Fast Bi-Level and Incremental Stochastic Approximation of the EM Algorithm

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.,*

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

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 (for ISAEM) and incremental EM like papers .As well as Optim papers (Variance reduction, SAGA etc.)

2 Expectation Maximization Methods

The basic "batch" EM (bEM) method iteratively computes a sequence of estimates $\{\boldsymbol{\theta}^k, k\in\mathbb{N}\}$ with an initial parameter $\boldsymbol{\theta}^0$. Each iteration of bEM is composed of two steps. In the E-step, a surrogate function is computed as $\boldsymbol{\theta}\mapsto Q(\boldsymbol{\theta},\boldsymbol{\theta}^{k-1}) = \sum_{i=1}^n Q_i(\boldsymbol{\theta},\boldsymbol{\theta}^{k-1})$ where $Q_i(\boldsymbol{\theta},\boldsymbol{\theta}') := -\int_{\mathsf{Z}}\log f(z_i,y_i;\theta)p(z_i|y_i;\boldsymbol{\theta}')\mu(\mathrm{d}z_i)$ such that $p(z_i|y_i;\boldsymbol{\theta}) := f(z_i,y_i;\boldsymbol{\theta})/g(y_i,\boldsymbol{\theta})$ is the conditional probability density of the latent variables z_i given the observations y_i. When $f(z_i,y_i;\boldsymbol{\theta})$ is a curved exponential family model, the E-step amounts to computing the conditional expectation of the complete data sufficient statistics,

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

In the M-step, the surrogate function is minimized producing a new fit of the parameter $\boldsymbol{\theta}^k = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{k-1})$.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

3 Monte Carlo Integration and Stochastic Approximation

4 Incremental and Bi-Level Inexact EM MEthods

We first describe the stochastic EM methods to be analyzed under a unified framework. The kth iteration of a generic stochastic EM method is composed of two sub-steps —

$$\textsf{sE-step}: \ \hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} - \gamma_{k+1}\big(\hat{\mathbf{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\big), \tag{4}$$

which is a stochastic version of the E-step in (3). Note $\{\gamma_k\}_{k=1}^{\infty} \in [0,1]$ is a sequence of step sizes, $\boldsymbol{\mathcal{S}}^{(k+1)}$ is a proxy for $\bar{\mathbf{s}}(\hat{\boldsymbol{\theta}}^{(k)})$, and $\bar{\mathbf{s}}$ is defined in (3). The M-step is given by

$$\textsf{M-step}: \ \hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}^{(k+1)}) := \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \ \big\{ \mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \big\langle \hat{\boldsymbol{s}}^{(k+1)} \,|\, \phi(\boldsymbol{\theta}) \big\rangle \big\}, \tag{5}$$

which is controlled by the sufficient statistics determined by the sE-step. The stochastic EM methods differ in the way that $\boldsymbol{\mathcal{S}}^{(k+1)}$ is computed. Existing methods employ stochastic approximation or variance reduction without the need to fully compute $\bar{\mathbf{s}}(\hat{\boldsymbol{\theta}}^{(k)})$. To simplify notations, we define

$$\bar{\mathbf{s}}_i^{(k)} := \bar{\mathbf{s}}_i(\hat{\boldsymbol{\theta}}^{(k)}) = \int_{\mathsf{Z}} S(z_i, y_i) p(z_i|y_i; \hat{\boldsymbol{\theta}}^{(k)}) \mu(\mathrm{d}z_i) \quad \text{and} \quad \bar{\mathbf{s}}^{(\ell)} := \bar{\mathbf{s}}(\hat{\boldsymbol{\theta}}^{(\ell)}) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathbf{s}}_i^{(\ell)}. \tag{6}$$

Note that if $\boldsymbol{\mathcal{S}}^{(k+1)} = \bar{\mathbf{s}}^{(k)}$ and $\gamma_{k+1} = 1$, eq. (4) reduces to the E-step in the classical bEM method. To describe the stochastic EM methods, let $i_k \in [\![1, n]\!]$ be a random index drawn at iteration k and $\tau_i^k = \max\{k' : i_{k'} = i,\ k' < k\}$ be the iteration index where $i \in [\![1, n]\!]$ is last drawn prior to iteration k, we have:

(iEM [Neal and Hinton, 1998]) $$\boldsymbol{\mathcal{S}}^{(k+1)} = \boldsymbol{\mathcal{S}}^{(k)} + \tfrac{1}{n}\big(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\tau_{i_k}^k)}\big) \tag{7}$$

(sEM [Cappé and Moulines, 2009]) $$\boldsymbol{\mathcal{S}}^{(k+1)} = \bar{\mathbf{s}}_{i_k}^{(k)} \tag{8}$$

(sEM-VR [Chen et al., 2018]) $$\boldsymbol{\mathcal{S}}^{(k+1)} = \bar{\mathbf{s}}^{(\ell(k))} + \big(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\ell(k))}\big) \tag{9}$$

The stepsize is set to $\gamma_{k+1} = 1$ for the iEM method; $\gamma_{k+1} = \gamma$ is constant for the sEM-VR method. In the original version of the sEM method, the sequence of step γ_{k+1} is a diminishing step size. Moreover, for iEM we initialize with $\boldsymbol{\mathcal{S}}^{(0)} = \bar{\mathbf{s}}^{(0)}$; for sEM-VR, 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.

fiEM Our analysis framework can handle a new, yet natural application of a popular variance reduction technique to the EM method. The new method, called fiEM, is developed from the SAGA method [Defazio et al., 2014] in a similar vein as in sEM-VR.

For iteration $k \geq 0$, the fiEM method draws *two* indices *independently* and uniformly as $i_k, j_k \in [\![1, n]\!]$. 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 $j \in [\![1, n]\!]$ is last drawn as j_k prior to iteration k. With the initialization $\overline{\boldsymbol{\mathcal{S}}}^{(0)} = \bar{\mathbf{s}}^{(0)}$, we use a slightly different update rule from SAGA inspired by [Reddi et al., 2016], as described by the following recursive updates

$$\boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + \big(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)}\big), \ \ \overline{\boldsymbol{\mathcal{S}}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}\big(\bar{\mathbf{s}}_{j_k}^{(k)} - \bar{\mathbf{s}}_{j_k}^{(t_{j_k}^k)}\big). \tag{10}$$

Algorithm 1 Stochastic EM methods.

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

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_{\mathsf{max}}-1} \gamma_\ell}. \tag{11}$$

for $k = 0, 1, 2, \ldots, K$ **do**
  Draw index $i_k \in [\![1, n]\!]$ uniformly (and $j_k \in [\![1, n]\!]$ for fiEM).
  Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using (8) or (7) or (9) or (10).
  Compute $\hat{\mathbf{s}}^{(k+1)}$ via the sE-step (4).
  Compute $\hat{\boldsymbol{\theta}}^{(k+1)}$ via the M-step (5).
end for
Return: $\hat{\boldsymbol{\theta}}^{(K)}$.

where we set a constant step size as $\gamma_{k+1} = \gamma$.

In the above, the update of $\boldsymbol{\mathcal{S}}^{(k+1)}$ corresponds to an *unbiased estimate* of $\bar{\mathbf{s}}^{(k)}$, while the update for $\overline{\boldsymbol{\mathcal{S}}}^{(k+1)}$ maintains the structure that $\overline{\boldsymbol{\mathcal{S}}}^{(k)} = n^{-1} \sum_{i=1}^{n} \bar{\mathbf{s}}_i^{(t_i^k)}$ for any $k \geq 0$. The two updates of (10) are based on two different and independent indices i_k, j_k that are randomly drawn from $[\![n]\!]$. This is used for our fast convergence analysis in Section 5.

5 Finite Time Analysis

6 Numerical Examples

7 Conclusion

References

O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

J. Chen, J. Zhu, Y. W. Teh, and T. Zhang. Stochastic expectation maximization with variance reduction. In *Advances in Neural Information Processing Systems*, pages 7978–7988, 2018.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

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

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

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 Lemma