Stochastic Approximation
Beyond Gradient

Gersende Fort
CNRS
Institut de Mathématiques de Toulouse, France

European Meeting of Statisticians 2023, Warsaw, Poland

In collaboration with

- Aymeric Dieuleveut, Ecole Polytechnique, CMAP, France
- Eric Moulines, Ecole Polytechnique, CMAP, France
- Hoi To Wai, Chinese Univ. of Hong-Kong, Hong-Kong

Talk based on the paper:

- *Stochastic Approximation Beyond Gradient for Signal Processing and Machine Learning*
  by A. Dieuleveut, G. Fort, E. Moulines and H.-T. Wai
  HAL-03979922 and arXiv:2302.11147

Partly funded by
Fondation Simone et Cino Del Duca, Project OpSiMorE

Outline

1. Stochastic Approximation, beyond gradient
2. Two examples
3. Non asymptotic convergence bounds in expectation
4. Variance Reduction by the `SPIDER` control variate

I. Stochastic Approximation

Stochastic Approximation: solve the root-finding problem

Solve

$$\omega \in \mathbb{R}^d \qquad \text{s.t.} \quad h(\omega) = 0$$

when **only stochastic estimates** of the *mean field* h are available.

by an iterative algorithm

$$\omega_{k+1} = \omega_k + \gamma_{k+1} \; H(\omega_k, X_{k+1})$$

where

- γ_{k+1} is a positive step size
- $H(\omega_k, X_{k+1})$ is a stochastic oracle for $h(\omega_k)$.

Time discretization of the ODE

$$\frac{d\omega}{dt} = h\left(\omega(t)\right)$$

yields $\tau_{k+1} = \tau_k + \gamma_{k+1} \; h(\tau_k)$.

Beyond the gradient case

The gradient case:

- Solve $\mathrm{argmin}_\omega f(\omega)$ "by" solving $\nabla f = 0$ when $-\nabla f(\omega) = \mathbb{E}\left[H(\omega, X)\right]$

Expected Risk Minimization	for batch data	$f(\omega) = (1/n)\sum_{i=1}^n \ell(\omega, Z_i)$
	for streaming data	$f(\omega) = \mathbb{E}\left[\ell(\omega, X)\right]$

- Available oracles given ω_k, a random variable X_{k+1} and the stochastic gradient term $H(\omega_k, X_{k+1})$.

Expected Risk Minimization	for batch data	$H(\omega, X_{k+1}) = -\mathsf{D}_{10}\ell(\omega, Z_{X_{k+1}}) \quad X_{k+1} \in \{1, \cdots, n\}$
	for streaming data	$H(\omega, X_{k+1}) = -\mathsf{D}_{10}\ell(\omega, X_{k+1})$

Two extensions:

- The function h is not necessarily a gradient
- The oracle can be **biased**

$$\mathbb{E}\left[H(\omega_k, X_{k+1})\middle|\mathcal{F}_k\right] \neq h(\omega_k) \qquad \mathcal{F}_k := \sigma(\omega_0, X_1, \ldots, X_k)$$

II. Two examples of *SA beyond gradient*

1st ex.: Compressed gradient

- **Compression operator** $\mathcal{C}(x, U)$, if the cost of storing/transmitting $\mathcal{C}(x, U)$ is less than the cost of storing/transmitting x

- First family:

$$\omega_{k+1} = \omega_k + \gamma_{k+1}\, \mathcal{C}(H(\omega_k, X_{k+1}), U_{k+1})$$

Ex. The Gauss-Southwell coordinate descent estimator $\quad \mathcal{C}(x, u) = x_u e_u \quad u \in \{1, \cdots, d\}$

- Second family:

$$\omega_{k+1} = \omega_k + \gamma_{k+1}\, H(\mathcal{C}(\omega_k, U_{k+1}), X_{k+1})$$

Ex. Stoch Gdt for deep learning, the Straight-Through Estimator quantizes the model ω_k before computing the oracle.

- Third family:

$$\omega_{k+1} = \mathcal{C}\left(\omega_k + \gamma_{k+1}\, H(\omega_k, X_{k+1}), U_{k+1}\right)$$

Ex. Low precision Stoch Gdt: the model is quantized after computing the oracle.

2nd ex.: Stochastic Expectation Maximization in the curved exponential family

- The goal

$$\mathrm{argmin}_\theta f(\theta) := -\log \int_{\mathcal{D}} p(x;\theta)\,\nu(\mathrm{d}x) \qquad p(x;\theta) = \xi(x)\exp(\langle S(x), \phi(\theta)\rangle - \psi(\theta))$$

- The EM algorithm

$$\theta_k \xrightarrow[\text{compute an expectation}]{\text{E-step}} \bar{\mathsf{S}}(\theta_k) \xrightarrow[\text{optimize}]{\text{M-step}} \theta_{k+1} := \mathsf{T}(\bar{\mathsf{S}}(\theta_k))$$

$$s_k := \bar{\mathsf{S}}(\theta_k) \xrightarrow[\text{optimize}]{\text{M-step}} \mathsf{T}(s_k) \xrightarrow{\text{E-step}} s_{k+1} := \bar{\mathsf{S}}(\mathsf{T}(s_k))$$

where

$$\bar{\mathsf{S}}(\theta) := \int_{\mathcal{D}} S(x)\,\pi(x;\theta)\,\nu(\mathrm{d}x) \qquad \mathsf{T}(s) := \mathrm{argmin}_\theta\, \psi(\theta) - \langle s, \phi(\theta)\rangle$$

- Limiting points of EM

EM finds $\theta_\star$ solving the root-finding pbm

$$\mathsf{T}(\bar{\mathsf{S}}(\theta)) = \theta$$

or

EM finds $s_\star$ solving the root-finding pbm

$$\bar{\mathsf{S}}(\mathsf{T}(s)) = s$$

and then set

$$\theta_\star = \mathsf{T}(s_\star).$$

Which oracle of $\bar{\mathsf{S}}$, in order to solve

$$\bar{\mathsf{S}}(\mathsf{T}(\omega)) - \omega = 0 \qquad\qquad \bar{\mathsf{S}}(\mathsf{T}(\omega)) := \int S(x)\, \pi(x; \mathsf{T}(\omega))\, \nu(\mathrm{d}x)$$

- Stochastic Approximation EM (SAEM) when $\bar{\mathsf{S}}$ intractable

$$\omega_{k+1} = \omega_k + \gamma_{k+1} \left(\frac{1}{M} \sum_{m=1}^{M} S(X_{k+1,m}) - \omega_k \right)$$

where $X_{k+1,m}$ obtained from self-normalized importance sampling, MCMC, $\cdots$

- Mini-batch EM: when $S(x) := (1/n) \sum_{i=1}^{n} S_i(x)$ and large n

$$\omega_{k+1} = \omega_k + \gamma_{k+1} \left(\frac{1}{\mathsf{b}} \sum_{i \in X_{k+1}} \bar{\mathsf{S}}_i(\mathsf{T}(\omega_k)) - \omega_k \right).$$

III. Non-asymptotic convergence bounds in expectation

The assumptions

$$\omega_{k+1} = \omega_k + \gamma_{k+1} H(\omega_k, X_{k+1})$$

On the oracles, for some $W \geq 0$:

- L^2-moment $\quad \mathbb{E}\left[\|H(\omega_k, X_{k+1})\|^2\right] < \infty$
- Growth of the mean field $\quad \exists c_0, c_1, \forall \omega \;\; \|h(\omega)\|^2 \leq c_0 + c_1 W(\omega)$
- Bias $\quad \exists \tau_0, \tau_1, \; \forall k \; \left\|\mathbb{E}\left[H(\omega_k, X_{k+1}) \middle| \mathcal{F}_k\right] - h(\omega_k)\right\|^2 \leq \tau_0 + \tau_1 W(\omega_k)$ a.s.
- Variance
  $\exists \sigma_0, \sigma_1, \forall k, \mathbb{E}\left[\left\|H(\omega_k, X_{k+1}) - \mathbb{E}\left[H(\omega_k, X_{k+1}) \middle| \mathcal{F}_k\right]\right\|^2 \middle| \mathcal{F}_k\right] \leq \sigma_0^2 + \sigma_1^2 W(\omega_k)$ a.s.

Ex. Gradient: $W(\omega) = \|\nabla f(\omega)\|^2$ $\qquad$ Stoch EM: $W(\omega) = \|\bar{\mathsf{S}}(\mathsf{T}(\omega)) - \omega\|^2$

A smooth Lyapunov function: V

- Lower bounded $\quad \inf_\omega V(\omega) > -\infty$
- Smooth fct $\quad V$ is C^1 and $\exists L_V, \;\; \|\nabla V(\omega) - \nabla V(\omega')\| \leq L_V \|\omega - \omega'\|$
- Lyapunov V and control W $\quad \exists \rho \geq 0, \forall \omega \;\; \langle \nabla V(\omega), h(\omega) \rangle \leq -\rho W(\omega)$

Ex. Gradient: $V(\omega) = f(\omega)$ $\qquad$ Stoch EM: $V(\omega) = f(\mathsf{T}(\omega))$

If **biased** oracles i.e. $\tau_0 + \tau_1 > 0$, additional conditions

$$c_V := \sup_\omega \frac{\|\nabla V(\omega)\|^2}{W(\omega)} < \infty, \qquad \sqrt{c_V}\,\left(\sqrt{\tau_0}/2 + \sqrt{\tau_1}\right) < \rho.$$

Main theorem

Theorem 1, Dieuleveut-F.-Moulines-Wai (2023)

In addition to the previous assumptions, assume that $\gamma_k \in (0, \gamma_{\max})$. Then, for any $T \geq 1$

$$\sum_{k=0}^{T-1} \frac{\gamma_{k+1}\mu_{k+1}}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\mu_{\ell+1}} \mathbb{E}\left[W(\omega_k)\right]$$

$$\leq 2\frac{\mathbb{E}\left[V(\omega_0)\right] - \min V}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\mu_{\ell+1}} \qquad \text{initial cond.}$$

$$+ L_V \eta_0 \frac{\sum_{k=0}^{T-1}\gamma_{k+1}^2}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\mu_{\ell+1}}$$

$$+ 2\mathfrak{b}_0 \frac{\sum_{k=0}^{T-1}\gamma_{k+1}}{\sum_{\ell=0}^{T-1}\gamma_{\ell+1}\mu_{\ell+1}} \qquad \mathfrak{b}_0 = 0 \text{ iff unbiased oracles}$$

- Meaningful results under the assumptions

$$\min_{\omega:\mathrm{d}(\omega,\{h=0\})>\epsilon} W(\omega) > 0 \qquad \forall \epsilon > 0$$

A Robbins-Siegmund type inequality

Lemma 9, Dieuleveut-F.-Moulines-Wai (2023)

$$\mathbb{E}\left[V(\omega_{k+1})\middle|\mathcal{F}_k\right] \le V(\omega_k) - \gamma_{k+1} \underbrace{\mu_{k+1}}_{\substack{\text{positive for} \\ \gamma_{k+1} \text{ small enough}}} W(\omega_k)$$
$$+ \gamma_{k+1} \underbrace{\mathsf{b}_0}_{\substack{\ge 0 \text{ and zero} \\ \text{iff unbiased oracles}}}$$
$$+ \gamma_{k+1}^2 \tilde{\mathsf{b}}$$

- From the assumptions on the Lyapunov function V

$$V(\omega_{k+1}) \le V(\omega_k) + \left\langle \nabla V(\omega_k), \omega_{k+1} - \omega_k \right\rangle + \frac{L_V}{2} \|\omega_{k+1} - \omega_k\|^2.$$

- Use

$$\omega_{k+1} - \omega_k = \gamma_{k+1} h(\omega_k) + \gamma_{k+1} \left(H(\omega_k, X_{k+1}) - h(\omega_k) \right).$$

- Negative term: $\left\langle \nabla V(\omega_k), \gamma_{k+1}\, h(\omega_k) \right\rangle \le -\rho\gamma_{k+1}\, W(\omega_k)$
- Apply the conditional expectations $\mathbb{E}\left[\cdot\middle|\mathcal{F}_k\right]$ and use the assumptions on the bias and variance of the oracles.

Corollary 1: which parameter ?

On the left hand side:

$$\sum_{k=0}^{T-1} \frac{\gamma_{k+1}\mu_{k+1}}{\sum_{\ell=0}^{T-1} \gamma_{\ell+1}\mu_{\ell+1}} \mathbb{E}\left[W(\omega_k)\right]$$

- If **W is convex**,

$$W\left(\sum_{k=0}^{T-1} \frac{\gamma_{k+1}\mu_{k+1}}{\sum_{\ell=0}^{T-1} \gamma_{\ell+1}\mu_{\ell+1}} \omega_k\right) \leq \sum_{k=0}^{T-1} \frac{\gamma_{k+1}\mu_{k+1}}{\sum_{\ell=0}^{T-1} \gamma_{\ell+1}\mu_{\ell+1}} W(\omega_k)$$

adopt a convex combination of the iterates

- Otherwise,

$$\sum_{k=0}^{T-1} \frac{\gamma_{k+1}\mu_{k+1}}{\sum_{\ell=0}^{T-1} \gamma_{\ell+1}\mu_{\ell+1}} W(\omega_k) = \mathbb{E}\left[W\left(\omega_R\right)\right] \qquad \mathbb{P}(R = k) \propto \gamma_{k+1}\mu_{k+1}.$$

stop at a random time / choose randomly one of the iterates

Corollary 2: Constant step size

$$\gamma := \frac{\gamma_{\max}}{2} \wedge O\left(\frac{1}{\sqrt{T}}\right)$$

then

$$\frac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}\left[W(\omega_k)\right] \leq \underbrace{B}_{\substack{\text{null} \\ \text{iff unbiased oracles}}} + \frac{A_1}{\sqrt{T}} \wedge \frac{A_2}{T}$$

- If **unbiased** oracles,
  the RHS goes to zero when $T \to +\infty$ by choosing $\gamma_k = \gamma_{\max}/\sqrt{T}$
  the convergence rate of SA is $O(1/\sqrt{T})$.
- If **biased oracles**: the RHS can not be made small when the step sizes are constant.

Corollary 3: ϵ-approximate stationarity

In non-convex optimization: in general, it is intractable to find a global minimum or to test if a point is a local minimum.

- Stationarity as a convergence criterion: For a precision ϵ, find a random stopping time R s.t.

$$\mathbb{E}\left[W(\omega_R)\right] \leq \epsilon.$$

- When the oracles are **unbiased**: choose R *uniform* on $\{1, \cdots, T\}$ for T larger than

$$T(\epsilon) := \frac{A_3}{\epsilon^2} \vee \frac{A_4}{\epsilon}.$$

high-precision regime:	$T(\epsilon) = O(1/\epsilon^2)$	step size $\gamma_\epsilon = O(\epsilon)$
low-precision regime:	$T(\epsilon) = O(1/\epsilon)$	step size $\gamma_\epsilon = \gamma_{\max}/2$

- Section III-A, Dieuleveut-F.-Moulines-Wai (2023) explicit constants, not detailed here

IV. Variance reduction by SPIDER

Control variates

- The oracles are not unique:

$$\mathbb{E}\left[H(\omega, X)\right] = h(\omega) \Longrightarrow \mathbb{E}\left[H(\omega, X) + U\right] = h(\omega) \quad \text{where} \quad \mathbb{E}[U] = 0.$$

- Choose U **correlated with the** *natural oracle* $H(\omega, X)$ s.t.

$$\mathrm{Var}\left(H(\omega, X) + U\right) << \mathrm{Var}\left(H(\omega, X)\right)$$

The SPIDER control variate when h is a finite sum

Adapted from the gradient case: **S**tochastic **P**ath-**I**ntegrated **D**ifferential **E**stimato**R**

For problems of the form

$$\omega : \qquad h(\omega) = 0 \qquad \text{when} \quad h(\omega) = \frac{1}{n}\sum_{i=1}^{n} h_i(\omega) \qquad \text{and } n \text{ large}$$

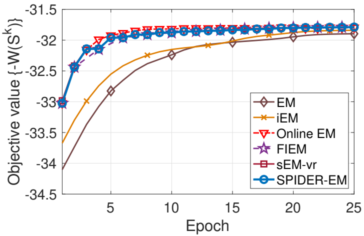
- At iteration $\#k$, a natural oracle for $h(\omega_k)$ is

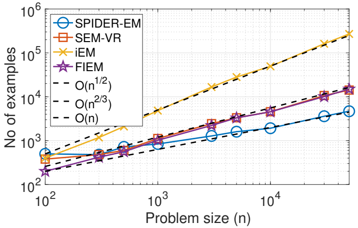
$$H(\omega_k, X_{k+1}) := \frac{1}{\mathsf{b}} \sum_{i \in X_{k+1}} h_i(\omega_k) \qquad X_{k+1} \text{mini-batch from } \{1, \ldots, n\}, \text{ of size b}$$

- The SPIDER oracle is

$$H_{k+1}^{\mathrm{sp}} := \frac{1}{\mathsf{b}} \sum_{i \in X_{k+1}} h_i(\omega_k) + \underbrace{H_k^{\mathrm{sp}}}_{\substack{\text{oracle} \\ \text{for } h(\omega_{k-1})}} - \underbrace{\frac{1}{\mathsf{b}} \sum_{i \in X_{k+1}} h_i(\omega_{k-1})}_{\substack{\text{oracle} \\ \text{for } h(\omega_{k-1})}}$$

Efficiency ... via plots (here)





Conclusion

- A unifying framework for SA, that covers gradient SA, non-gradient SA, possibly with biased oracles is introduced.

- Explicit controls of convergence in expectation are provided.

- From which are deduced: stopping rules strategies, constant step sizes strategies, rates of convergence.

Stochastic Approximation Beyond Gradient for Signal Processing and Machine Learning
by A. Dieuleveut, G. Fort, E. Moulines and H.-T. Wai
HAL-03979922 and arXiv:2302.11147