Stochastic Optimization beyond Gradient for Machine Learning

Gersende Fort

CNRS - Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS)
Toulouse, France

Outline of the talk

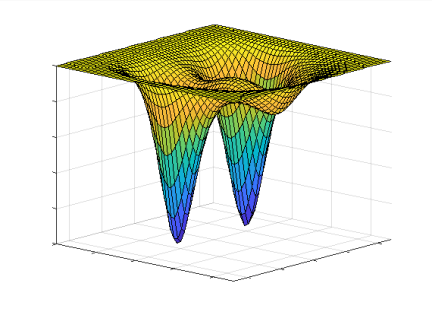
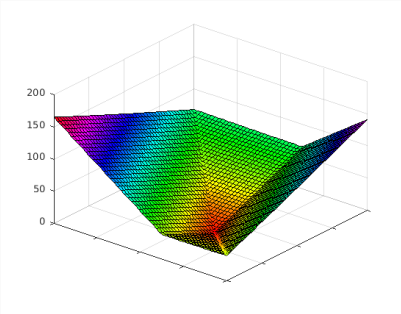
- Mathematics for Stochastic Optimization
  - from Optimization to Stochastic Optimization
  - Objective function in ML
  - Sources of randomness
- Focus: Stochastic Approximation

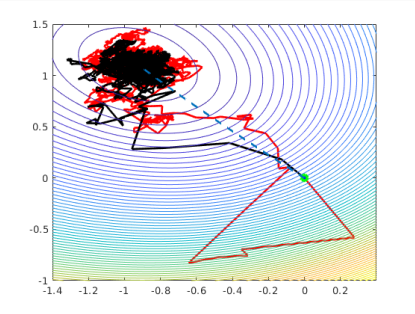
Stochastic Optimization

$$\operatorname{argmin}_{\omega \in \mathbb{R}^d} L(\omega)$$

when L has no closed form expression but can be written as an expectation

$$\operatorname{argmin}_{\omega \in \mathbb{R}^d} \mathbb{E}\left[\ell(\omega; Z)\right]$$

Objective function: convex or not, smooth or not

Method: produce iterates $\omega_1, \omega_2, \cdots$ by using random sources.

Stochastic Optimization in Machine Learning

learn a system i.e. find $\omega \in \mathbb{R}^d$, from examples $X_1, \cdots, X_t, X_{t+1}, \cdots$

Batch learning. the ML model is trained using a batch of examples $X_1, \cdots, X_n$.

$$\text{Ex.} \qquad \operatorname{argmin}_\omega \frac{1}{n} \sum_{i=1}^{n} \ell_i(\omega; X_i)$$

Online learning. the ML model is adjusted sequentially from fresh data in order to optimize a long time observation-based criterion.

$$\text{Ex.} \qquad \operatorname{argmin}_\omega \mathbb{E}\left[\ell(\omega; X)\right]$$

from $X_1, \cdots, X_t, X_{t+1}, \cdots$ s.t. $\mathbb{E}\left[\ell(\omega; X)\right] = \lim_t \frac{1}{t} \sum_{i=1}^{t} \ell(\omega; X_i)$ a.s..

Randomness in Optimization algorithm: why, and safe ?

$$\omega_{t+1} = \mathsf{M}\left(\omega_t, Z_{t+1}, t\right)$$

Internal randomness. $Z_{t+1} = X_{t+1}$ is given by the system to be learnt

- reinforcement learning
- online learning

External randomness. Z_{t+1} is a numerical tool for learning the system

- subsampling (data in large batch learning, directions for vector valued parameters ω, $\cdots$)
- random quantization
- intractable integrals/sum in the definition of M

Safe ? A *stochastic* error

- a bias, w.r.t. some "ideal" iterative scheme $\tilde{\omega}_{t+1} = \mathsf{M}^\star(\tilde{\omega}_t, t)$
- a variance, which may cause unstability and slow down the convergence

The Mathematics of Stochastic Optimization

1. Propose new SO algorithms
2. Understand the role of *design* parameters
3. (limiting) Behavior of the iterative scheme
4. Comparison of algorithms

Case: Stochastic Approximation algorithms

- What is *Stochastic Approximation*
- An optimization method in Machine Learning
- Finite time analysis
- Best strategies, Variance reduction

Stochastic Approximation

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 $\{\gamma_t\}_t$ and an initial value $\omega_0 \in \mathbb{R}^d$.
Repeat:

$$\omega_{t+1} = \omega_t + \gamma_{t+1}\ H(\omega_t, Z_{t+1})$$

where $H(\omega_t, Z_{t+1})$ is a stochastic oracle of $h(\omega_t)$.

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; $H(\omega, Z_{t+1})$ is a Monte Carlo approximation.

Stochastic Approximation in Machine Learning (1/3)

- Stochastic Gradient algorithm $\qquad h(\omega) = -\nabla R(\omega)$

$$\omega_{t+1} = \omega_t + \gamma_{t+1}\ H(\omega_t, Z_{t+1}) \qquad \mathbb{E}\left[H(\omega_t, Z_{t+1})|\text{past}_t\right] = h(\omega_t)$$

- Compression with Stochastic Gradient, when frugal algorithms are mandatory

Compression operator $x \mapsto \mathcal{C}(x, U)$, random or deterministic;

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

increasing interest in distributed optimization

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

gradient at a perturbed iterate: Straight-Through Estimator

Stochastic Approximation in Machine Learning (2/3)

$\omega_{t+1} = \omega_t + \gamma_{t+1} H(\omega_t, Z_{t+1}) \qquad H(\omega_t, Z_{t+1}) \approx h(\omega_t)$

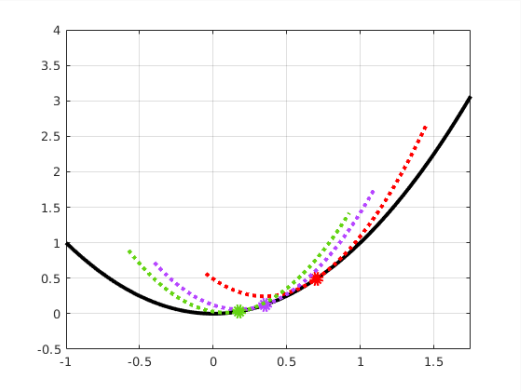
- Stochastic Majorization-Minimization algorithm

in the *structured* case

$$\mathbb{E}\left[\ell(\cdot, Z)\right] \leq C_{\tilde{\omega}} + \psi(\cdot) + \left\langle \mathbb{E}\left[\mathsf{S}(\tilde{\omega}, Z)\right], \phi(\cdot) \right\rangle$$

and unique minimizer

$$\mathsf{T}(s) := \operatorname{argmin}_{\omega} \psi(\cdot) + \langle s, \phi(\cdot) \rangle$$



The limiting points solve $\omega^\star = \mathsf{T}(s^\star)$, where $\mathbb{E}\left[\mathsf{S}(\mathsf{T}(s^\star), Z)\right] - s^\star = 0$

Examples:
- Proximal gradient algorithm
- Mirror descent algorithm
- When ℓ is an intractable integral of a positive function ($\rightarrow$ EM algorithm)
- Training some Mixture of Experts models
- Dictionary Learning
- Variational inference

Stochastic Approximation in Machine Learning (3/3)

$\omega_{t+1} = \omega_t + \gamma_{t+1} H(\omega_t, Z_{t+1}) \qquad H(\omega_t, Z_{t+1}) \approx h(\omega_t)$

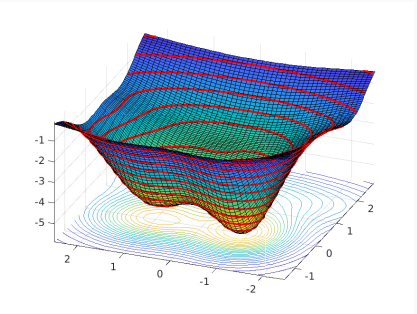
- Fixed point algorithms with a Lyapunov function

When the map M in intractable and the goal is to solve $\omega = \mathsf{M}(\omega)$, run a SA algorithm

$$h(\omega) = \mathsf{M}(\omega) - \omega.$$

It *may* work if there exists a Lyapunov function V

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



Example: Reinforcement learning: value function $\mathcal{V}$ of a policy with linear function approximation $\mathcal{V}(\cdot) = \Phi\omega$, by the TD(0) algorithm.

$$H(\omega, (S_t, S_{t+1}, R(S_t, S_{t+1}))) = \left(R(S_t, S_{t+1}) + \lambda \left\langle \Phi(S_{t+1}, :), \omega \right\rangle - \langle \Phi(S_t, :), \omega \rangle\right) \Phi(S_t, :)^\top$$

A finite time analysis of SA algorithms

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 $T \geq 1$

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

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

$W(\omega_{t-1})$ quantifies how far the iterate ω_{t-1} is, from the limiting set $\mathcal{L} \supseteq \{h = 0\}$.

This term: the impact of the initial value ω_0.

This term: due to the bias and variance of the oracle.

This term: exists when the oracle is biased and the bias does not vanish when $W = 0$.

Design parameters, Variance Reduction

Is there an optimal strategy for selecting the γ_t's and the output of the algorithm ?

When unbiased oracles, the strategy "constant step size" and "return an iterate chosen at random in $\{\omega_0, \cdots, \omega_{T-1}\}$" is optimal

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

Complexity analysis: ϵ-approximate stationary point

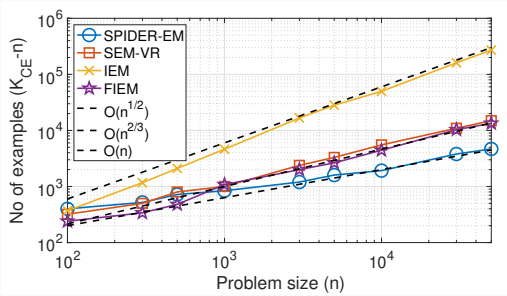
It holds $\mathbb{E}\left[W(\omega_{\mathcal{R}_T})\right] \leq \epsilon$ for any $T \geq T_\epsilon$

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

Variance reduction when h is a finite sum

The oracle is not unique

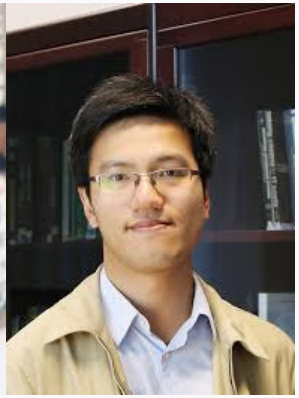
- Variance reduction scheme for SA: adapted from SVRG, SAGA, SPIDER (SARAH)
- SPIDER is the most efficient



Works in collaboration







A. Dieuleveut — IPP, France

F. Forbes — INRIA, France

E. Moulines — IPP, France

H. D. Nguyen — La Trobe Univ., Australia

H.-T. Wai — Hong-Kong Univ.

- Sequential Sample Average Majorization-Minimization. *Submitted*
- Federated Majorize-Minimization for large scale learning. *Submitted*
- Stochastic Approximation beyond Gradient for Signal Processing and Machine Learning. *IEEE Trans Signal Processing, 2023.*
- Stochastic Variable Metric Proximal Gradient with variance reduction for non-convex composite optimization. *Statistics and Computing, 2023.*
- An online Minorization-Maximization algorithm. *IFCS 2022 proceedings*.
- Federated Expectation Maximization with heterogeneity mitigation and variance reduction. *NeurIPS, 2021.*
- The Fast Incremental Expectation Maximization for finite-sum optimization: nonasymptotic convergence. *Statistics and Computing, 2021.*
- The Perturbed Prox-Preconditioned SPIDER algorithm: non-asymptotic convergence bounds. *IEEE Statistical Signal Processing Workshop proceedings, 2021.*
- Geom-SPIDER-EM: Faster Variance Reduced Stochastic Expectation Maximization for Nonconvex Finite-Sum Optimization. *IEEE International Conference on Acoustics, Speech and Signal Processing proceedings, 2021.*
-A Stochastic Path Integrated Differential Estimator Expectation Maximization Algorithm. *NeurIPS, 2020.*

Sketch of proof

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

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

The definition of the iterative scheme

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

Sketch of proof

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

The conditional expectation

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

Sketch of proof

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

The mean field h and the bias term

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

Sketch of proof

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

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

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

Sketch of proof

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

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 $k \geq 0$,

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

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