Stochastic approximation-based algorithms, when the Monte Carlo bias does not vanish

Gersende Fort

Institut de Mathématiques de Toulouse
CNRS
Toulouse, France

Based on joint works with

- Yves Atchadé (Univ. Michigan, USA)
- Eric Moulines (Ecole Polytechnique, France)
- Edouard Ollier (ENS Lyon, France)
- Laurent Risser (IMT, France).
- Adeline Samson (Univ. Grenoble Alpes, France).

and published in the papers (or works in progress)

- Convergence of the Monte-Carlo EM for curved exponential families (Ann. Stat., 2003)
- On Perturbed Proximal-Gradient algorithms (JMLR, 2017)
- Stochastic Proximal Gradient Algorithms for Penalized Mixed Models (Statistics and Computing, 2018)
- Stochastic FISTA algorithms : so fast ? (IEEE workshop SSP, 2018)

This talk : answer a computationnel issue

- Find

$$\theta^* \in \mathrm{argmin}_{\theta \in \Theta} \left(f(\theta) + g(\theta) \right) \tag{1}$$

where

- $\Theta \subseteq \mathbb{R}^d$ *(extension to any Hilbert possible; not done)*
- g is **not smooth**, but is **convex** and proper, lower semi-continuous (*"prox" operator*)
- f is **is not explicit / is untractable**, ∇f exists but **is not explicit / is untractable**
  When proving results : f is convex and ∇f is Lipschitz

- In this talk : numerical tools to solve (1) based on first order methods; convergence analysis.

Outline

Example 1 : large scale learning

Minimization of a composite function

- $g = 0$ or g is a penalty / regularization / constraint condition on the parameter θ
- f is an (empirical) loss function associated to N examples

$$f(\theta) = \frac{1}{N} \sum_{i=1}^{N} f_i(\theta)$$

when N is large

For any i, f_i and ∇f_i can be evaluated at any point θ **but** the computation of the sum over N terms is too expensive.

Rmk that $\nabla f(\theta) = \mathbb{E}\left[\nabla f_I(\theta)\right]$ where I r.v. uniform on $\{1, \cdots, N\}$.

Example 2 : binary graphical model

Minimization of a composite function

- Observation $y \in \{-1, 1\}^p$ *(a binary vector of length p, collecting the binary values of p nodes)*, with statistical model

$$\pi_\theta(y) \propto \exp\left(\sum_{i=1}^p \theta_i y_i + \sum_{i=1}^p \sum_{j=i+1}^p \theta_{ij} y_i y_j\right)$$

  with an **untractable** normalizing constant $\exp(Z_\theta)$. θ collects the "weights".

- f is the negative log-likelihood of N indep. observations

$$f(\theta) = -\log Z_\theta + \sum_{i=1}^p \theta_i \left(N^{-1} \sum_{n=1}^N Y_i^{(n)}\right) + \sum_{i=1}^p \sum_{j=i+1}^p \theta_{ij} \left(N^{-1} \sum_{n=1}^N \mathbb{1}_{Y_i^{(n)} = Y_j^{(n)}}\right)$$

  In this model $\nabla f(\theta) = \mathbb{E}_\theta \left[H(X, \theta)\right]$ where $X \sim \pi_\theta$

- $g = 0$ or g is a penalty / regularization / constraint condition on the parameter θ (the number of observations $N << p^2/2$)

Example 3 : Parametric inference in Latent variable models

Minimization of a composite function

- g is a penalty function (e.g. for sparsity condition on θ)
- f is the negative log-likelihood of the N observations

$$f(\theta) = -\log \int_{\mathsf{X}} h(x, Y_{1:N}; \theta)\, \nu(\mathsf{d}x)$$

and the gradient is of the form

$$\nabla f(\theta) = \int_{\mathsf{X}} \partial_\theta \log h(x, Y_{1:N}; \theta)\ \frac{h(x, Y_{1:N}; \theta)}{\int_{\mathsf{X}} h(u, Y_{1:N}; \theta) \nu(\mathsf{d}u)}\, \nu(\mathsf{d}x)$$

i.e. an expectation w.r.t. the a posteriori distribution *(known up to a normalizing constant in these models)*

Outline

Numerical solution : the ingredient

$$\mathrm{argmin}_{\theta \in \Theta} F(\theta) \qquad \text{with } F(\theta) = \underbrace{f(\theta)}_{\text{smooth}} + \underbrace{g(\theta)}_{\text{non smooth}}$$

The Proximal Gradient algorithm

Given a stepsize sequence $\{\gamma_n, n \geq 0\}$, iterative algorithm:

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1} \nabla f(\theta_n)\right)$$

where

$$\mathrm{Prox}_{\gamma,g}(\tau) \stackrel{\text{def}}{=} \mathrm{argmin}_{\theta \in \Theta} \left(g(\theta) + \frac{1}{2\gamma} \|\theta - \tau\|^2 \right)$$

Proximal map: Moreau(1962)

Proximal Gradient algorithm: Beck-Teboulle(2010); Combettes-Pesquet(2011); Parikh-Boyd(2013)

- A generalization of the gradient algorithm to a composite objective fct.
- A Majorize-Minimize algorithm from a quadratic majorization of f (since Lipschitz gradient) which produces a sequence $\{\theta_n, n \geq 0\}$ such that

$$F(\theta_{n+1}) \leq F(\theta_n).$$

In our frameworks, $\nabla f(\theta)$ is not available.

Numerical solution : a perturbed proximal-gradient algorithm

The Perturbed Proximal Gradient algorithm

Given a stepsize sequence $\{\gamma_n, n \geq 0\}$, iterative algorithm:

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\mathbf{H_{n+1}}\right)$$

where H_{n+1} is an approximation of $\nabla f(\theta_n)$.

Useful for the proof: observe

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n) - \underbrace{\gamma_{n+1}\left(H_{n+1} - \nabla f(\theta_n)\right)}_{perturbation}\right)$$

Convergence result : the assumptions (1/2)

$$\mathrm{argmin}_{\theta \in \Theta} F(\theta) \qquad \text{with } F(\theta) = f(\theta) + g(\theta)$$

where

- the function g: $\mathbb{R}^d \to [0, \infty]$ is convex, non smooth, not identically equal to $+\infty$, and lower semi-continuous
- the function f: $\mathbb{R}^d \to \mathbb{R}$ is a smooth **convex** function

  i.e. f is continuously differentiable and there exists $L > 0$ such that

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L \, \|\theta - \theta'\| \qquad \forall \theta, \theta' \in \mathbb{R}^d$$

- $\Theta \subseteq \mathbb{R}^d$ is the domain of g: $\Theta = \{\theta \in \mathbb{R}^d : g(\theta) < \infty\}$.
- The set $\mathrm{argmin}_\Theta F$ is a non-empty subset of Θ.

Convergence results (2/2)

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1} \; H_{n+1}\right) \qquad \text{with } H_{n+1} \approx \nabla f(\theta_n)$$

Set: $\mathcal{L} = \mathrm{argmin}_\Theta (f+g)$ $\qquad \eta_{n+1} = H_{n+1} - \nabla f(\theta_n)$

Theorem (Atchadé, F., Moulines (2017))

Assume

- g *convex, lower semi-continuous;* f *convex,* C^1 *and its gradient is Lipschitz with constant* L*;* $\mathcal{L}$ *is non empty.*
- $\sum_n \gamma_n = +\infty$ *and* $\gamma_n \in (0, 1/L]$.
- *Convergence of the series*

$$\sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^2, \qquad \sum_n \gamma_{n+1}\eta_{n+1}, \qquad \sum_n \gamma_{n+1} \left\langle \mathsf{T}_n, \eta_{n+1} \right\rangle$$

where $\mathsf{T}_n = \mathrm{Prox}_{\gamma_{n+1},g}(\theta_n - \gamma_{n+1}\nabla f(\theta_n))$.

Then there exists $\theta_\star \in \mathcal{L}$ *such that* $\lim_n \theta_n = \theta_\star$.

Sketch of proof

Its proof relies on

1. a deterministic Lyapunov inequality

$$\|\theta_{n+1}-\theta_\star\|^2 \leq \|\theta_n-\theta_\star\|^2 - \underbrace{2\gamma_{n+1}\left(F(\theta_{n+1}) - \min F\right)}_{\text{non-negative}} \underbrace{-2\gamma_{n+1}\left\langle \mathsf{T}_n - \theta_\star, \eta_{n+1}\right\rangle + 2\gamma_{n+1}^2 \|\eta_{n+1}\|^2}_{\text{signed noise}}$$

2. (an extension of) the Robbins-Siegmund lemma

   Let $\{v_n, n \geq 0\}$ and $\{\chi_n, n \geq 0\}$ be non-negative sequences and $\{\xi_n, n \geq 0\}$ be such that $\sum_n \xi_n$ exists. If for any $n \geq 0$,

$$v_{n+1} \leq v_n - \chi_{n+1} + \xi_{n+1}$$

   then $\sum_n \chi_n < \infty$ and $\lim_n v_n$ exists.

Rmk: deterministic lemma, signed noise.

What about Nesterov-based acceleration ? (FISTA)

Let $\{t_n, n \geq 0\}$ be a positive sequence s.t.

$$\gamma_{n+1} t_n (t_n - 1) \leq \gamma_n t_{n-1}^2$$

Nesterov acceleration of the Proximal Gradient algorithm

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g} \left(\tau_n - \gamma_{n+1} \nabla f(\tau_n) \right)$$
$$\tau_{n+1} = \theta_{n+1} + \frac{t_n - 1}{t_{n+1}} \left(\theta_{n+1} - \theta_n \right)$$

Nesterov(2004), Tseng(2008), Beck-Teboulle(2009)

Zhu-Orecchia (2015); Attouch-Peypouquet(2015); Bubeck-Lee-Singh(2015); Su-Boyd-Candes(2015)

(deterministic) Proximal-gradient $\quad F(\theta_n) - \min F = O\left(\frac{1}{n}\right)$

(deterministic) Accelerated Proximal-gradient $\quad F(\theta_n) - \min F = O\left(\frac{1}{n^2}\right)$

Convergence results for perturbed FISTA

When $\nabla f(\tau_n)$ is replaced with H_{n+1}

Perturbed FISTA

$$\begin{aligned} H_{n+1} &\approx \nabla f(\tau_n) \\ \theta_{n+1} &= \mathrm{Prox}_{\gamma_{n+1},g}\left(\tau_n - \gamma_{n+1} H_{n+1}\right) \\ \tau_{n+1} &= \theta_{n+1} + \frac{t_n - 1}{t_{n+1}}\left(\theta_{n+1} - \theta_n\right) \end{aligned}$$

Under conditions on γ_n, t_n and on the perturbation $\tilde{\eta}_{n+1} \stackrel{\text{def}}{=} H_{n+1} - \nabla f(\tau_n)$

$$\sum_n \gamma_{n+1} t_n \left\langle z_n - \theta^*, \tilde{\eta}_{n+1} \right\rangle < \infty$$

we have (F., Risser, Atchadé, Moulines; 2018)

- $\lim_n \gamma_{n+1} t_n^2 F(\theta_n)$ exists
- Explicit control of this quantity.

Outline

Monte Carlo approximation

- We consider the case when

$$\nabla f(\theta) = \int_{\mathsf{X}} H(x,\theta)\ \pi_\theta(\mathrm{d}x)$$

and the approximation relies on **a Monte Carlo approximation**

$$H_{n+1} \stackrel{\text{def}}{=} \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} H(X_{j,n};\theta_n)$$

- In our motivating examples 2 and 3
  - π_θ is known up to a normalization constant
  - **exact sampling** from π_θ is **not possible**
  - MCMC techniques can always be used : at iteration n, the points $X_{1,n}, X_{2,n}, \cdots$ are from a Markov chain with invariant distribution π_{θ_n}.

Convergence results on Markov chains F., Moulines (2003)

- The approximation is biased

$$\mathbb{E}\left[\frac{1}{m_{n+1}}\sum_{i=1}^{m_{n+1}} H(X_{i,n},\theta)|\mathcal{F}_n\right] \neq \int H(x,\theta)\ \pi_{\theta_n}(\mathrm{d}x)$$

- The bias may vanish when the number of points tends to infinity

$$\left|\mathbb{E}\left[\frac{1}{m_{n+1}}\sum_{i=1}^{m_{n+1}} H(X_{i,n},\theta)\Big|\mathcal{F}_n\right] - \int H(x,\theta)\ \pi_{\theta_n}(\mathrm{d}x)\right| \leq \frac{C(\theta_n, X_{0,n})}{m_{n+1}}$$

$$\mathbb{E}\left[\left|\frac{1}{m_{n+1}}\sum_{i=1}^{m_{n+1}} H(X_{i,n},\theta) - \int H(x,\theta)\ \pi_{\theta_n}(\mathrm{d}x)\right|^p \Big|\mathcal{F}_n\right] \leq \frac{\tilde{C}(\theta_n, X_{0,n})}{m_{n+1}^{p/2}}$$

- The control of this bias depends on the current value of the parameter θ_n

These results depend on the **ergodic properties** of the Markov chain: assumptions on the target density π_θ and on the transition kernel P_θ of the Markov chain are required.
Assumptions of the form $\sup_\theta \sup_x |H(x,\theta)|/W(x) < \infty$ are also used in these bounds.

Impact of the bias (1/2)

let us check the condition "$\sum_n \gamma_n \eta_n < \infty$ w.p.1":

$$\sum_n \gamma_{n+1} \eta_{n+1} = \sum_n \gamma_{n+1} \left(H_{n+1} - \nabla f(\theta_n) \right)$$

- The RHS

$$\sum_n \gamma_{n+1} \left\{ H_{n+1} - \mathbb{E}\left[H_{n+1} | \mathcal{F}_n \right] \right\} + \sum_n \gamma_{n+1} \underbrace{\left\{ \mathbb{E}\left[H_{n+1} | \mathcal{F}_n \right] - \nabla f(\theta_n) \right\}}_{\substack{\text{unbiased MC: null} \\ \text{biased MC: } O(1/m_n)}}$$

- The most technical case: the biased case with constant batch size $m_n = m$

Impact of the bias (2/2) - case $m_n = m = 1$

- Let P_θ be the Markov transition kernel of the chain with inv. dstribution π_θ.
- Solution $\widehat{H}_\theta$ to the Poisson equation

$$H(x,\theta) - \int H(y,\theta)\pi_\theta(\mathsf{d}y) = \widehat{H}_\theta - P_\theta \widehat{H}_\theta(x)$$

- This yields, by choosing $X_{0,n} = X_{1,n-1}$

$$\begin{aligned} H(X_{1,n},\theta_n) - \int_{\mathsf{X}} H(y,\theta_n)\pi_{\theta_n}(\mathsf{d}y) &= \widehat{H}_{\theta_n}(X_1) - P_{\theta_n}\widehat{H}_{\theta_n}(X_{1,n}) \\ &= \widehat{H}_{\theta_n}(X_{1,n}) - P_{\theta_n}\widehat{H}_{\theta_n}(X_{0,n}) + P_{\theta_n}\widehat{H}_{\theta_n}(X_{0,n}) - P_{\theta_n}\widehat{H}_{\theta_n}(X_{1,n}) \\ &= \widehat{H}_{\theta_n}(X_{1,n}) - P_{\theta_n}\widehat{H}_{\theta_n}(X_{0,n}) \qquad \text{Martingale increment} \\ &+ P_{\theta_n}\widehat{H}_{\theta_n}(X_{1,n-1}) - P_{\theta_{n-1}}\widehat{H}_{\theta_{n-1}}(X_{1,n-1}) \qquad \text{Regularity in } \theta \\ &+ P_{\theta_{n-1}}\widehat{H}_{\theta_{n-1}}(X_{1,n-1}) - P_{\theta_n}\widehat{H}_{\theta_n}(X_{1,n}) \qquad \text{telescopic} \end{aligned}$$

Strategy 1: vanishing bias $m_n \to \infty$ (1/2)

- For almost-sure convergence of $\{\theta_n, n \geq 0\}$

Conditions on the batch size m_n and the stepsize γ_n for the convergence

$$\sum_n \gamma_n = +\infty, \qquad \sum_n \frac{\gamma_n^2}{m_n} < \infty; \qquad \sum_n \frac{\gamma_n}{m_n} < \infty \textit{ (biased case)}$$

Conditions on the Markov kernels: *There exist $\lambda \in (0,1)$, $b < \infty$, $p \geq 2$ and a measurable function $W : \mathsf{X} \to [1, +\infty)$ such that*

$$\sup_{\theta \in \Theta} |H_\theta|_W < \infty, \qquad \sup_{\theta \in \Theta} P_\theta W^p \leq \lambda W^p + b.$$

In addition, for any $\ell \in (0, p]$, there exist $C < \infty$ and $\rho \in (0,1)$ such that for any $x \in \mathsf{X}$,

$$\sup_{\theta \in \Theta} \| P_\theta^n(x, \cdot) - \pi_\theta \|_{W^\ell} \leq C \rho^n W^\ell(x). \tag{2}$$

Condition on Θ: Θ is bounded.

Constant step sizes $\gamma_n = \gamma$ are allowed as soon as $\sum_n m_n^{-1} < \infty$.

Strategy 1: vanishing bias $m_n \to \infty$ (2/2)

- For rates of convergence in L^q on the functional

$$\left\| F\left(\frac{1}{n}\sum_{k=1}^{n}\theta_k\right) - \min F\right\|_{L^q} \leq \left\| \frac{1}{n}\sum_{k=1}^{n} F(\theta_k) - \min F\right\|_{L^q} \leq u_n$$

$u_n = O(\ln n / n)$

with increasing batch size and constant stepsize

$$\gamma_n = \gamma_\star \qquad\qquad m_n \propto n.$$

Rate with $O(n^2)$ Monte Carlo samples !

After n iterations : the rate of the perturbed Proximal-Gradient is $O(1/n)$, using n^2 Monte Carlo simulations.

Given n Monte Carlo simulations: the rate is $O(1/\sqrt{n})$.

Strategy 2: **NON**-vanishing bias $m_n = m$. (1/2)

- "Stochastic Approximation" framework Benveniste, Metivier, Priouret (1990)
- For almost-sure convergence of $\{\theta_n, n \geq 0\}$

Conditions on the stepsize γ_n for the convergence

Condition on the step size:

$$\sum_n \gamma_n = +\infty \qquad \sum_n \gamma_n^2 < \infty \qquad \sum_n |\gamma_{n+1} - \gamma_n| < \infty$$

Condition on the Markov chain: *same as in the case "increasing batch size" and there exists a constant C such that for any $\theta, \theta' \in \Theta$*

$$|H_\theta - H_{\theta'}|_W + \sup_x \frac{\|P_\theta(x,\cdot) - P_{\theta'}(x,\cdot)\|_W}{W(x)} + \|\pi_\theta - \pi_{\theta'}\|_W \leq C\,\|\theta - \theta'\|.$$

Condition on the Prox:

$$\sup_{\gamma \in (0,1/L]} \sup_{\theta \in \Theta} \gamma^{-1} \, \|\mathrm{Prox}_{\gamma,g}(\theta) - \theta\| < \infty.$$

Condition on Θ: Θ is bounded.

Strategy 2: **NON**-vanishing bias $m_n = m$. (2/2)

- For rates of convergence in L^q on the functional

$$\left\| F\left(\frac{1}{n}\sum_{k=1}^{n}\theta_k\right) - \min F \right\|_{L^q} \leq \left\| \frac{1}{n}\sum_{k=1}^{n} F(\theta_k) - \min F \right\|_{L^q} \leq u_n$$

$u_n = O(1/\sqrt{n})$

with (slowly) decaying stepsize

$$\gamma_n = \frac{\gamma_\star}{n^a}, a \in [1/2, 1] \qquad\qquad m_n = m_\star.$$

With averaging: optimal rate, even with slowly decaying stepsize $\gamma_n \sim 1/\sqrt{n}$.

After n iterations : the rate of the perturbed Proximal-Gradient is $O(1/\sqrt{n})$, using n Monte Carlo simulations.

What about Stochastic FISTA ?

- We prove F., Risser, Atchadé, Moulines (2018)

$$\lim_n n^2 F(\theta_n) < \infty \quad \text{a.s.} \qquad \sup_n n^2 \mathbb{E}\left[F(\theta_n)\right] < \infty$$

with

$$t_n = O(n), \qquad \gamma_n = \gamma \qquad m_n = O(n^3)$$

- After n Monte Carlo simulations :
  - the rate is $O(1/\sqrt{n})$
  - the same rate as the (perturbed) Proximal-Gradient with an averaging strategy.

Outline

Latent variable models, curved exponential family

- One motivation was "penalized inference in latent variable models"

$$\mathrm{argmin}_\theta - \log \int_{\mathsf{X}} h(x,\theta)\nu(\mathsf{d}x) + g(\theta)$$

- When curved exponential family

$$h(x,\theta) = \exp(\phi(\theta) + \langle S(x), \psi(\theta)\rangle)$$

- In that case, Proximal-Gradient algo gets into

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1} g}\left(\theta_n - \gamma_{n+1}\{\nabla\phi(\theta_n) + \Psi(\theta_n)\bar{S}(\theta_n)\}\right)$$

where

$$\bar{S}(\theta_n) = \int S(z)\ \pi_{\theta_n}(\mathsf{d}z).$$

EM and Gdt-Prox

- Expectation-Maximization: a famous algorithm to solve this optimization issue in these models
- It can be shown Ollier, F., Samson (2018) that the proximal-gradient algorithm is a (Generalized) EM algorithm under regularity conditions on $\phi, \psi, \bar{S}$.

Stochastic EM and Stochastic Gdt-Prox

- Stochastic proximal-gradient algorithm

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1} g}\left(\theta_n - \gamma_{n+1}\{\nabla \phi(\theta_n) + \Psi(\theta_n) S_{n+1}\}\right)$$

where

$$S_{n+1} \approx \bar{S}(\theta_n)$$

- Strategy 1

$$S_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n})$$

- Strategy 2

$$S_{n+1} = (1-\delta_n) S_n + \frac{\delta_n}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n})$$

- These two strategies correspond resp. to a (generalized) MCEM and a (generalized) SAEM.