, and that the gradients agree, ie. $\nabla h_k(\lambda^{k-1}) = 0$, then we can derive a similar convergence rate in terms of the norm of the gradient of f. The differentiability will typically follow from using an NLL f that is differentiable on its domain. The assumption that ∇h_k is L-Lipschitz continuous is weak provided that our iterates λ^k are not diverging to an infinite value, or converging to a point on the boundary of the domain. In common cases of mixtures of exponential families, it is true that $\nabla h_k(\lambda^{k-1}) = 0$. The convergence proof under these assumptions essentially relies on the observation that since λ^k is a global minimizer of g_k, $g_k(\lambda^k) \leq g_k(\lambda^{k-1} - \frac{1}{L}\nabla g(\lambda^{k-1}))$. Using the standard gradient descent progress bound (Theorem 1, Karimi et al. [8]), and that $h_k(\lambda^{k-1}) = 0$ and $\nabla h_k(\lambda^{k-1}) = 0$, we can follow the proofs of the above theorems and arrive at the result that

$$\min_{k\in\{1,2,\ldots,t\}} \|\nabla f(\lambda^{k-1})\|_2^2 \leq \frac{2L(f(\lambda^0) - f(\lambda^*))}{t}.$$

Similar to [5] and [6], we view EM in an optimization framework. Doing so allows future work to use numerical optimization techniques to develop improved variants of EM. In particular, accelerated EM and an SVRG version of EM could be worth exploring with these tools.

References

[1] Mairal, J., 2013. Optimization with first-order surrogate functions. In Proceedings of the 30th International Conference on Machine Learning (ICML-13) (pp. 783-791).

[2] Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), pp.1-38.

[3] Khan, M.E., Babanezhad, R., Lin, W., Schmidt, M. and Sugiyama, M., 2015. Faster stochastic variational inference using proximal-gradient methods with general divergence functions. arXiv preprint arXiv:1511.00146.

[4] Balakrishnan, S., Wainwright, M.J. and Yu, B., 2017. Statistical guarantees for the EM algorithm: From population to sample-based analysis. The Annals of Statistics, 45(1), pp.77-120.

[5] Salakhutdinov, R., Roweis, S. and Ghahramani, Z., 2002, August. On the convergence of bound optimization algorithms. In Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence (pp. 509-516). Morgan Kaufmann Publishers Inc..

[6] Wu, C.J., 1983. On the convergence properties of the EM algorithm. The Annals of statistics, pp.95-103.

[7] Figueiredo, M.A. and Nowak, R.D., 2003. An EM algorithm for wavelet-based image restoration. IEEE Transactions on Image Processing, 12(8), pp.906-916.

[8] Karimi, H., Nutini, J. and Schmidt, M., 2016, September. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 795-811). Springer International Publishing.