

Stochastic Approximation: Finite-time analyses and Variance Reduction

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse

In collaboration with

- Aymeric Dieuleveut,
- Eric Moulines,
- Hoi-To Wai,

Ecole Polytechnique, CMAP, France

Ecole Polytechnique, CMAP, France

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

HAI-03029700 NeurIPS, 2020

Partly funded by

Fondation Simone et Cino Del Duca, Project OpSiMorE



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

Examples of SA: Stochastic Gradient and beyond

Non-asymptotic analysis

Variance Reduction within SA

Conclusion

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** $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$, solve

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

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 $\omega \in \mathbb{R}^d$.

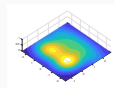
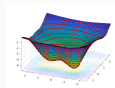
Example: $h(\omega)$ is an expectation; $H(\omega, X_{k+1})$ is a Monte Carlo approximation.

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**. $V : \mathbb{R}^d \rightarrow \mathbb{R}_{>0}$, C^1 and inf-compact s.t.

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

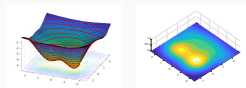


Stochastic Approximation: root-finding method in a Lyapunov setting

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**. $V : \mathbb{R}^d \rightarrow \mathbb{R}_{>0}$, 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)

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

ρ_k depends on the conditional L^2 -moment (bias and variance) 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 Approximation

Examples of SA: Stochastic Gradient and beyond

Non-asymptotic analysis

Variance Reduction within SA

Conclusion

Stochastic Gradient is a SA method (1/2)

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

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

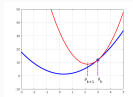
Stochastic Gradient is a SA method (2/2)

		oracles given by the system	oracles built by the user	Biased oracle	Unbiased oracle
SGD	batch online	✓	✓	(✓)	✓ ✓

Batch learning: $\operatorname{argmin}_{\omega} \frac{1}{n} \sum_{i=1}^n \ell(\omega, Z_i)$

Online learning: $\operatorname{argmin}_{\omega} \mathbb{E}[\ell(\omega, Z)]$ from examples $Z_1, Z_2, \dots$

Majorization-Minimization algorithms (Expectation-Maximization algorithms) with structured majorizing functions (1/3)

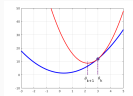


MM algorithms for the minimization of $F : \mathbb{R}^p \rightarrow \mathbb{R}$

$$F(\cdot) \leq Q(\cdot, \tau), \quad \forall \tau, \quad F(\tau) = Q(\tau, \tau)$$

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

Majorization-Minimization algorithms (Expectation-Maximization algorithms) with structured majorizing functions (1/3)

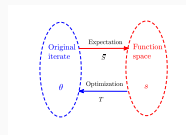


MM algorithms for the minimization of $F : \mathbb{R}^p \rightarrow \mathbb{R}$

$$F(\cdot) \leq Q(\cdot, \tau), \quad \forall \tau, \quad F(\tau) = Q(\tau, \tau)$$

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

$$\begin{aligned} w_k &\xrightarrow{\text{Minimize}} T(w_k) := \operatorname{argmin}_\theta \langle w_k, \phi(\theta) \rangle \\ &\xrightarrow{\text{Majorize}} w_{k+1} := \mathbb{E}_{T(w_k)} [S(X)] \\ &\xrightarrow{\text{Minimize}} T(w_{k+1}) := \operatorname{argmin}_\theta \langle w_{k+1}, \phi(\theta) \rangle \\ &\dots \end{aligned}$$



A root-finding algorithms: $\mathbb{E}_{T(\omega)} [S(X)] - \omega = 0$

SA-MM The oracles are Monte Carlo approximations of the intractable expectations.

Majorization-Minimization algorithms (Expectation-Maximization algorithms) with structured majorizing functions (2/3)

EM algorithm for the maximization of $F : \mathbb{R}^p \rightarrow \mathbb{R}$

$$F(\omega) := \frac{1}{n} \sum_{i=1}^n \log \int p_{\text{joint}}(Z_i, h; \omega) \nu(dh)$$

Structured minorizing functions (curved exponential family)

$$\mathcal{Q}(\cdot, \tau) := \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tau} [\log p_{\text{joint}}(Z_i, H; \cdot)] \quad \text{w.r.t. } p_{\text{joint}}(\cdot; \tau)$$

$$\log p_{\text{joint}}(Z_i, H; \cdot) = \left\langle \frac{1}{n} \sum_{i=1}^n S_i(H), \phi(\cdot) \right\rangle$$

A root-finding algorithms: $\mathbb{E}_{T(\omega)} \left[\frac{1}{n} \sum_{i=1}^n S_i(H) \right] - \omega = 0$

SA within EM The oracles are Monte Carlo approximations of the intractable expectations.

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)

Majorization-Minimization algorithms (Expectation-Maximization algorithms) with structured majorizing functions (3/3)

		oracles given by the system	oracles built by the user	Biased oracles	Unbiased oracles
EM	batch		✓	✓	(✓)
	online	✓	(✓)	✓	(✓)

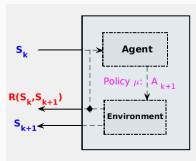
Value function in a Reward Markov process via Bellman equation (1/2)

Value function in a Reward Markov process:

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

$$\lambda \in (0, 1)$$

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



The Bellman equation $B[V] - V = 0$

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

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

TD(0) is a SA

Sutton (1987); Tsitsiklis and Van Roy (1997)

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

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

Value function in a Reward Markov process via Bellman equation (2/2)

	oracles given by the system	oracles built by the user	Biased oracle	Unbiased oracle
TD(0)	✓		✓	(✓)

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: $\mathbb{E}[H(\omega, X)] = h(\omega)$

From time homogeneous iterative algorithm to SA

$$\omega_{k+1} = M(\omega_k) \quad \Longrightarrow \quad \text{the fixed points: } M(\omega) - \omega = 0.$$

When M is not explicit but stochastic oracles are available, run

$$\omega_{k+1} = \omega_k + \gamma_{k+1} (\mathcal{M}(\omega_k, X_{k+1}) - \omega_k)$$

Does it converge to the same limiting points ? $\dots$ Lyapunov function !

Stochastic Approximation

Examples of SA: Stochastic Gradient and beyond

Non-asymptotic analysis

Variance Reduction within SA

Conclusion

► Asymptotic convergence analysis, when the horizon tends to infinity

Benveniste et al (1987/2012), Benâïm (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, ...

► 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

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

The assumptions

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

Lyapunov function V and control W

There exist $V : \mathbb{R}^d \rightarrow [0, +\infty)$, $W : \mathbb{R}^d \rightarrow [0, +\infty)$ 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	ω_* solution	$-\nabla R(\omega)$	$0.5 \ \omega - \omega_*\ ^2$	$-\langle \omega - \omega_*, h(\omega) \rangle$
and R strongly cvx	ω_* solution	$-\nabla R(\omega)$	$0.5 \ \omega - \omega_*\ ^2$	$W = V$ or, as above
Stochastic EM		$\bar{s}(\mathbf{T}(\omega)) - \omega$	$F(\mathbf{T}(\omega))$	$\ h(\omega)\ ^2$
TD(0)	$\Phi \omega_*$ solution	$\Phi' D(\mathbf{B} \Phi \omega - \Phi \omega)$	$0.5 \ \omega - \omega_*\ ^2$	$(\omega - \omega_*)' \Phi' D \Phi (\omega - \omega_*)$

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

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

- Variance

$$\mathbb{E} [\|H(\omega_k, X_{k+1}) - \mathbb{E} [H(\omega_k, X_{k+1}) | \mathcal{F}_k]\|^2 | \mathcal{F}_k] \leq \sigma_0^2 + \sigma_1^2 W(\omega_k)$$

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

$$\sqrt{c_V} (\sqrt{\tau_0}/2 + \sqrt{\tau_1}) < \rho,$$

$$c_V := \sup_{\omega} \frac{\|\nabla V(\omega)\|^2}{W(\omega)} < \infty.$$

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 $\gamma_k \in (0, \gamma_{\max})$,

$$\eta_1 \geq \sigma_1^2 + c_1 > 0$$

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

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

$$\begin{aligned} \sum_{k=1}^T \frac{\gamma_k \mu_k}{\sum_{\ell=1}^T \gamma_{\ell} \mu_{\ell}} \mathbb{E}[W(\omega_{k-1})] &\leq 2 \frac{\mathbb{E}[V(\omega_0)]}{\sum_{\ell=1}^T \gamma_{\ell} \mu_{\ell}} \\ &\quad + L_V \eta_0 \frac{\sum_{k=1}^T \gamma_k^2}{\sum_{\ell=1}^T \gamma_{\ell} \mu_{\ell}} \\ &\quad + c_V \sqrt{\tau_0} \frac{\sum_{k=1}^T \gamma_k}{\sum_{\ell=1}^T \gamma_{\ell} \mu_{\ell}} \\ \mu_{\ell} &= 2(\rho - \mathbf{b}_1) - \gamma_{\ell} L_V \eta_1 > 0 \end{aligned}$$

- η_{ℓ} depends on the bias and variance of the oracles; $\eta_0 > 0$.
- For unbiased oracles: $\tau_0 = \mathbf{b}_1 = 0$
- Better bounds when $V = W$; not discussed here

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

Case unbiased oracles: 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 $\mathcal{R}_T \sim \mathcal{U}(\{0, \dots, T-1\})$

or when W is convex: return the averaged iterate

$$T^{-1} \sum_{k=0}^{T-1} \omega_k$$

yields

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

- The left hand side comes from $T^{-1} \sum_{t=0}^{T-1} \mathbb{E}[W(\omega_t)] = \mathbb{E}[W(\omega_{\mathcal{R}_T})]$
- The right hand side: it is an *optimal* control in expectation.

Case unbiased oracles: ϵ -approximate stationary point

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

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: $\epsilon > 2\eta_0/\eta_1$,

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

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

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

"Biased oracles" mean :

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

Specific assumptions: $\rho > (\sqrt{\tau_0}/2 + \sqrt{\tau_1}) \left(\sup \|\nabla V\|/\sqrt{W} \right)$

where

$$\exists \rho > 0, \quad \forall \omega, \quad \langle \nabla V(\omega), h(\omega) \rangle \leq -\rho W(\omega)$$

When $\tau_0 \neq 0$

- Difficult !
- The previous strategy "constant step size, uniform random stopping time" does not hold: the RHS can not be made small by any choice of γ .

Example. SAEM with self-normalized Importance Sampling (m draws per iterations):

- $\tau_0 = O(1/m)$
- ϵ -approximate stationary point with $m \leftarrow m_\epsilon$, $T \leftarrow T_\epsilon$, $\text{cost}_\epsilon = O(T_\epsilon m_\epsilon)$

Case biased oracles: ϵ -approximate stationary point on an example

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

The SA-EM oracle

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

Make the bias small by choosing $m = m(\epsilon)$.

Complexity

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

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

For low precision regime,

$$T_\epsilon = \frac{C_1}{\epsilon}, \quad m_\epsilon = \frac{C_2}{\epsilon}, \quad \text{cost}_{\text{comp}} = T_\epsilon (nm_\epsilon \text{cost}_{\text{MC}} + \text{cost}_{\text{opt}})$$

Other rates for high precision regime.

Variance Reduction within SA

Stochastic Approximation

Examples of SA: Stochastic Gradient and beyond

Non-asymptotic analysis

Variance Reduction within SA

Conclusion

- Add a random variable to the *natural oracle* $H(\omega, X)$
- *Control variates* U , classical in Monte Carlo:

$$\mathbb{E}[H(\omega, X) + U] = \mathbb{E}[H(\omega, X)] \quad \text{Var}(H(\omega, X) + U) < \text{Var}(H(\omega, X)).$$

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)

Finite sum case: which "control variate" ?

"Finite sum" case. $h(\omega) = n^{-1} \sum_{i=1}^n h_i(\omega)$ and h_i globally Lipschitz.

Usual oracle for $h(\omega_k)$.

$$H(\omega_k, X_{k+1}) := \frac{1}{b} \sum_{i \in X_{k+1}} h_i(\omega_k) \quad X_{k+1} \text{ mini batch of size } b \text{ in } \{1, \dots, n\}$$

SVRG approach fix ℓ and for $k \geq \ell$: add the term

$$\frac{1}{n} \sum_{i=1}^n h_i(\omega_\ell) - \frac{1}{b} \sum_{i \in X_{k+1}} h_i(\omega_\ell)$$

SAGA approach Add the term

$$\frac{1}{n} \sum_{i=1}^n \tilde{h}_{i,k} - \frac{1}{b} \sum_{i \in X_{k+1}} \tilde{h}_{i,k}$$

and update the auxiliary quantity

$$\tilde{h}_{i,k+1} := h_i(\omega_k) \quad i \in X_{k+1}, \quad \tilde{h}_{i,k+1} := \tilde{h}_{i,k} \quad \text{otherwise.}$$

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)

The SPIDER oracle is

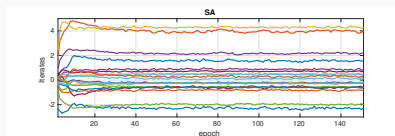
$$H_{k+1}^{\text{SP}} := \frac{1}{b} \sum_{i \in X_{k+1}} h_i(\omega_k) + \underbrace{H_k^{\text{SP}}}_{\text{oracle for } h(\omega_{k-1})} - \underbrace{\frac{1}{b} \sum_{i \in X_{k+1}} h_i(\omega_{k-1})}_{\text{oracle for } h(\omega_{k-1})}$$

- Implementation: Run $K_{\text{out}} K_{\text{in}}$ iterations and *refresh* the control variate every K_{in} iterations
- It holds: at outer loop $\#t$,

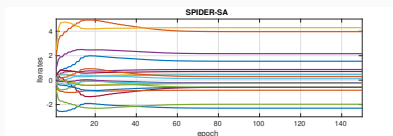
$$\left(1 - 2\gamma^2 L^2 \frac{K_{\text{in}}}{b}\right) \sum_{k=0}^{K_{\text{in}}} \mathbb{E} [\|H_{t,k}^{\text{SP}} - h(\omega_{t,k-1})\|^2 | \text{past}_{t,0}] \leq 2\gamma^2 L^2 \frac{K_{\text{in}}}{b} \sum_{k=1}^{K_{\text{in}}} \mathbb{E} [\|h(\omega_{t,k})\|^2 | \text{past}_{t,0}]$$

Efficiency ... via plots (here)

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

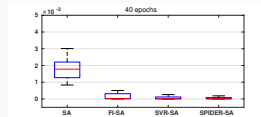
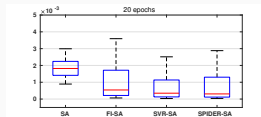


Estimation of 20 parameters, one path of SA

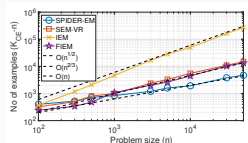


Estimation of 20 parameters, one path of SPIDER-SA

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

Stochastic Approximation

Examples of SA: Stochastic Gradient and beyond

Non-asymptotic analysis

Variance Reduction within SA

Conclusion

- SA methods with non-gradient mean field and/or biased oracles - in ML and computational 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, ...