An online Minorization-Maximization algorithm

Hien Duy Nguyen, Florence Forbes, Gersende Fort, Olivier Cappé

An online Minorization–Maximization algorithm

Hien Duy Nguyen, Florence Forbes, Gersende Fort, and Olivier Cappé

Abstract Modern statistical and machine learning settings often involve high data volume and data streaming, which require the development of online estimation algorithms. The online Expectation–Maximization (EM) algorithm extends the popular EM algorithm to this setting, via a stochastic approximation approach. We show that an online version of the Minorization–Maximization (MM) algorithm, which includes the online EM algorithm as a special case, can also be constructed in a similar manner. We demonstrate our approach via an application to the logistic regression problem and compare it to existing methods.

Keywords: Expectation–Maximization, Minorization–Maximization, parameter estimation, online algorithms, Stochastic Approximation

1 Introduction

Expectation–Maximization (EM) [6, 17] and Minorization–Maximization (MM) algorithms [15] are important classes of optimization procedures that allow for the construction of estimation routines for many data analytic models, including

Hien Duy Nguyen
School of Mathematics and Physics, University of Queensland, St. Lucia, 4067 QLD, Australia, e-mail: h.nguyen7@uq.edu.au

Florence Forbes
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000, Grenoble, France, e-mail: florence.forbes@inria.fr

Gersende Fort
Institut de Mathématiques de Toulouse, CNRS, Toulouse, France, e-mail: gersende.fort@math.univ-toulouse.fr,

Olivier Cappé
ENS Paris, Universite PSL, CNRS, INRIA, France, e-mail: Olivier.Cappe@cnrs.fr

many finite mixture models. The benefit of such algorithms comes from the use of computationally simple surrogates in place of difficult optimization objectives.

Driven by high volume of data and streamed nature of data acquisition, there has been a rapid development of online and mini-batch algorithms that can be used to estimate models without requiring data to be accessed all at once. Online and mini-batch versions of EM algorithms can be constructed via the classic Stochastic Approximation framework (see, e.g., [2, 13]) and examples of such algorithms include those of [3, 7, 8, 10, 11, 12, 19]. Via numerical assessments, many of the algorithms above have been demonstrated to be effective in mixture model estimation problems. Online and mini-batch versions of MM algorithms on the other hand have largely been constructed following convex optimizations methods (see, e.g., [9, 14, 23]) and examples of such algorithms include those of [4, 16, 18, 22].

In this work, we provide a stochastic approximation construction of an online MM algorithm using the framework of [3]. The main advantage of our approach is that we do not make convexity assumptions and instead replace them with oracle assumptions regarding the surrogates. Compared to the online EM algorithm of [3] that this work is based upon, the `Online MM` algorithm extends the approach to allow for surrogate functions that do not require latent variable stochastic representations, which is especially useful for constructing estimation algorithms for mixture of experts (MoE) models (see, e.g. [20]). We demonstrate the `Online MM` algorithm via an application to the MoE-related logistic regression problem and compare it to competing methods.

Notations. By convention, vectors are column vectors. For a matrix A, $A^\top$ denotes its transpose. The Euclidean scalar product is denoted by $\langle a, b\rangle$. For a continuously differentiable function $\theta \mapsto h(\theta)$ (resp. twice continuously differentiable), $\nabla_\theta h$ (or simply ∇ when there is no confusion) is its gradient (resp. $\nabla^2_{\theta\theta}$ is its Hessian).

2 The online MM algorithm

Consider the optimization problem

$$\arg\max_{\theta\in\mathbb{T}} \mathbb{E}\left[f\left(\theta; X\right)\right], \tag{1}$$

where $\mathbb{T}$ is a measurable open subset of $\mathbb{R}^p$, $\mathbb{X}$ is a topological space endowed with its Borel sigma-field, $f : \mathbb{T}\times\mathbb{X} \to \mathbb{R}$ is a measurable function and X is a $\mathbb{X}$-valued random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In this paper, we are interested in the setting when the expectation $\mathbb{E}\left[f\left(\theta; X\right)\right]$ has no closed form, and the optimization problem is solved by an MM-based algorithm.

Following the terminology of [15], we say that $g : \mathbb{T}\times\mathbb{X}\times\mathbb{T}, (\theta, x, \tau) \mapsto g\left(\theta, x; \tau\right)$ is a *minorizer of* f, if for any $\tau \in \mathbb{T}$ and for any $(\theta, x) \in \mathbb{T}\times\mathbb{X}$, it holds that

$$f(\theta; x) - f(\tau; x) \geq g(\theta, x; \tau) - g(\tau, x; \tau). \tag{2}$$

In our work, we consider the case when the minorizer function g has the following structure:

A1 The minorizer surrogate g is of the form:

$$g(\theta, x; \tau) = -\psi(\theta) + \left\langle \bar{S}(\tau; x), \phi(\theta) \right\rangle, \tag{3}$$

where $\psi : \mathbb{T} \to \mathbb{R}$, $\phi : \mathbb{T} \to \mathbb{R}^d$ and $\bar{S} : \mathbb{T} \times \mathbb{X} \to \mathbb{R}^d$ are measurable functions. In addition, ϕ and ψ are continuously differentiable on $\mathbb{T}$.

We also make the following assumptions:

A2 There exists a measurable open and convex set $\mathbb{S} \subseteq \mathbb{R}^d$ such that for any $s \in \mathbb{S}$, $\gamma \in [0, 1)$ and any $(\tau, x) \in \mathbb{T} \times \mathbb{X}$:

$$s + \gamma \left\{ \bar{S}(\tau; x) - s \right\} \in \mathbb{S}.$$

A3 The expectation $\mathbb{E}[\bar{S}(\theta; X)]$ exists, is in $\mathbb{S}$, and is finite whatever $\theta \in \mathbb{T}$ but it may have no closed form. Online independent oracles $\{X_n, n \geq 0\}$, with the same distribution as X, are available.

A4 For any $s \in \mathbb{S}$, there exists a unique root to $\theta \mapsto -\nabla\psi(\theta) + \nabla\phi(\theta)^\top s$, which is the unique maximum on $\mathbb{T}$ of the function $\theta \mapsto -\psi(\theta) + \langle s, \phi(\theta) \rangle$. This root is denoted by $\bar{\theta}(s)$.

Seen as a function of θ, $g(\cdot, x; \tau)$ is the sum of two functions: $-\psi$ and a linear combination of the components of $\phi = (\phi_1, \ldots, \phi_d)$. Assumption A1 implies that the minorizer surrogate is in a functional space spanned by these $(d + 1)$ functions. By (2) and A1–A3, it follows that

$$\mathbb{E}\left[f(\theta; X)\right] - \mathbb{E}\left[f(\tau; X)\right] \geq \psi(\tau) - \psi(\theta) + \left\langle \mathbb{E}\left[\bar{S}(\tau; X)\right], \phi(\theta) - \phi(\tau) \right\rangle, \tag{4}$$

thus providing a minorizer function for the objective function $\theta \mapsto \mathbb{E}\left[f(\theta; X)\right]$. By A4, the usual MM algorithm would define iteratively the sequence $\theta_{n+1} = \bar{\theta}\left(\mathbb{E}\left[\bar{S}(\theta_n; X)\right]\right)$. Since the expectation may not have closed form but infinite datasets are available (see A3), we propose a novel `Online MM` algorithm. It defines the sequence $\{s_n, n \geq 0\}$ as follows: given positive step sizes $\{\gamma_{n+1}, n \geq 1\}$ in $(0, 1)$ and an initial value $s_0 \in \mathbb{S}$, set for $n \geq 0$:

$$s_{n+1} = s_n + \gamma_{n+1} \left\{ \bar{S}\left(\bar{\theta}(s_n); X_{n+1}\right) - s_n \right\}. \tag{5}$$

The update mechanism (5) is a Stochastic Approximation iteration, which defines an $\mathbb{S}$-valued sequence (see A2). It consists of the construction of a sequence of minorizer functions through the definition of their *parameter* s_n in the functional space spanned by $-\psi, \phi_1, \ldots, \phi_d$.

If our algorithm (5) converges, any limiting point $s_\star$ satisfies $\mathbb{E}\left[\bar{S}(\bar{\theta}(s_\star); X)\right] = s_\star$. Hence, our algorithm is designed to approximate the intractable expectation, evaluated at $\bar{\theta}(s_\star)$, where $s_\star$ satisfies a fixed point equation. The following lemma establishes the relation between the limiting points of (5) and the optimization problem (1) at hand. Namely, it implies that any limiting value $s_\star$ provides a stationary

point $\theta_\star := \bar{\theta}(s_\star)$ of the objective function $\mathbb{E}\left[f\left(\theta; X\right)\right]$ (i.e., $\theta_\star$ is a root of the derivative of the objective function). The proof follows the technique of [3]. Set

$$\mathsf{h}(s) := \mathbb{E}\left[\bar{S}\left(\bar{\theta}\left(s\right); X\right)\right] - s, \qquad \Gamma := \{s \in \mathbb{S} : \mathsf{h}(s) = 0\}.$$

Lemma 1 *Assume that $\theta \mapsto \mathbb{E}\left[f(\theta; X)\right]$ is continuously differentiable on $\mathbb{T}$ and denote by $\mathcal{L}$ the set of its stationary points. If $s_\star \in \Gamma$, then $\bar{\theta}(s_\star) \in \mathcal{L}$. Conversely, if $\theta_\star \in \mathcal{L}$, then $s_\star := \mathbb{E}\left[\bar{S}\left(\theta_\star; X\right)\right] \in \Gamma$.*

Proof A4 implies that

$$-\nabla\psi(\bar{\theta}(s)) + \nabla\phi(\bar{\theta}(s))^\top s = 0, \qquad s \in \mathbb{S}. \tag{6}$$

Use (2) and A1, and apply the expectation w.r.t. X (under A3). This yields (4), which is available for any $\theta, \tau \in \mathbb{T}$. This inequality provides a minorizer function for $\theta \mapsto \mathbb{E}\left[f(\theta; X)\right]$: the difference is nonnegative and minimal (i.e. equal to zero) at $\theta = \tau$. Under the assumptions and A1, this yields

$$\nabla\mathbb{E}\left[f(\cdot; X)\right]|_{\theta=\tau} + \nabla\psi(\tau) - \nabla\phi(\tau)^\top\mathbb{E}\left[\bar{S}(\tau; X)\right] = 0. \tag{7}$$

Let $s_\star \in \Gamma$ and apply (7) with $\tau \leftarrow \bar{\theta}(s_\star)$. It then follows that

$$\nabla\mathbb{E}\left[f(\cdot; X)\right]|_{\theta=\bar{\theta}(s_\star)} + \nabla\psi(\bar{\theta}(s_\star)) - \nabla\phi(\bar{\theta}(s_\star))^\top s_\star = 0,$$

which implies $\bar{\theta}(s_\star) \in \mathcal{L}$ by (6). Conversely, if $\theta_\star \in \mathcal{L}$, then by (7), we have

$$\nabla\psi(\theta_\star) - \nabla\phi(\theta_\star)^\top\mathbb{E}\left[\bar{S}(\theta_\star; X)\right] = 0,$$

which, by A3 and A4, implies that $\theta_\star = \bar{\theta}\left(\mathbb{E}\left[\bar{S}(\theta_\star; X)\right]\right) = \bar{\theta}(s_\star)$. By definition of $s_\star$, this yields $s_\star = \mathbb{E}\left[\bar{S}\left(\bar{\theta}(s_\star); X\right)\right]$; i.e. $s_\star \in \Gamma$. □

By applying the results of [5] regarding the asymptotic convergence of Stochastic Approximation algorithms, additional regularity assumptions on $\phi, \psi, \bar{\theta}$ imply that the algorithm (5) possesses a continuously differentiable Lyapunov function V defined on $\mathbb{S}$ and given by $V : s \mapsto \mathbb{E}\left[f(\bar{\theta}(s); X)\right]$, satisfying $\langle\nabla V(s), \mathsf{h}(s)\rangle \leq 0$, where the inequality is strict outside the set Γ (see [3, Prop. 2]). In addition to Lemma 1, assumptions on the distribution of X and on the stability of the sequence $\{s_n, n \geq 0\}$ are provided in [5, Thm. 2 and Lem. 1], which, combined with the usual conditions on the step sizes: $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < \infty$, yields the almost-sure convergence of the sequence $\{s_n, n \geq 0\}$ to the set Γ, and the almost-sure convergence of the sequence $\{\bar{\theta}(s_n), n \geq 0\}$ to the set $\mathcal{L}$ of the stationary points of the objective function $\theta \mapsto \mathbb{E}\left[f(\theta; X)\right]$. Due to the limited space, the exact statement of these convergence results for our `Online MM` framework is omitted.

3 Example application

As an example, we consider the logistic regression problem, where we solve (1) with

$$f(\theta; x) := yw^\top\theta - \log\left\{1 + \exp\left(w^\top\theta\right)\right\}, \qquad x := (y, w),$$

where $y \in \{0, 1\}$, $w \in \mathbb{R}^p$, and $\theta \in \mathbb{T} := \mathbb{R}^p$. Here, we assume that $X = (Y, W)$ is a random variable such that $\mathbb{E}[f(\theta; X)]$ exists for each θ.

Denote by λ the standard logistic function $\lambda(\cdot) := \exp\{\cdot\}/(1+\exp\{\cdot\})$. Following [1], (2) and A1 are verified by taking

$$\psi(\theta) := 0, \qquad \phi(\theta) := \begin{bmatrix} \theta \\ \mathrm{vec}\,(\theta\theta^\top) \end{bmatrix}, \qquad \bar{S}(\tau; x) = \begin{bmatrix} \bar{s}_1(\tau; x) \\ \mathrm{vec}\left(\bar{S}_2(\tau; x)\right) \end{bmatrix}$$

where

$$\bar{s}_1(\tau; x) := \left\{y - \lambda\left(\tau^\top w\right)\right\} w + \frac{1}{4} ww^\top\tau, \quad \bar{S}_2(\tau; x) = -\frac{1}{8} ww^\top.$$

With $\mathbb{S} := \{(s_1, \mathrm{vec}\,(S_2)) : s_1 \in \mathbb{R}^p$ and $S_2 \in \mathbb{R}^{p\times p}$ is symmetric positive definite$\}$, it follows that $\bar{\theta}(s) := -(2S_2)^{-1}s_1$.

Online MM. Let $s_n = \left(s_{1,n}, S_{2,n}\right) \in \mathbb{S}$. The corresponding `Online MM` recursion is then

$$s_{1,n+1} = s_{1,n} + \gamma_{n+1}\left(Y_{n+1} - \lambda\left(\bar{\theta}(s_n)^\top W_{n+1}\right) W_{n+1} + \frac{1}{4} W_{n+1}W_{n+1}^\top\bar{\theta}(s_n) - s_{1,n}\right) \quad (8)$$

$$S_{2,n+1} = S_{2,n} + \gamma_{n+1}\left(-\frac{1}{8} W_{n+1}W_{n+1}^\top - S_{2,n}\right), \quad (9)$$

where $\{(Y_{n+1}, W_{n+1}), n \geq 0\}$ are i.i.d. pairs with the same distribution as $X = (Y, W)$. Parameter estimates can then be deduced by setting $\theta_{n+1} := \bar{\theta}(s_{n+1})$.

For comparison, we also consider two Stochastic Approximation schemes directly on θ in the parameter-space: a stochastic gradient (`SG`) algorithm and a Stochastic Newton Raphson (`SNR`) algorithm.

Stochastic gradient. `SG` requires the gradient of $f(\theta; x)$ with respect to θ: $\nabla f(\theta; x) = \{y - \lambda(\theta^\top w)\} w$, which leads to the recursion

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1}\left\{Y_{n+1} - \lambda(\hat{\theta}_n^\top W_{n+1})\right\} W_{n+1}. \quad (10)$$

Stochastic Newton-Raphson. In addition `SNR` requires the Hessian with respect to θ, given by $\nabla^2_{\theta\theta} f(\theta; x) = -\lambda(\theta^\top w)\{1 - \lambda(\theta^\top w)\} ww^\top$. The `SNR` recursion is then

$$\hat{A}_{n+1} = \hat{A}_n + \gamma_{n+1}\left\{\nabla^2_{\theta\theta} f(\hat{\theta}_n; X_{n+1}) - \hat{A}_n\right\} \quad (11)$$

$$G_{n+1} = -\hat{A}_{n+1}^{-1} \quad (12)$$

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1} G_{n+1}\left\{Y_{n+1} - \lambda(\hat{\theta}_n^\top W_{n+1})\right\} W_{n+1}\ . \quad (13)$$

Equation (12) assumes that $\hat{A}_{n+1}$ is invertible. In this logistic example, we can guarantee this by choosing $\hat{A}_0$ to be invertible. Otherwise $\hat{A}_n$ is invertible after some n sufficiently large, with probability one. Again in the logistic case, observe that, from the structure of $\nabla^2_{\theta\theta} f$ and from the Woodbury matrix identity, Equations (11–12) can be replaced by

$$G_{n+1} = \frac{G_n}{1-\gamma_{n+1}} - \frac{\gamma_{n+1}}{1-\gamma_{n+1}} \frac{a_{n+1} G_n W_{n+1} W_{n+1}^\top G_n}{\left\{(1-\gamma_{n+1}) + \gamma_{n+1} a_{n+1} W_{n+1}^\top G_n W_{n+1}\right\}} \ .$$

where $a_{n+1} := \lambda(\hat{\theta}_n^\top W_{n+1}) \left\{1 - \lambda(\hat{\theta}_n^\top W_{n+1})\right\}$,

It appears that the `Online MM` recursion in the s-space defined by (8) and (9) is equivalent to the `SNR` recursion above (i.e., (11)–(13)) when the Hessian $\nabla^2_{\theta\theta} f(\theta; x)$ is replaced by the lower bound $-\frac{1}{4} w w^\top$. This observation holds whenever g is quadratic in $(\theta - \tau)$.

Polyak averaging. In practice, for `Online MM`, `SG`, and `SNR` recursions, it is common to consider Polyak averaging [21], starting from some iteration n_0, chosen such as to avoid the initial highly volatile estimates. Set $\hat{\theta}^A_{n_0} := 0$, and for $n \geq n_0$,

$$\hat{\theta}^A_{n+1} = \hat{\theta}^A_n + \alpha_{n-n_0+1}(\hat{\theta}_n - \hat{\theta}^A_n), \tag{14}$$

where α_n is usually set to $\alpha_n := n^{-1}$.

Numerical illustration. We now demonstrate the performance of the `Online MM` algorithm for logistic regression – defined by (5) and the derivations above. To do so, a sequence $\{X_i = (Y_i, W_i)\,, i \in \{1, \ldots, n_{\max}\}\}$ of $n_{\max} = 10^5$ i.i.d. replicates of $X = (Y, W)$ is simulated: $W = (1, U)$, where $U \sim \mathrm{N}(0, 1)$ and $[Y|W = w] \sim \mathrm{Ber}\left(\lambda\left(\theta_0^\top w\right)\right)$, where $\theta_0 = (3, -3)$. `Online MM` is run using the learning rate $\gamma_n = n^{-0.6}$, as suggested in [3]. The algorithm is initialized with $\hat{\theta}_0 = (0, 0)$ and $s_0 = \sum_{i=1}^2 \bar{S}\left(\hat{\theta}_0; X_i\right)/2$.

For comparison, we also show, on Figure 1, the `SG`, `SNR` estimates and their Polyak averaged values in θ-space. As is usually recommended with Stochastic Approximation, the first few volatile estimations are discarded. Similarly, for Polyak averaging, we set $n_0 = 10^3$. As expected, we observe that the `Online MM` and the `SNR` recursions are very close but with the `SNR` showing more variability. Their comparison after Polyak averaging shows very close trajectories while the `SG` trajectory is clearly different and shows more bias. Final estimates [Polyak averaged estimates] of θ_0 from the `SG`, `SNR`, and `Online MM` algorithms are respectively: $(2.67, -2.66)$ $[(2.51, -2.48)]$, $(3.03, -3.03)$ $[(2.99, -3.03)]$, and $(3.01, -3.03)$ $[(2.98, -3.02)]$, which we can compare to the batch maximum likelihood estimate $(3.00, -3.05)$ (obtained via the `glm` function in `R`). Notice the remarkable closeness between the `online MM` and batch estimates.

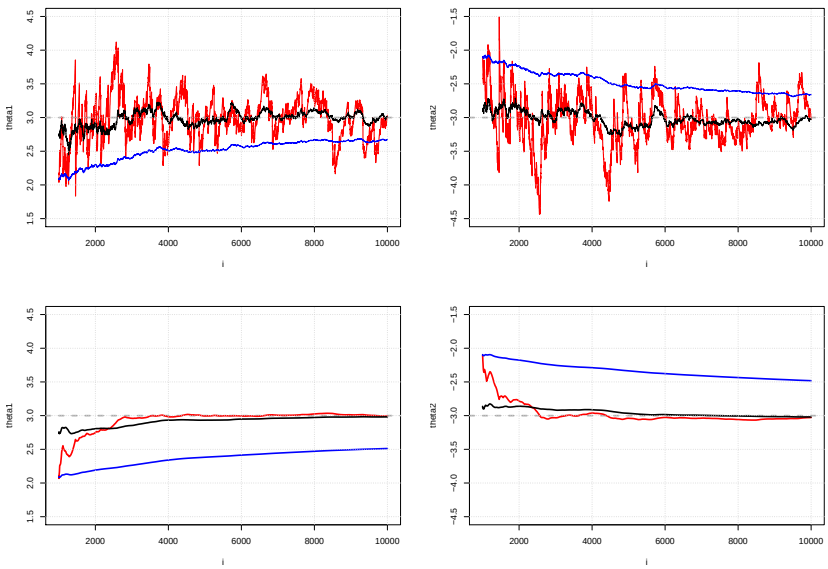

Fig. 1 Logistic regression example: the first row shows `Online MM` (black), `SG` (blue), and `SNR` (red) recursions. The second row shows the respective Polyak averaging recursions. The estimates of the first θ (first column) and the second (second column) components of θ are plotted started from $n = 10^3$ for readability.

4 Final remarks

Remark 1 For a parametric statistical model indexed by θ, let $f(\theta; x)$ be the log-density of a random variable X with stochastic representation $f(\theta; x) = \log \int_{\mathbb{Y}} p_\theta(x, y)\, \mu(\mathrm{d}y)$, where $p_\theta(x, y)$ is the joint density of (X, Y) with respect to the positive measure μ for some latent variable $Y \in \mathbb{Y}$. Then, via [15, Sec. 4.2], we recover the `Online EM` algorithm by using the minorizer function g:

$$g(\theta, x; \tau) := \int_{\mathbb{Y}} \log p_\theta(x, y)\ p_\tau(x, y) \exp(-f(\tau; x))\, \mu(\mathrm{d}y).$$

Remark 2 Via the minorization approach of [1] (as used in Section 3) and the mixture representation from [19], we can construct an `Online MM` algorithm for MoE models, analogous to the MM algorithm of [20]. We shall provide exposition on such an algorithm in future work.

Acknowledgements. Part of the work by G. Fort is funded by the *Fondation Simone et Cino Del Duca, Institut de France*. H. Nguyen is funded by ARC Grant DP180101192. The work is supported by Inria project LANDER.

References

1. Bohning, D.: Multinomial logistic regression algorithm. Annals of the Institute of Statistical Mathematics (1992)
2. Borkar, V.S.: Stochastic Approximation: A Dynamical Systems Viewpoint. Springer (2009)
3. Cappé, O., Moulines, E.: On-line expectation-maximization algorithm for latent data models. Journal of the Royal Statistical Society B **71**, 593–613 (2009)
4. Cui, Y., Pang, J.: Modern Nonconvex nondifferentiable optimization. SIAM, Philadelphia (2022)
5. Delyon, B., Lavielle, M., Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm. Annals of Statistics **27**, 94–128 (1999)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B **39**, 1–38 (1977)
7. Fort, G., Gach, P., Moulines, E.: Fast incremental expectation maximization for finite-sum optimization: nonasymptotic convergence. Statistics and Computing **31**, 1–24 (2021)
8. Fort, G., Moulines, E., Wai, H.T.: A stochastic path-integrated differential estimator expectation maximization algorithm. In: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS) (2020)
9. Hazan, E.: Introduction to Online Convex Optimization. Foundations and Trends in Optimization **2** (2015)
10. Karimi, B., Miasojedow, B., Moulines, E., Wai, H.T.: Non-asymptotic analysis of biased stochastic approximation scheme. Proceedings of Machine Learning Research **99**, 1–31 (2019)
11. Karimi, B., Wai, H.T., Moulines, R., Lavielle, M.: On the global convergence of (fast) incremental expectation maximization methods. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS) (2019)
12. Kuhn, E., Matias, C., Rebafka, T.: Properties of the stochastic approximation EM alpgorithm with mini-batch sampling. Statistics and Computing **30**, 1725–1739 (2020)
13. Kushner, H.J., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications. Springer, New York (2003)
14. Lan, G.: First-order and Stochastic Optimization Methods for Machine Learning. Springer, Cham (2020)
15. Lange, K.: MM Optimization Algorithms. SIAM, Philadelphia (2016)
16. Mairal, J.: Stochastic majorization-minimization algorithm for large-scale optimization. In: Advances in Neural Information Processing Systems, pp. 2283–2291 (2013)
17. McLachlan, G.J., Krishnan, T.: The EM Algorithm And Extensions. Wiley, New York (2008)
18. Mokhtari, A., Koppel, A.: High-dimensional nonconvex stochastic optimization by doubly stochastic successive convex approximation. IEEE Transactions on Signal Processing **68**, 6287–6302 (2020)
19. Nguyen, H.D., Forbes, F., McLachlan, G.J.: Mini-batch learning of exponential family finite mixture models. Statistics and Computing **30**, 731–748 (2020)
20. Nguyen, H.D., McLachlan, G.J.: Laplace mixture of linear experts. Computational Statistics and Data Analysis **93**, 177–191 (2016)
21. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM Journal of Control and Optimization **30**, 838–855 (1992)
22. Razaviyayn, M., Sanjabi, M., Luo, Z.: A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. Mathematical Programming Series B pp. 515–545 (2016)
23. Shalev-Shwartz, S.: Online learning and online convex optimization. Foundations and Trends in Machine Learning **4**, 107–194 (2011)