Fast Bi-Level and Incremental Noisy EM Algorithms

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

T.B.C

1 Introduction

We formulate the following empirical risk minimization as:

$$\min_{\boldsymbol{\theta} \in \Theta} \overline{\mathcal{L}}(\boldsymbol{\theta}) := \mathrm{R}(\boldsymbol{\theta}) + \mathcal{L}(\boldsymbol{\theta}) \;\text{ with }\; \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \big\{ -\log g(y_i;\boldsymbol{\theta}) \big\} \,, \tag{1}$$

where $\{y_i\}_{i=1}^n$ are the observations, Θ is a convex subset of $\mathbb{R}^d$ for the parameters, $\mathrm{R} : \Theta \to \mathbb{R}$ is a smooth convex regularization function and for each $\boldsymbol{\theta} \in \Theta$, $g(y;\boldsymbol{\theta})$ is the (incomplete) likelihood of each individual observation. The objective function $\overline{\mathcal{L}}(\boldsymbol{\theta})$ is possibly *non-convex* and is assumed to be lower bounded $\overline{\mathcal{L}}(\boldsymbol{\theta}) > -\infty$ for all $\boldsymbol{\theta} \in \Theta$. In the latent variable model, $g(y_i;\boldsymbol{\theta})$, is the marginal of the complete data likelihood defined as $f(z_i, y_i;\boldsymbol{\theta})$, i.e. $g(y_i;\boldsymbol{\theta}) = \int_{\mathsf{Z}} f(z_i, y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i)$, where $\{z_i\}_{i=1}^n$ are the (unobserved) latent variables. We make the assumption of a complete model belonging to the curved exponential family, *i.e.,*

$$f(z_i, y_i;\boldsymbol{\theta}) = h(z_i, y_i) \exp\big(\big\langle S(z_i, y_i) \,|\, \phi(\boldsymbol{\theta})\big\rangle - \psi(\boldsymbol{\theta})\big) \,, \tag{2}$$

where $\psi(\boldsymbol{\theta})$, $h(z_i, y_i)$ are scalar functions, $\phi(\boldsymbol{\theta}) \in \mathbb{R}^k$ is a vector function, and $S(z_i, y_i) \in \mathbb{R}^k$ is the complete data sufficient statistics.

Prior Work Cite Kuhn [Kuhn et al., 2019] (for ISAEM) and incremental EM like papers. As well as Optim papers (Variance reduction, SAGA etc.)

2 Expectation Maximization Algorithm

Full batch EM is a two steps procedure. The $\textsf{E-step}$ amounts to computing the conditional expectation of the complete data sufficient statistics,

$$\overline{\mathbf{s}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \overline{\mathbf{s}}_i(\boldsymbol{\theta}) \quad \text{where} \quad \overline{\mathbf{s}}_i(\boldsymbol{\theta}) = \int_{\mathsf{Z}} S(z_i, y_i) p(z_i|y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i) \,. \tag{3}$$

The $\textsf{M-step}$ is given by

$$\textsf{M-step:}\;\; \hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}(\overline{\mathbf{s}}(\boldsymbol{\theta})) := \underset{\vartheta \in \Theta}{\arg\min} \;\big\{ \mathrm{R}(\vartheta) + \psi(\vartheta) - \big\langle \overline{\mathbf{s}}(\boldsymbol{\theta}) \,|\, \phi(\vartheta) \big\rangle \big\}, \tag{4}$$

3 Monte Carlo Integration and Stochastic Approximation

For complex and possibly nonlinear models, the expectation under the posterior distribution defined in (3) is not tractable. In that case, the first solution involves computing a Monte Carlo integration of that latter term. For all $i \in \llbracket 1, n \rrbracket$, draw for $m \in \llbracket 1, M \rrbracket$, samples $z_{i,m} \sim p(z_i|y_i;\theta)$ and compute the MC integration $\hat{\mathrm{s}}$ of the deterministic quantity $\overline{\mathrm{s}}(\boldsymbol{\theta})$:

$$\textsf{MC-step}: \ \hat{\mathrm{s}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=1}^{M} S(z_{i,m}, y_i)$$

and compute $\hat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}(\hat{\mathrm{s}})$.

This algorithm bypasses the intractable expectation issue but is rather computationally expensive in order to reach point wise convergence (M needs to be large).

As a result, an alternative to that stochastic algorithm is to use a Robbins-Monro (RM) type of update. We denote

$$\hat{S}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\hat{S}_i^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=1}^{M} S(z_{i,m}^{(k)}, y_i) \tag{5}$$

where $z_{i,m}^{(k)} \sim p(z_i|y_i;\theta^{(k-1)})$. At iteration k, the sufficient statistics $\hat{\mathrm{s}}^{(k)}$ is approximated as follows:

$$\textsf{SA-step}: \ \hat{\mathrm{s}}^{(k)} = \hat{\mathrm{s}}^{(k-1)} + \gamma_k(\hat{S}^{(k)} - \hat{\mathrm{s}}^{(k-1)}) \tag{6}$$

where $\{\gamma_k\}_{k=1}^{\infty} \in [0,1]$ is a sequence of decreasing step sizes to ensure asymptotic convergence. This is called the Stochastic Approximation of the EM (SAEM), see [Delyon et al., 1999] and allows a smooth convergence to the target parameter. It represents the *first level* of our algorithm (needed to temper the variance and noise implied by MC integration).

In the next section, we derive variants of this algorithm to adapt of the sheer size of data of today's applications.

4 Incremental and Bi-Level Inexact EM Methods

Strategies to scale to large datasets include classical incremental and variance reduced variants. We will explicit a general update that will cover those variants and that represents the *second level* of our algorithm, namely the incremental update of the noisy statistics $\hat{S}^{(k)}$ inside the RM type of update.

$$\textsf{Inexact-step}: \ \hat{S}^{(k)} = \hat{S}^{(k-1)} + \rho_{k+1}\big(\boldsymbol{\mathcal{S}}^{(k)} - \hat{S}^{(k-1)}\big), \tag{7}$$

Note $\{\rho_k\}_{k=1}^{\infty} \in [0,1]$ is a sequence of step sizes, $\boldsymbol{\mathcal{S}}^{(k+1)}$ is a proxy for $\hat{S}^{(k)}$, If the stepsize is equal to one and the proxy $\boldsymbol{\mathcal{S}}^{(k+1)} = \hat{S}^{(k)}$, i.e., computed in a full batch manner as in (5), then we recover the SAEM algorithm. Also if $\rho_k = 1$, $\gamma_k = 1$ and $\boldsymbol{\mathcal{S}}^{(k+1)=\hat{S}^{(k)}}$, then we recover the Monte Carlo EM algorithm.

We now introduce three variants of the SAEM update depending on different definitions of the proxy $\boldsymbol{\mathcal{S}}^{(k)}$ and the choice of the stepsize ρ_k. Let $i_k \in \llbracket 1, n \rrbracket$ be a random index drawn at iteration k and $\tau_i^k = \max\{k' : i_{k'} = i, \ k' < k\}$ be the iteration index where $i \in \llbracket 1, n \rrbracket$ is last drawn prior to iteration k. For iteration $k \geq 0$, the fiSAEM method draws *two* indices *independently* and uniformly as $i_k, j_k \in \llbracket 1, n \rrbracket$. In addition to τ_i^k which was defined *w.r.t.* i_k, we define $t_j^k = \{k' : j_{k'} = j, k' < k\}$ to be the iteration index where the sample $j \in \llbracket 1, n \rrbracket$ is last drawn as j_k prior to iteration k. With the initialization $\overline{\boldsymbol{\mathcal{S}}}^{(0)} = \overline{\mathrm{s}}^{(0)}$, we use a slightly different update rule from SAGA inspired by [Reddi et al., 2016]. Then, we obtain:

$$\begin{aligned}
&\textit{(iSAEM [Kuhn et al., 2019, Karimi, 2019])} && \boldsymbol{\mathcal{S}}^{(k+1)} = \boldsymbol{\mathcal{S}}^{(k)} + \tfrac{1}{n}\big(\hat{S}_{i_k}^{(k)} - \hat{S}_{i_k}^{(\tau_{i_k}^k)}\big) && (8)\\
&\textit{(vrSAEM This paper)} && \boldsymbol{\mathcal{S}}^{(k+1)} = \hat{S}^{(\ell(k))} + \big(\hat{S}_{i_k}^{(k)} - \hat{S}_{i_k}^{(\ell(k))}\big) && (9)\\
&\textit{(fiSAEM This paper)} && \boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + \big(\hat{S}_{i_k}^{(k)} - \hat{S}_{i_k}^{(t_{i_k}^k)}\big) && (10)\\
& && \overline{\boldsymbol{\mathcal{S}}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}\big(\hat{S}_{j_k}^{(k)} - \hat{S}_{j_k}^{(t_{j_k}^k)}\big). && (11)
\end{aligned}$$

The stepsize is set to $\rho_{k+1} = 1$ for the iSAEM method; $\rho_{k+1} = \gamma$ is constant for the vrSAEM and fiSAEM methods. Moreover, for iSAEM we initialize with $\boldsymbol{\mathcal{S}}^{(0)} = \hat{S}^{(0)}$; for vrSAEM we set an epoch size of m and define $\ell(k) := m\lfloor k/m \rfloor$ as the first iteration number in the epoch that iteration k is in.

Algorithm 1 Bi-Level Stochastic Approximation EM methods.

Input: initializations $\hat{\boldsymbol{\theta}}^{(0)} \leftarrow 0$, $\hat{\mathbf{s}}^{(0)} \leftarrow \hat{S}^{(0)}$, $K_{\sf max} \leftarrow$ max. iteration number.
Set the terminating iteration number, $K \in \{0, \dots, K_{\sf max} - 1\}$, as a discrete r.v. with:

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_{\sf max}-1} \gamma_\ell}. \tag{12}$$

for $k = 0, 1, 2, \dots, K$ **do**
  Draw index $i_k \in [\![1, n]\!]$ uniformly (and $j_k \in [\![1, n]\!]$ for fiSAEM).
  Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using (8) or (9) or (10).
  Compute $\hat{S}^{(k+1)}$ via the Inexact-step (7).
  Compute $\hat{\mathbf{s}}^{(k+1)}$ via the SA-step (6).
  Compute $\hat{\boldsymbol{\theta}}^{(k+1)}$ via the M-step (4).
end for
Return: $\hat{\boldsymbol{\theta}}^{(K)}$.

5 Finite Time Analysis

Finite analysis of iSAEM vrSAEM and fiSAEM .

Analysis in the curved exponential family assumption.

Suboptimality condition would be: $\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}^{(K)})\|^2]$ where

$$\min_{\mathbf{s} \in \mathsf{S}} V(\mathbf{s}) := \overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) = \mathrm{R}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) \tag{13}$$

is the Lyapunov function minimized here.

6 Numerical Examples

6.1 Gaussian Mixture Models

Graphs obtained and relevant

6.2 Logistic Regression with Missing values OR random effects

To Be Done

7 Conclusion

References

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. doi: 10.1214/aos/1018031103. URL https://doi.org/10.1214/aos/1018031103.

B. Karimi. *Non-Convex Optimization for Latent Data Models: Algorithms, Analysis and Applications*. PhD thesis, 2019.

E. Kuhn, C. Matias, and T. Rebafka. Properties of the stochastic approximation em algorithm with mini-batch sampling. *arXiv preprint arXiv:1907.09164*, 2019.

S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Fast incremental method for nonconvex optimization. *arXiv preprint arXiv:1603.06159*, 2016.

A Proof of Theorem