Stochastic Approximation Beyond Gradient

Gersende Fort

CNRS
Institut de Mathématiques de Toulouse
France

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

Publications:

Stochastic Approximation Beyond Gradient for Signal Processing and Machine Learning

HAL-03979922 arXiv:2302.11147 IEEE Trans. on Signal Processing, 2023

A Stochastic Path Integrated Differential Estimator Expectation Maximization Algorithm

HAl-03029700 NeurIPS, 2020

Partly funded by

Fondation Simone et Cino Del Duca, Project OpSiMorE

Outline

1. Stochastic Approximation: the algorithm and the Lyapunov framework

   Stochastic Approximation:
   an iterative stochastic algorithm, for finding zeros of a vector field.

2. Examples of SA: stochastic gradient and beyond

   Stochastic Gradient is an example of SA, but SA encompasses broader scenarios

3. Non-asymptotic analysis

   best strategy after T iterations, complexity analysis

4. Variance reduction

5. Conclusion

Stochastic Approximation

Stochastic Approximation: a root-finding method

Robbins and Monro (1951) **Wolfowitz (1952), Kiefer and Wolfowitz (1952), Blum (1954), Dvoretzky (1956)**

Problem:

Given a vector field $h : \mathbb{R}^d \to \mathbb{R}^d$, solve

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

Available: for all ω, stochastic oracles of $h(\omega)$.

The *Stochastic Approximation* **method**:

Choose: a sequence of positive step sizes $\{\gamma_k\}_k$ and an initial value $\omega_0 \in \mathbb{R}^d$.
Repeat:

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

where $H(\omega_k, X_{k+1})$ is a stochastic oracle of $h(\omega_k)$.

Rmk: here, the field h is defined on $\mathbb{R}^d$; and for all $\omega \in \mathbb{R}^d$.

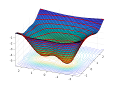
Example: $h(\omega)$ is an expectation; $H(\omega, X_{k+1})$ is a Monte Carlo approximation.

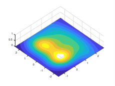
Stochastic Approximation: root-finding method in a Lyapunov setting

SA: $\omega_{k+1} = \omega_k + \gamma_{k+1}\, H(\omega_k, X_{k+1})$ **with an oracle** $H(\omega_k, X_{k+1}) \approx h(\omega_k)$

A Lyapunov function. $V : \mathbb{R}^d \to \mathbb{R}_{>0}$, C^1 and inf-compact s.t.

$$\langle \nabla V(\omega), h(\omega) \rangle \leq 0$$



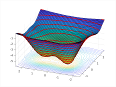
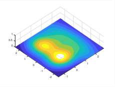


Stochastic Approximation: root-finding method in a Lyapunov setting

SA: $\omega_{k+1} = \omega_k + \gamma_{k+1} H(\omega_k, X_{k+1})$ **with an oracle** $H(\omega_k, X_{k+1}) \approx h(\omega_k)$

A Lyapunov function. $V : \mathbb{R}^d \to \mathbb{R}_{>0}$, C^1 and inf-compact s.t.

$$\langle \nabla V(\omega), h(\omega) \rangle \leq 0$$

- **Key property**

> A Robbins-Siegmund type inequality — Robbins and Siegmund (1971)
>
> $$\mathbb{E}\left[V(\omega_{k+1}) | \text{past}_k\right] \leq V(\omega_k) + \gamma_{k+1} \langle \nabla V(\omega_k), h(\omega_k) \rangle + \gamma_{k+1} \rho_k$$

ρ_k depends on the conditional bias and conditional L^2-moment of the oracles.

- The Lyapunov fct is **not monotone** along the random path $\{\omega_k, k \geq 0\}$
- Key property for the (a.s.) boundedness of the random path, and its convergence.

- SA is an *optimization* method for the minimization of V

... but, converges to $\{\langle \nabla V(\cdot), h(\cdot) \rangle = 0\}$.

Examples of SA: Stochastic Gradient and beyond

Stochastic Gradient is a SA method

Find a root of h: $\omega_{k+1} = \omega_k + \gamma_{k+1}\, H(\omega_k, X_{k+1})$ **where** $H(\omega_k, X_{k+1}) \approx h(\omega_k)$

SG is a root finding algorithm

- designed to solve $\nabla R(\omega) = 0$
- for convex and **non-convex** optimization.

SG is a SA algorithm

$$\omega_{k+1} = \omega_k - \gamma_{k+1}\, \widehat{\nabla R(\omega_k)}$$

see e.g. survey by Bottou (2003, 2010); Lan (2020). Non-convex case: Bottou et al (2018); Ghadimi and Lan (2013)

Empirical Risk Minimization for batch data $R(\omega) = \frac{1}{n}\sum_{i=1}^{n} \ell(\omega, Z_i)$

Vector field: $$h(\omega) = -\frac{1}{n}\sum_{i=1}^{n} \nabla_\omega \ell(\omega, Z_i)$$

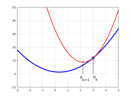
Oracle: $$H(\omega, X_{k+1}) = -\frac{1}{\mathsf{b}}\sum_{i \in X_{k+1}} \nabla_\omega \ell(\omega, Z_i);$$ X_{k+1} is a random mini-batch, cardinal b.

Unbiased oracles: $\mathbb{E}[H(\omega, X_{k+1})] = h(\omega)$

Majorization-Minimization algorithms, with structured majorizing functions

Expectation-Maximization, for curved exponential family — Dempster et al (1977)

- `SAEM`, SA with biased or unbiased oracles — Delyon et al (1999)
- `Mini-batch EM`, SA with unbiased oracles — adapted from Online EM - Cappé and Moulines (2009)



MM algorithms for the minimization of $F : \mathbb{R}^p \to \mathbb{R}$

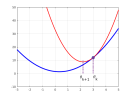
$$F(\cdot) \leq \mathcal{Q}(\cdot, \tau), \qquad \forall \tau, \qquad\qquad F(\tau) = \mathcal{Q}(\tau, \tau)$$

Structured majorizing fcts: parametric family, $\mathcal{Q}(\cdot, \tau) = \langle \mathbb{E}\left[\mathsf{S}(X, \tau)\right], \phi(\cdot) \rangle$

Majorization-Minimization algorithms, with structured majorizing functions

Expectation-Maximization, for curved exponential family — Dempster et al (1977)

- `SAEM`, SA with biased or unbiased oracles — Delyon et al (1999)
- `Mini-batch EM`, SA with unbiased oracles — adapted from Online EM - Cappé and Moulines (2009)

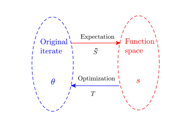


MM algorithms for the minimization of $F : \mathbb{R}^p \to \mathbb{R}$

$$F(\cdot) \leq \mathcal{Q}(\cdot, \tau), \qquad \forall \tau, \qquad F(\tau) = \mathcal{Q}(\tau,\tau)$$

Structured majorizing fcts: parametric family, $\mathcal{Q}(\cdot,\tau) = \langle \mathbb{E}\left[\mathsf{S}(X,\tau)\right], \phi(\cdot)\rangle$

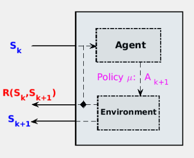
$$w_k \xrightarrow{\text{Minimize}} \mathsf{T}(w_k) := \mathrm{argmin}_\theta \, \langle w_k, \phi(\theta)\rangle$$
$$\xrightarrow{\text{Majorize}} w_{k+1} := \mathbb{E}\left[\mathsf{S}(X, \mathsf{T}(w_k))\right]$$



- A root-finding algorithm: $\mathbb{E}\left[\mathsf{S}(X, \mathsf{T}(\omega))\right] - \omega = 0$
- Oracles = Monte Carlo approximations of the intractable expectation

Value function in a Reward Markov process via Bellman equation

Value function in a Reward Markov process:



- Markov process $(s_t)_t$ with stationary distribution π
- taking values in $\mathcal{S}$, $\quad \mathrm{Card}(\mathcal{S}) = n$.
- Reward $\mathsf{R}(s, s')$
- Value function: $\qquad \lambda \in (0,1)$

$$\forall\ s \in \mathcal{S}, \qquad V_\star(s) := \sum_{t \geq 0} \lambda^t\, \mathbb{E}\left[\mathsf{R}(S_t, S_{t+1}) \middle| S_0 = s\right].$$

The Bellman equation $\qquad \mathsf{B}[V] - V = 0$

$$\mathbb{E}\left[\,\mathsf{R}(S_0, S_1) + \lambda V(S_1)\,|\,S_0 = s\,\right] - V(s) = 0, \qquad \forall s \in \mathcal{S}$$

Algorithm TD(0): with linear fct approximation: $V^\omega := \Phi\omega = \omega_1\Phi_1(\cdot) + \cdots + \omega_d\Phi_d(\cdot)$

TD(0) is a SA — Sutton (1987); Tsitsiklis and Van Roy (1997)

with mean field $\qquad h(\omega) := \Phi'\,\mathrm{diag}(\pi)\ (\mathsf{B}[\Phi\omega] - \Phi\omega)$

Oracle: $\qquad H(\omega, (S_k, S_{k+1}, R(S_k, S_{k+1}))) := \left(\mathsf{R}(S_k, S_{k+1}) + \lambda V^\omega(S_{k+1}) - V^\omega(S_k)\right)\ (\Phi_{S_k,:})'$

SA beyond the gradient case

Understanding the behavior of SA algorithms and designing improved algorithms require new insights that depart from the study of *traditional SG* algorithms.

What is the "gradient case" ?

- the mean field h is a gradient: $h(\omega) = -\nabla R(\omega)$
- the oracle is unbiased: $\mathbb{E}\left[H(\omega, X)\right] = h(\omega)$

Non-asymptotic analysis

Analyses

- Asymptotic convergence analysis, when the horizon tends to infinity

Benveniste et al (1987/2012), Benaïm (1999), Kushner and Yin (2003), Borkar (2009)

  - almost-sure convergence of the sequence $\{\omega_k, k \geq 0\}$
  - to (a connected component of) the set $\mathcal{L} := \{\omega : \langle \nabla V(\omega), h(\omega) \rangle = 0\}$
  - CLT, $\cdots$

- Non-asymptotic analysis

Given a total number of iterations T

  - After T calls to an oracle, what can be obtained ?

ϵ-approximate stationary point and sample complexity

  - How many iterations to reach an ϵ-approximate stationary point

$$\forall \epsilon > 0, \quad \mathbb{E}\left[W(\omega_\bullet)\right] \leq \epsilon$$

The assumptions

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

Lyapunov function V and control W

There exist $V : \mathbb{R}^d \to [0, +\infty)$, $W : \mathbb{R}^d \to [0, +\infty)$ and positive constants s.t.

- V and W: $\forall \omega \quad \langle \nabla V(\omega), h(\omega) \rangle \leq -\rho\, W(\omega)$
- V smooth $\forall \omega, \omega' \quad \|\nabla V(\omega) - \nabla V(\omega')\| \leq L_V \|\omega - \omega'\|$

		$h(\omega)$	$V(\omega)$	$W(\omega)$
Gradient case		$-\nabla R(\omega)$	$R(\omega)$	$\|h(\omega)\|^2$
and R convex	$\omega_\star$ solution	$-\nabla R(\omega)$	$0.5\|\omega - \omega_\star\|^2$	$-\langle \omega - \omega_\star, h(\omega) \rangle$
and R strongly cvx	$\omega_\star$ solution	$-\nabla R(\omega)$	$0.5\|\omega - \omega_\star\|^2$	$W = V$ or, as above
Stochastic EM		$\bar{s}(\mathsf{T}(\omega)) - \omega$	$F(\mathsf{T}(\omega))$	$\|h(\omega)\|^2$
TD(0)	$\Phi\omega_\star$ solution	$\Phi' D(\mathsf{B}\Phi\omega - \Phi\omega)$	$0.5\|\omega - \omega_\star\|^2$	$(\omega - \omega_\star)'\Phi' D \Phi(\omega - \omega_\star)$

The assumptions

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

On the oracles and the mean field

There exist non-negative constants s.t.

- The mean field $\quad \forall \omega \quad \|h(\omega)\|^2 \leq c_0 + c_1 W(\omega)$

for all k, almost-surely,

- Bias $\quad \|\mathbb{E}\left[H(\omega_k, X_{k+1}) \middle| \mathcal{F}_k\right] - h(\omega_k)\|^2 \leq \tau_0 + \tau_1 W(\omega_k)$
- Variance $\quad \mathbb{E}\left[\|H(\omega_k, X_{k+1}) - \mathbb{E}\left[H(\omega_k, X_{k+1}) \middle| \mathcal{F}_k\right]\|^2 \middle| \mathcal{F}_k\right] \leq \sigma_0^2 + \sigma_1^2 W(\omega_k)$

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

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

Includes cases:

- Biased oracles, unbiased oracles
- Bounded variance of the oracles, unbounded variance of the oracles

A non-asymptotic convergence bound in expectation

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

Assume also that $\gamma_k \in (0, \gamma_{\max})$, $\quad \eta_1 \geq \sigma_1^2 + c_1 > 0$

$$\gamma_{\max} := \frac{2(\rho - \mathtt{b}_1)}{L_V \, \eta_1}$$

Then, there exist non-negative constants s.t. for any $T \geq 1$

$$\sum_{k=1}^{T} \frac{\gamma_k \mu_k}{\sum_{\ell=1}^{T} \gamma_\ell \mu_\ell} \, \mathbb{E}\left[W(\omega_{k-1})\right] \leq 2 \, \frac{\mathbb{E}\left[V(\omega_0)\right]}{\sum_{\ell=1}^{T} \gamma_\ell \mu_\ell} + L_V \, \eta_0 \, \frac{\sum_{k=1}^{T} \gamma_k^2}{\sum_{\ell=1}^{T} \gamma_\ell \mu_\ell} + c_V \sqrt{\tau_0} \, \frac{\sum_{k=1}^{T} \gamma_k}{\sum_{\ell=1}^{T} \gamma_\ell \mu_\ell}$$

$\mu_\ell = 2(\rho - \mathtt{b}_1) - \gamma_\ell L_V \eta_1 > 0$

- η_ℓ depends on the bias and variance of the oracles; $\eta_0 > 0$.
- For unbiased oracles: $\tau_0 = \mathtt{b}_1 = 0$
- Better bounds when $V = W$; not discussed here $\quad$ ex.: SGD for strongly cvx fct; TD(0)

After T iterations

The strategy

- Choose a constant stepsize $\gamma_k = \gamma := \frac{\gamma_{\max}}{2} \wedge \frac{\sqrt{2\mathbb{E}[V(\omega_0)]}}{\sqrt{\eta_0 L_V}\sqrt{T}}$
- Random stopping: return $\omega_{\mathcal{R}_T}$ where $\mathcal{R}_T \sim \mathcal{U}(\{0, \cdots, T-1\})$
  or when W is convex: return the averaged iterate $T^{-1}\sum_{k=0}^{T-1}\omega_k$

yields

$$\mathbb{E}\left[W(\omega_{\mathcal{R}_T})\right] \leq \frac{2\sqrt{2L_V\eta_0}\sqrt{\mathbb{E}\left[V(\omega_0)\right]}}{(\rho - b_1)\sqrt{T}} \vee \frac{8\mathbb{E}\left[V(\omega_0)\right]}{\gamma_{\max}(\rho - b_1)T} + c_V\frac{\sqrt{\tau_0}}{\rho - b_1}$$

When $\tau_0 = 0$ i.e. unbiased oracles, or bias scaling with W, it is an *optimal* control in expectation.

When $\tau_0 > 0$:
- the term can not be made small with constant step size
- ad-hoc strategies: play with "design parameters" to make this term small.

ϵ-approximate stationary point, for unbiased oracles

For all $\epsilon > 0$, let $\mathcal{T}(\epsilon) \subset \mathbb{N}$ s.t. for all $T \in \mathcal{T}(\epsilon)$, $\quad \mathbb{E}\left[W(\omega_{\mathcal{R}_T})\right] \leq \epsilon$.

For unbiased oracles,

$\mathcal{T}(\epsilon) = [T_\epsilon, +\infty)$ with

$$T_\epsilon := 8\,\mathbb{E}[V(\omega_0)]\,\frac{\eta_0 L_V}{\rho^2}\left(\frac{1}{\epsilon^2} \vee \frac{\eta_1}{2\eta_0 \epsilon}\right)$$

- Low precision regime: $\epsilon > 2\eta_0/\eta_1$,

$$T_\epsilon = 4\,\mathbb{E}[V(\omega_0)]\,\frac{\eta_1 L_V}{\rho^2\,\epsilon}, \qquad \gamma = \frac{\gamma_{\max}}{2}$$

- High precision regime: $\epsilon \in (0, 2\eta_0/\eta_1]$,

$$T_\epsilon = 8\,\mathbb{E}[V(\omega_0)]\,\frac{\eta_0 L_V}{\rho^2\,\epsilon^2}, \qquad \gamma = \frac{\rho\,\epsilon}{2\eta_0 L_V}$$

ϵ-approximate stationary point, when biased oracles: on an example

EM $\quad h(\omega) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathsf{S}}_i(\mathsf{T}(\omega)) - \omega \quad$ where $\quad \bar{\mathsf{S}}_i(\tau) := \int_{\mathcal{X}} S_i(x)\pi(x;\tau)\mathrm{d}x$

The SA-EM oracle

- Monte Carlo sum with m points,
- case "Self-normalized Importance Sampling": bias β_0/m and variance β_1/m.

Complexity

For all $\epsilon > 0$, let $\mathcal{T}(\epsilon) \subset \mathbb{N}^2$ s.t. for all $(T, m) \in \mathcal{T}(\epsilon)$, $\quad \mathbb{E}\left[W(\omega_{\mathcal{R}_T})\right] \leq \epsilon$.

$$T \geq \frac{16\mathbb{E}[V(\omega_0)](1+\bar{\sigma_1}^2/m)}{v_{\min}^2\kappa\epsilon} \vee \frac{32\mathbb{E}[V(\omega_0)]\bar{\sigma}_0^2 L_V}{m v_{\min}^2\kappa^2\epsilon^2} \qquad m \geq \frac{4c_b}{(1-\kappa)v_{\min}\epsilon}$$

For low precision regime,

$$T_\epsilon = \frac{C_1}{\epsilon}, \qquad m_\epsilon = \frac{C_2}{\epsilon}, \qquad \mathrm{cost}_{\mathrm{comp}} = T_\epsilon\left(n m_\epsilon \,\mathrm{cost}_{\mathrm{MC}} + \mathrm{cost}_{\mathrm{opt}}\right)$$

Other rates for low precision regime.

Variance Reduction within SA

Control variates for variance reduction

- Add a random variable to the *natural oracle* $H(\omega, X)$

- *Control variates* U, classical in Monte Carlo:

$$\mathbb{E}\left[H(\omega, X) + U\right] = \mathbb{E}\left[H(\omega, X)\right] \qquad \mathrm{Var}\left(H(\omega, X) + U\right) < \mathrm{Var}\left(H(\omega, X)\right).$$

Introduced in Stochastic Gradient, in the case *finite sum*

$$h(\omega) = \frac{1}{n}\sum_{i=1}^{n} h_i(\omega)$$

Extended to SA

Survey on Variance Reduction in ML: Gower et al (2020)
Gradient case: Johnson and Zhang (2013), Defazio et al (2014), Nguyen et al (2017), Fang et al (2018), Wang et al (2018), Shang et al (2020)
Riemannian non-convex optimization: Han and Gao (2022)
Mirror Descent: Luo et al (2022)
Stochastic EM: Chen et al (2018), Karimi et al (2019), Fort et al. (2020, 2021), Fort and Moulines (2021,2023)

The SPIDER control variate when h is a finite sum

Adapted from the gradient case: Stochastic Path-Integrated Differential EstimatoR

Nguyen et al (2017), Fang et al (2018), Wang et al (2019)

In the finite sum setting: $h(\omega) = \frac{1}{n}\sum_{i=1}^{n} h_i(\omega)$ and n large

- At iteration $\#(k+1)$, a natural oracle for $h(\omega_k)$ is

$$H(\omega_k, X_{k+1}) := \frac{1}{\mathsf{b}} \sum_{i \in X_{k+1}} h_i(\omega_k)$$

X_{k+1} mini-batch from $\{1, \ldots, n\}$, of size b

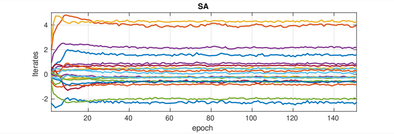
- The SPIDER oracle is

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

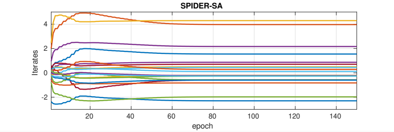
- Implementation: *refresh* the control variate every K_{in} iterations

Efficiency ... via plots (here)

Application: Stochastic EM with ctt step size, mixture of twelve Gaussian in $\mathbb{R}^{20}$; unknown weights, means and covariances.

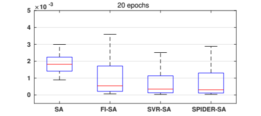


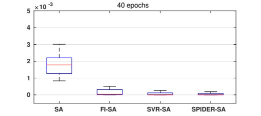
Estimation of 20 parameters, one path of SA



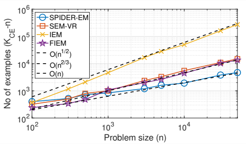
Estimation of 20 parameters, one path of SPIDER-SA

Squared norm of the mean field h, after 20 and 40 epochs; for SA and three variance reduction methods





Application: Stochastic EM with ctt step size, mixture of two Gaussian in $\mathbb{R}$, unknown means.



For a fixed accuracy level, for different values of the problem size n, display the number of examples processed to reach the accuracy level (mean nbr over 50 indep runs).

Conclusion

Conclusion

- SA methods with non-gradient mean field and/or biased oracles - in ML and compurational statistics.
- A non-asymptotic analysis for *general Stochastic Approximation schemes*
- For *finite sum field* h*:* variance reduction within SA via control variates.

- Oracles, from *Markovian* examples
- Roots of $h = 0$, on $\Omega \subset \mathbb{R}^d$

- Federated SA: compression, control variateS, partial participation, heterogeneity, local iterations, . . .