Stochastic Approximation Beyond Gradient

Gersende Fort
CNRS
Institut de Mathématiques de Toulouse

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

ANR AAPG-2019, Project MASDOL

Outline

- Stochastic Approximation
- Examples of SA: stochastic gradient and beyond

  Stochastic Gradient is an example of SA, but SA encompasses broader scenarios (compressed stochastic gradient; Reinforcement Learning via TD learning; Computational Statistics via EM)

  Understanding the behavior of these algorithms and designing improved algorithms require new insights that depart from the study of traditional SG algorithms.
- Non-asymptotic analysis

  best strategy after T iterations, complexity analysis
- Variance Reduction for SA

  Improved SA schemes.
- Conclusion

Stochastic Approximation

Stochastic Approximation: is a root-finding method

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

Problem:

Given a mean 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 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 $h(\omega_k)$.

Stochastic Approximation: the intuition

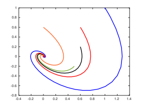
SA: $\omega_{k+1} = \omega_k + \gamma_{k+1} H(\omega_k, X_{k+1})$ **with an oracle** $H(\omega_k, X_{k+1}) \approx h(\omega_k)$

ODE with vector field h

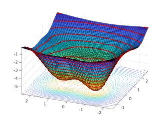
- A function $t \in [0, +\infty) \mapsto \overline{w}_t \in \mathbb{R}^d$ s.t.

$$\overline{w}_0 = \omega_0, \qquad \frac{d\overline{w}_t}{dt} = h(\overline{w}_t).$$

- A fixed point $\omega^\star$ is a root of h.
- Under assumptions (Lyapunov), $\lim_t \mathrm{dist}(\overline{w}_t, \mathcal{L}) = 0$.
- $\{h = 0\} \subseteq \mathcal{L}$.



$d = 2$. For five initial values ω_0, the solution $t \mapsto \overline{w}_t$.



A Lyapunov function for h

- $V : \mathbb{R}^d \to [0, +\infty)$, continuously differentiable, and inf-compact.
- $t \mapsto V(\overline{w}_t)$ decreasing i.e. $\langle \nabla V(\overline{w}_t), h(\overline{w}_t) \rangle \leq 0$

Stochastic Approximation: the intuition

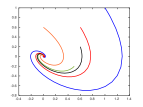
SA: $\omega_{k+1} = \omega_k + \gamma_{k+1} H(\omega_k, X_{k+1})$ **with an oracle** $H(\omega_k, X_{k+1}) \approx h(\omega_k)$

ODE with vector field h

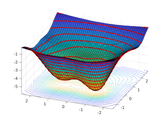
- A function $t \in [0, +\infty) \mapsto \overline{w}_t \in \mathbb{R}^d$ s.t.

$$\overline{w}_0 = \omega_0, \qquad \frac{d\overline{w}_t}{dt} = h(\overline{w}_t).$$

- A fixed point $\omega^\star$ is a root of h.
- Under assumptions (Lyapunov), $\lim_t \mathrm{dist}(\overline{w}_t, \mathcal{L}) = 0$.
- $\{h = 0\} \subseteq \mathcal{L}$.



$d = 2$. For five initial values ω_0, the solution $t \mapsto \overline{w}_t$.



A Lyapunov function for h

- $V : \mathbb{R}^d \to [0, +\infty)$, continuously differentiable, and inf-compact.
- $t \mapsto V(\overline{w}_t)$ decreasing i.e. $\langle \nabla V(\overline{w}_t), h(\overline{w}_t) \rangle \leq 0$

SA is an approximation ($\times 2$): Euler and oracle

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

Stochastic Approximation: the step sizes and the oracles

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

- $\gamma_k > 0$
- $\sum_k \gamma_k = +\infty$

- The oracles can be unbiased $\mathbb{E}\left[H(\omega_k, X_{k+1})|\mathrm{past}_k\right] = h(\omega_k)$
  or biased $\mathbb{E}\left[H(\omega_k, X_{k+1})|\mathrm{past}_k\right] \neq h(\omega_k)$

- $\lim_K \sum_{k=0}^K \gamma_k \, (H(\omega_k, X_{k+1}) - h(\omega_k))$ exists (wp1)

  unbiased case with bounded variance: $\sum_k \gamma_k^2 < \infty$

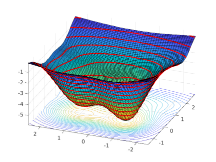
- $\lim_k \gamma_k = 0$

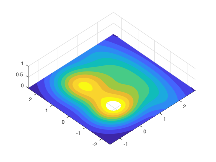
Stochastic Approximation: stability and convergence via a Lyapunov function

Lyapunov for ODE

- $t \mapsto V(\overline{w}_t)$ decreasing i.e.

$$\langle \nabla V(\overline{w}_t), h(\overline{w}_t) \rangle \leq 0$$





Lyapunov for the theory of SA

- The Lyapunov fct **is not monotone** along the random path $\{\omega_k, k \geq 0\}$

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

A Robbins-Siegmund type inequality — Robbins and Siegmund (1971)

$$\mathbb{E}\left[V(\omega_{k+1}) | \mathrm{past}_k\right] \leq V(\omega_k) + \gamma_{k+1} \, \langle \nabla V(\omega_k), h(\omega_k) \rangle + \gamma_{k+1} \, \rho_k$$

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

- For the (a.s.) boundedness of the random path, and its convergence.

Stochastic Approximation and Stochastic Optimization

- Stochastic Approximation for solving:

$$\nabla R(\omega) = 0 \qquad \text{when} \qquad \nabla R(\omega) = \mathbb{E}\left[H(\omega, X)\right]$$

- Stochastic Approximation for solving a fixed point equation:

$$\mathsf{T}(\omega) = \omega \qquad \text{when} \qquad \mathsf{T}(\omega) = \mathbb{E}\left[\tilde{H}(\omega, X)\right]$$

- The Lyapunov function assumption: minimizes V through steps in the directions given by the vector field h

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

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 $\nabla R(\omega) = 0$

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)

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 $\nabla R(\omega) = 0$

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

$$R(\omega) = \frac{1}{n}\sum_{i=1}^{n} \ell(\omega, Z_i) \qquad h(\omega) = -\frac{1}{n}\sum_{i=1}^{n} \mathsf{D}_{10}\ell(\omega, Z_i)$$

$$H(\omega, X_{k+1}) = -\frac{1}{\mathsf{b}}\sum_{i \in X_{k+1}} \mathsf{D}_{10}\ell(\omega, Z_i)$$

X_{k+1} a random subset of $\{1, \ldots, n\}$, cardinal b.

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 $\nabla R(\omega) = 0$

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

$$R(\omega) = \frac{1}{n}\sum_{i=1}^{n} \ell(\omega, Z_i) \qquad h(\omega) = -\frac{1}{n}\sum_{i=1}^{n} \mathsf{D}_{10}\ell(\omega, Z_i)$$

$$H(\omega, X_{k+1}) = -\frac{1}{\mathsf{b}}\sum_{i \in X_{k+1}} \mathsf{D}_{10}\ell(\omega, Z_i)$$

X_{k+1} a random subset of $\{1, \ldots, n\}$, cardinal b.

SG is a SA algorithm with goal: optimization

- for convex and **non-convex** optimization.

SA beyond the gradient case

The "gradient case":

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

SA beyond the gradient case: two examples.

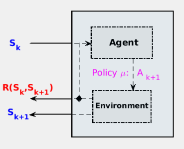
Policy evaluation of a Markov Reward Process

by a Temporal Difference (TD) method with linear function approximation

A Markov Reward Process:

- State $s \in \mathcal{S}$, $\quad \mathrm{Card}(\mathcal{S}) = n$.
- Markov process: transition matrix P, $\pi P = \pi$
- Reward $\mathsf{R}(s, s')$ P, π and R depend on the policy μ
- Value function: $\lambda \in (0, 1)$

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



▶ The value function evaluation is a root-finding problem

Bellman equation: $\quad \mathsf{B}V_\star - V_\star = 0 \qquad \mathsf{B}V(s) := \mathbb{E}\left[\mathsf{R}(S_0, S_1) + \lambda V(S_1) | S_0 = s\right]$

Linear Function Approximation: $\quad V^\omega \in \mathrm{Span}(\phi_1, \cdots, \phi_d)$

find V^ω $\Leftrightarrow$ find $\Phi\omega$ $\Leftrightarrow$ find $\omega \in \mathbb{R}^d$

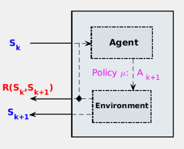
Policy evaluation of a Markov Reward Process

by a Temporal Difference (TD) method with linear function approximation

A Markov Reward Process:

- State $s \in \mathcal{S}$, $\mathrm{Card}(\mathcal{S}) = n$.
- Markov process: transition matrix P, $\pi P = \pi$
- Reward $\mathsf{R}(s, s')$ P, π and R depend on the policy μ
- Value function: $\lambda \in (0,1)$

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



▶ The value function evaluation is a root-finding problem

Bellman equation: $\mathsf{B}V_\star - V_\star = 0$ $\qquad \mathsf{B}V(s) := \mathbb{E}\left[\mathsf{R}(S_0, S_1) + \lambda V(S_1) | S_0 = s\right]$

Linear Function Approximation: $V^\omega \in \mathrm{Span}(\phi_1, \cdots, \phi_d)$

find V^ω $\Leftrightarrow$ find $\Phi\omega$ $\Leftrightarrow$ find $\omega \in \mathbb{R}^d$

▶ TD(0) with linear function approximation is SA $\qquad$ Sutton (1987); Tsitsiklis and Van Roy (1997)

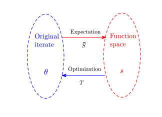
TD(0) is a SA with mean field $\qquad h(\omega) := \Phi' \, \mathrm{diag}(\pi) \ (\mathsf{B}\Phi\omega - \Phi\omega)$

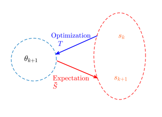
Oracle: $\qquad H(\omega, (S_k, S_{k+1}, R(S_k, S_{k+1}))) := \left(\mathsf{R}(S_k, S_{k+1}) + \lambda[\Phi\omega]_{S_{k+1}} - [\Phi\omega]_{S_k}\right) (\Phi_{S_k,:})'$

Stochastic Expectation-Maximization

In the curved exponential family **Dempster et al (1977)**

$$\text{argmin}_\theta - \log \int_{\mathcal{X}} p(x;\theta)\,\nu(\mathsf{d}x) \qquad p(x;\theta) > 0$$





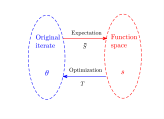
- EM is a root-finding algorithm
  - EM is a Majorize-Minimization algorithm
  - The majorizing function defined by $\int_{\mathcal{X}} S(x)\pi(x;\theta_k)\nu(\mathsf{d}x)$
  - Fixed points of EM: Delyon et al (1999)

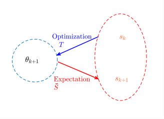
$$\theta_\star := \mathsf{T}(s_\star) \quad \text{with} \quad s_\star \text{ s.t. } \bar{\mathsf{S}}\left(\mathsf{T}\left(s_\star\right)\right) - s_\star = 0$$

Stochastic Expectation-Maximization

In the curved exponential family **Dempster et al (1977)**

$$\mathrm{argmin}_\theta - \log \int_{\mathcal{X}} p(x;\theta)\, \nu(\mathrm{d}x) \qquad p(x;\theta) > 0$$





- EM is a root-finding algorithm
  - EM is a Majorize-Minimization algorithm
  - The majorizing function defined by $\int_{\mathcal{X}} S(x)\pi(x;\theta_k)\nu(\mathrm{d}x)$
  - Fixed points of EM: Delyon et al (1999)

$$\theta_\star := \mathsf{T}(s_\star) \quad \text{with} \quad s_\star \text{ s.t. } \bar{\mathsf{S}}\left(\mathsf{T}\left(s_\star\right)\right) - s_\star = 0$$

- When $\bar{\mathsf{S}}$ intractable, the most popular/efficient Stochastic EM is SA

$$\bar{\mathsf{S}}(\cdot) := \int_{\mathcal{X}} S(x)\pi(x;\cdot)\,\nu(\mathrm{d}x) \qquad \text{or (and)} \qquad \bar{\mathsf{S}}(\cdot) := \frac{1}{n}\sum_{i=1}^{n} \bar{\mathsf{S}}_i(\cdot),$$

Stochastic EM is a SA with mean field $h(\omega) := \bar{\mathsf{S}}(\mathsf{T}(\omega)) - \omega$

[U,B] Oracle for SAEM: $H(\omega, X_{k+1}) := m^{-1}\sum_{\ell=1}^{m} S(X_{k+1,\ell}) - \omega \quad X_{k+1,\cdot} \sim$ MCMC $\pi(\cdot;\mathsf{T}(\omega))$

[U] Oracle for mini-batch EM: $H(\omega, X_{k+1}) := \mathsf{b}^{-1}\sum_{i\in X_{k+1}} \bar{\mathsf{S}}_i(\mathsf{T}(\omega)) - \omega$

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

for all k, almost-surely,

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

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

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

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 ex.: SGD for strongly cvx fct; TD(0)

After T iterations

- Reached with a constant step size

$$\gamma_k = \gamma := \frac{\gamma_{\max}}{2} \wedge \frac{\sqrt{2\mathbb{E}[V(\omega_0)]}}{\sqrt{\eta_0 L_V}\sqrt{T}}$$

$$\underbrace{\frac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}\left[W(\omega_k)\right]}_{\mathbb{E}\left[W(\omega_{\mathcal{R}_T})\right]} \leq \frac{2\sqrt{2L_V\eta_0}\sqrt{\mathbb{E}\left[V(\omega_0)\right]}}{(\rho - b_1)\sqrt{T}} \vee \frac{8\mathbb{E}\left[V(\omega_0)\right]}{\gamma_{\max}(\rho - b_1)T} + c_V\frac{\sqrt{\tau_0}}{\rho - b_1}$$

When $\tau_0 = 0$ i.e. unbiased oracles, or bias scaling with W

- Random stopping: return $\omega_{\mathcal{R}_T}$ where $\mathcal{R}_T \sim \mathcal{U}(\{0, \cdots, T-1\})$
- When W is convex: return the *Polyak-Ruppert-Juditsky* averaged iterate $T^{-1}\sum_{k=0}^{T-1}\omega_k$
- Upper bound depending on T: $\propto 1/\sqrt{T}$

ϵ-approximate stationary point, for unbiased oracles

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

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 $\quad h(\omega) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathsf{S}}_i(\mathsf{T}(\omega)) - \omega$ 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: biased oracles, with 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 $(T, m) \in \mathcal{T}(\epsilon)$, $\qquad \mathbb{E}\left[W(\omega_{\mathcal{R}_T})\right] \leq \epsilon$.

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

For high 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.

Sketch of proof of the Theorem

A Lyapunov function V with L_V-Lipschitz gradient

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

Sketch of proof of the Theorem

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

The definition of the iterative scheme

$$V(\omega_{k+1}) \leq V(\omega_k) + \gamma_{k+1} \langle \nabla V(\omega_k), H(\omega_k, X_{k+1}) \rangle + \frac{L_V}{2} \gamma_{k+1}^2 \|H(\omega_k, X_{k+1})\|^2$$

Sketch of proof of the Theorem

$$V(\omega_{k+1}) \leq V(\omega_k) + \gamma_{k+1} \left\langle \nabla V(\omega_k), H(\omega_k, X_{k+1}) \right\rangle + \frac{L_V}{2} \gamma_{k+1}^2 \left\| H(\omega_k, X_{k+1}) \right\|^2$$

The conditional expectation

$$\mathbb{E}\left[V(\omega_{k+1}) | \mathcal{F}_k\right] \leq V(\omega_k) + \gamma_{k+1} \left\langle \nabla V(\omega_k), \mathbb{E}\left[H(\omega_k, X_{k+1}) | \mathcal{F}_k\right] \right\rangle + \frac{L_V}{2} \gamma_{k+1}^2 \, \mathbb{E}\left[\left\| H(\omega_k, X_{k+1}) \right\|^2 | \mathcal{F}_k \right]$$

Sketch of proof of the Theorem

$$\mathbb{E}\left[V(\omega_{k+1})|\mathcal{F}_k\right] \leq V(\omega_k) + \gamma_{k+1} \left\langle \nabla V(\omega_k), \mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right] \right\rangle + \frac{L_V}{2}\gamma_{k+1}^2 \mathbb{E}\left[\|H(\omega_k, X_{k+1})\|^2 |\mathcal{F}_k\right]$$

The mean field h and the bias term

$$\begin{aligned}\mathbb{E}\left[V(\omega_{k+1})|\mathcal{F}_k\right] \leq V(\omega_k) &+ \gamma_{k+1} \left\langle \nabla V(\omega_k), \mathsf{h}(\omega_k) \right\rangle \\ &+ \gamma_{k+1} \left\langle \nabla V(\omega_k), \mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right] - \mathsf{h}(\omega_k) \right\rangle \\ &+ \frac{L_V}{2}\gamma_{k+1}^2 \mathbb{E}\left[\|H(\omega_k, X_{k+1})\|^2 |\mathcal{F}_k\right]\end{aligned}$$

Sketch of proof of the Theorem

$$\begin{aligned}\mathbb{E}\left[V(\omega_{k+1})|\mathcal{F}_k\right] \le V(\omega_k) &+ \gamma_{k+1}\ \langle \nabla V(\omega_k), \mathsf{h}(\omega_k)\rangle \\ &+ \gamma_{k+1}\ \langle \nabla V(\omega_k), \mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right] - \mathsf{h}(\omega_k)\rangle \\ &+ \frac{L_V}{2}\gamma_{k+1}^2\ \mathbb{E}\left[\|H(\omega_k, X_{k+1})\|^2|\mathcal{F}_k\right]\end{aligned}$$

Cond L^2 = Cond Var + (Cond Exp)^2

$$\begin{aligned}\mathbb{E}\left[V(\omega_{k+1})|\mathcal{F}_k\right] \le V(\omega_k) &+ \gamma_{k+1}\ \langle \nabla V(\omega_k), \mathsf{h}(\omega_k)\rangle \\ &+ \gamma_{k+1}\ \langle \nabla V(\omega_k), \mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right] - \mathsf{h}(\omega_k)\rangle \\ &+ \frac{L_V}{2}\gamma_{k+1}^2\ \mathbb{E}\left[\|H(\omega_k, X_{k+1}) - \mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right]\|^2|\mathcal{F}_k\right] \\ &+ \frac{L_V}{2}\gamma_{k+1}^2\ \|\mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right]\|^2\end{aligned}$$

Sketch of proof of the Theorem

$$\begin{aligned}\mathbb{E}\left[V(\omega_{k+1})|\mathcal{F}_k\right] \leq V(\omega_k) &+ \gamma_{k+1} \langle \nabla V(\omega_k), \mathsf{h}(\omega_k)\rangle \\ &+ \gamma_{k+1} \left\langle \nabla V(\omega_k), \mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right] - \mathsf{h}(\omega_k) \right\rangle \\ &+ \frac{L_V}{2}\gamma_{k+1}^2 \, \mathbb{E}\left[\| H(\omega_k, X_{k+1}) - \mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right]\|^2 |\mathcal{F}_k\right] \\ &+ \frac{L_V}{2}\gamma_{k+1}^2 \, \| \mathbb{E}\left[H(\omega_k, X_{k+1})|\mathcal{F}_k\right] - \mathsf{h}(\omega_k) + \mathsf{h}(\omega_k) \|^2\end{aligned}$$

By assumptions: the drift term, the bias and variance of the oracles, and the mean field are controlled by W.
Apply the expectation.

There exist constants s.t. for any $k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[V(\omega_{k+1})\right] \leq \mathbb{E}\left[V(\omega_k)\right] &- \gamma_{k+1} \left(\rho - \mathsf{b}_1 - \gamma_k \frac{L_V \eta_1}{2}\right) \mathbb{E}\left[W(\omega_k)\right] \\ &+ \gamma_{k+1}\mathsf{b}_0 + \gamma_{k+1}^2 \frac{L_V\,\eta_0}{2}\end{aligned}$$

A drift term for γ_k small enough. Sum from $k = 0$ to $k = T-1$; conclude.

Variance Reduction within SA

Control variates for variance reduction

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

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

- Bias

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

- *Control variates* classical in Monte Carlo; introduced in Stochastic Gradient; 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

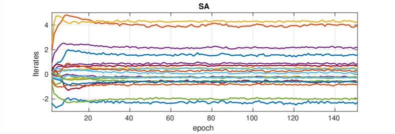
- The SPIDER oracle is

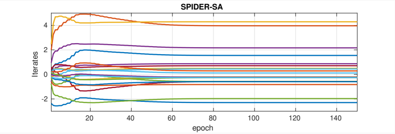
$$H^{\mathrm{sp}}_{k+1} := \frac{1}{\mathsf{b}} \sum_{i \in X_{k+1}} h_i(\omega_k) + \underbrace{H^{\mathrm{sp}}_{k}}_{\text{oracle for } h(\omega_{k-1})} - \underbrace{\frac{1}{\mathsf{b}} \sum_{i \in X_{k+1}} h_i(\omega_{k-1})}_{\text{oracle for } h(\omega_{k-1})}$$

- 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.

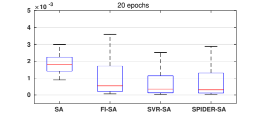


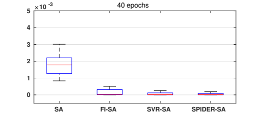


Estimation of 20 parameters, one path of SA

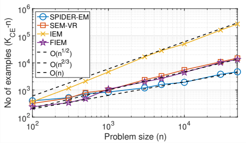
Estimation of 20 parameters, one path of SPIDER-SA

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*, and variance reduction via control variates.

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

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

Compressed Stochastic Gradient

Compression operator $\mathcal{C}$:

- a mapping $x \mapsto \mathcal{C}(x, U)$
- s.t. for any $x \in \mathbb{R}^d$, the cost for storing/transmitting $\mathcal{C}(x, U)$ is lower than the cost for storing/transmitting x.
- examples: projection, quantization
- random or deterministic

Compressed Stochastic Gradient

Compression operator $\mathcal{C}$:

- a mapping $x \mapsto \mathcal{C}(x, U)$
- s.t. for any $x \in \mathbb{R}^d$, the cost for storing/transmitting $\mathcal{C}(x, U)$ is lower than the cost for storing/transmitting x.
- examples: projection, quantization
- random or deterministic

Compression within a Stochastic Gradient step:

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

increasing interest in distributed optimization

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

gradient at a perturbed iterate: Straight-Through Estimator

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

low-precision SG