Convergence de méthodes de gradient stochastiques à biais persistant

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

- 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 ? (SSP, 2018)

Outline

Stochastic optimization methods: why ? (1/3)

Example 1 (large scale learning)

- optimization of

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

  when N is large
- through an iterative algorithm.

Stochastic since:

- at iteration k, choose **at random** a subset $\mathcal{I}_k$ of the N examples

Stochastic optimization methods: why ? (2/3)

Example 2 (inference in hidden variable models)

- Maximization of a function defined through an intractable integral

$$F(\theta) \stackrel{\text{def}}{=} \int_{\mathsf{X}} h(x,\theta)\,\mu(\mathrm{d}x) \qquad h \geq 0$$

- through the *Expectation-Maximization* algorithm, a MM-algorithm.

Stochastic since:

- at each iteration, given θ_n, write

$$\log F(\theta) \geq \log F(\theta_n) + \int_{\mathsf{X}} \log h(x,\theta)\,\frac{h(x,\theta_n)}{F(\theta_n)}\,\mu(\mathrm{d}x) - \int_{\mathsf{X}} \log h(x,\theta_n)\,\frac{h(x,\theta_n)}{F(\theta_n)}\,\mu(\mathrm{d}x)$$

  and define θ_{n+1} as the maximum of the RHS.

- When the expectation is not explicit, replace it with **a Monte Carlo sum**

$$\frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} \log h(X_{i,n},\theta) \qquad X_{i,n} \text{"related to"}\ \pi_{\theta_n}(\mathrm{d}x) \stackrel{\text{def}}{=} \frac{h(x,\theta_n)}{F(\theta_n)}\,\mu(\mathrm{d}x)$$

Stochastic optimization methods: why ? (3/3)

Example 3 (gradient-based methods with a non-explicit gradient)

- Gradient-based methods with a (non explicit) gradient of the form

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

where π_θ is a probability distribution.

Stochastic since:

- Replace the exact gradient with a **Monte Carlo sum**

$$\frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} H(X_{i,n}, \theta_n)$$

where $X_{i,n}$ are "related to" $\pi_{\theta_n}(\mathrm{d}x)$.

Outline

Biased stochastic approximation (1/2)

In many cases, the approximation is **unbiased**:

- at each iteration n, the exact quantity $G(\theta)$ is approximated by $\widehat{G}_n(\theta)$ such that
$$\mathbb{E}\left[\widehat{G}_n(\theta)|\mathcal{F}_n\right] - G(\theta) = 0.$$
- for example, in a Monte Carlo sum, if

  conditionnally to the past $X_{1,n}, \cdots, X_{i,n} \cdots$ i.i.d. $\pi_{\theta_n}(\mathrm{d}x)$

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

Biased stochastic approximation (2/2)

Nevertheless, in many Statistical Learning problems, the approximation is biased.

For example, when the approximation relies on a Monte Carlo sum,

- exact sampling under $\pi_{\theta_n}(\mathrm{d}x)$ is not possible
- Markov chain Monte Carlo sampling is possible: conditionnally to the past, $X_{1,n}, \cdots, X_{i,n} \cdots$ is a Markov chain with stationary distribution $\pi_{\theta_n}(\mathrm{d}x)$.
- We have:

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

What about MCMC-based Stochastic approximations ? (1/2)

- 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

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

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

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

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.

What about MCMC-based Stochastic approximations ? (2/2)

Difficulties:

- The bias may vanish to zero when $m_n \to 0$ (prohibitive computational cost).
- If $m_n = m$ (constant Monte Carlo number of points), then the bias **can not vanish**.
- In both cases, the control of the stochastic error depends on the current value θ_n - which is stochastic, and not necessarily "bounded".

Example (stochastic gradient):

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \frac{1}{m_{n+1}} \sum_{i=1}^{m_{n+1}} H(X_{i,n}, \theta_n)$$

$$= \theta_n - \gamma_{n+1} \int H(x, \theta_n)\, \pi_{\theta_n}(\mathrm{d}x) + \gamma_{n+1} \eta_{n+1}$$

with

$$\mathbb{E}\left[\eta_{n+1} | \mathcal{F}_n\right] \neq 0, \qquad |\eta_{n+1}| \not\to 0, \qquad \mathbb{E}\left[|\eta_{n+1}|^p \, | \mathcal{F}_n\right] \not\to 0.$$

Outline

The problem

Problem:

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

when

- $\theta \in \Theta \subseteq \mathbb{R}^d$
- the function g convex non-smooth nonnegative function (explicit)

The problem

Problem:

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

when

- $\theta \in \Theta \subseteq \mathbb{R}^d$
- the function g convex non-smooth nonnegative function (explicit)
- the function f is
  - · not necessarily convex,
  - · C^1 and ∇f is L-Lipschitz

$$\exists L > 0,\ \forall \theta, \theta' \qquad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\|.$$

  - · with an untractable gradient of the form

$$\nabla f(\theta) = \int H_\theta(x)\, \pi_\theta(\mathrm{d}x);$$

which can be approximated by **biased** Monte Carlo techniques.

The Proximal-Gradient algorithm (1/2)

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

The Proximal Gradient algorithm

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

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

where

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

Proximal map: Moreau(1962)

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

The Proximal-Gradient algorithm (1/2)

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

The Proximal Gradient algorithm

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

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

where

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

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 function.
- A MM/Majorize-Minimize algorithm from a quadratic majorization of f (since Lipschitz gradient) which produces a sequence $\{\theta_n, n \geq 0\}$ such that

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

The proximal-gradient algorithm (2/2)

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

The Proximal Gradient algorithm

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

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

where

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

About the Prox-step:

- when $g = 0$: $\operatorname{Prox}(\tau) = \tau$
- when g is the $\{0, +\infty\}$-valued indicator fct of a closed set: the algorithm is the projected gradient.
- in some cases, Prox is explicit (e.g. elastic net penalty). Otherwise, numerical approximation:

$$\theta_{n+1} = \operatorname{Prox}_{\gamma_{n+1},g}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n)\right) + \epsilon_{n+1} \qquad \text{in this talk, } \epsilon_{n+1} = 0$$

The 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* $\nabla f(\theta_n)$.

Monte Carlo-Proximal Gradient algorithm

In the case:

$$\nabla f(\theta) = \int H_\theta(x)\, \pi_\theta(x) \mu(\mathrm{d}x),$$

The MC-Proximal Gradient algorithm

Choose a stepsize sequence $\{\gamma_n, n \geq 0\}$ *and a batch size sequence* $\{m_n, n \geq 0\}$.

Given the current value θ_n,

1. *Sample a Markov chain* $\{X_{j,n}, j \geq 0\}$ *from a MCMC sampler with kernel* $P_{\theta_n}(x, \mathrm{d}x')$, *and unique invariant distribution* $\pi_{\theta_n}\, \mathrm{d}\mu$.
2. *Set*

$$H_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}).$$

3. *Update the value of the parameter*

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

Stochastic Approximation-Proximal Gradient algorithm

If in addition,

$$H_\theta(x) = \Phi(\theta) + \Psi(\theta) S(x)$$

which implies

$$\nabla f(\theta) = \Phi(\theta) + \Psi(\theta) \left(\int S(x)\, \pi_\theta(x) \mu(\mathrm{d}x) \right),$$

The SA-Proximal Gradient algorithm

Choose two stepsize sequences $\{\gamma_n, \delta_n, n \geq 0\}$ *and a batch size sequence* $\{m_n, n \geq 0\}$

Given the current value θ_n*,*

1. *Sample a Markov chain* $\{X_{j,n}, j \geq 0\}$ *from a MCMC sampler with kernel* $P_{\theta_n}(x, \mathrm{d}x')$*, and unique invariant distribution* $\pi_{\theta_n}\, \mathrm{d}\mu$*.*
2. *Set* $H_{n+1} = \Phi(\theta_n) + \Psi(\theta_n) S_{n+1}$ *with*

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

3. *Update the value of the parameter*

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

Design "parameters"

- Stepsize γ_n: constant or not ?
- Monte Carlo batch size m_n: constant or increasing (computational cost) ?
- Ergodicity of the MCMC sampler

(*) Penalized Expectation-Maximization (EM) vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in latent variable models.
- (Stochastic) EM algorithms

$$\tau_{n+1} = \operatorname{argmax}_\theta \int \log p_\theta(x)\ \pi_\theta(x)\, \mathrm{d}\mu(x) = \operatorname{argmax}_\theta \left\{ A(\theta) + \langle B(\theta), S_{n+1} \rangle \right\}$$

with

$$S_{n+1} = \int S(x)\ \pi_{\tau_n}(x)\, \mathrm{d}\mu(x) \qquad \text{EM}$$

$$S_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Monte Carlo EM} \quad \text{Wei and Tanner (1990)}$$

$$S_{n+1} = (1-\delta_{n+1}) S_n + \frac{\delta_{n+1}}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Stoch. Approx. EM} \quad \text{Delyon et al. (1999)}$$

(*) Penalized Expectation-Maximization (EM) vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in latent variable models.
- Generalized (Stochastic) EM algorithms

$$\tau_{n+1} = \operatorname{argmax}_\theta \int \log p_\theta(x)\ \pi_\theta(x)\, \mathrm{d}\mu(x) = \operatorname{argmax}_\theta \left\{ A(\theta) + \langle B(\theta), S_{n+1} \rangle \right\}$$

$$A(\tau_{n+1}) + \langle B(\tau_{n+1}), S_{n+1} \rangle \geq A(\tau_n) + \langle B(\tau_n), S_{n+1} \rangle$$

with

$$S_{n+1} = \int S(x)\ \pi_{\tau_n}(x)\, \mathrm{d}\mu(x) \qquad \text{EM}$$

$$S_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Monte Carlo EM} \quad \text{Wei and Tanner (1990)}$$

$$S_{n+1} = (1-\delta_{n+1}) S_n + \frac{\delta_{n+1}}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Stoch. Approx. EM} \quad \text{Delyon et al. (1999)}$$

(*) Penalized Expectation-Maximization (EM) vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in latent variable models.
- Generalized Penalized (Stochastic) EM algorithms

$$\tau_{n+1} = \operatorname{argmax}_\theta \int \log p_\theta(x)\ \pi_\theta(x)\, \mathrm{d}\mu(x) = \operatorname{argmax}_\theta \left\{ A(\theta) + \langle B(\theta), S_{n+1} \rangle \right\}$$

$$A(\tau_{n+1}) + \langle B(\tau_{n+1}), S_{n+1} \rangle - g(\tau_{n+1}) \geq A(\tau_n) + \langle B(\tau_n), S_{n+1} \rangle - g(\tau_n)$$

with

$$S_{n+1} = \int S(x)\ \pi_{\tau_n}(x)\, \mathrm{d}\mu(x) \qquad \text{EM}$$

$$S_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Monte Carlo EM} \quad \text{Wei and Tanner (1990)}$$

$$S_{n+1} = (1 - \delta_{n+1}) S_n + \frac{\delta_{n+1}}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Stoch. Approx. EM} \quad \text{Delyon et al. (1999)}$$

(*) Penalized Expectation-Maximization (EM) vs Proximal-Gradient

- EM Dempster et al. (1977) is a Majorize-Minimize algorithm for the computation of the ML estimate in latent variable models.
- (Stochastic) EM algorithms

$$\tau_{n+1} = \operatorname{argmax}_\theta \int \log p_\theta(x)\ \pi_\theta(x)\, \mathrm{d}\mu(x) = \operatorname{argmax}_\theta \left\{ A(\theta) + \langle B(\theta), S_{n+1} \rangle \right\}$$

with

$$S_{n+1} = \int S(x)\ \pi_{\tau_n}(x)\, \mathrm{d}\mu(x) \qquad \text{EM}$$

$$S_{n+1} = \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Monte Carlo EM} \quad \text{Wei and Tanner (1990)}$$

$$S_{n+1} = (1-\delta_{n+1}) S_n + \frac{\delta_{n+1}}{m_{n+1}} \sum_{j=1}^{m_{n+1}} S(X_{j,n}) \qquad \text{Stoch. Approx. EM} \quad \text{Delyon et al. (1999)}$$

- MC-Prox Gdt and SA-Prox GDT are Generalized Penalized EM algorithms (in the convex case).

Outline

The assumptions

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

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 $\operatorname{argmin}_\Theta F$ is a non-empty subset of Θ.

Existing results in the literature

There exist results under (some of) the assumptions

i.i.d. Monte Carlo approx, $\inf_n \gamma_n > 0$, $\sum_n \|H_{n+1} - \nabla f(\theta_n)\| < \infty$,

i.e. results for

- unbiased sampling. Almost no conditions for the biased sampling, such as the MCMC one.
- non vanishing stepsize sequence $\{\gamma_n, n \geq 0\}$.
- increasing batch size: when H_{n+1} is a Monte Carlo sum i.e.

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

the assumptions imply that $\lim_n m_n = +\infty$ at some rate.

Combettes (2001) Elsevier Science.
Combettes-Wajs (2005) Multiscale Modeling and Simulation.
Combettes-Pesquet (2015, 2016) SIAM J. Optim, arXiv
Lin-Rosasco-Villa-Zhou (2015) arXiv
Rosasco-Villa-Vu (2014,2015) arXiv
Schmidt-Leroux-Bach (2011) NIPS

Convergence of the perturbed proximal gradient algorithm (1/3)

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

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

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 $\lim_n \theta_n = \theta_\star$.

Convergence of the perturbed proximal gradient algorithm (2/3)

This convergence result

- for the convex case: f and g are convex.
- is a deterministic result.
  Covered: deterministic and random approximations H_{n+1} of $\nabla f(\theta_n)$.

Proof / Convergence of the perturbed proximal gradient algorithm (3/3)

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.

Note: deterministic lemma, signed noise.

Convergence: when H_{n+1} is a Monte-Carlo approximation (1/3)

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

$$\sum_n \gamma_{n+1}\eta_{n+1} = \sum_n \gamma_{n+1} \left(\frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}) - \int H_{\theta_n}(x)\ \pi_{\theta_n}(\mathrm{d}x) \right)$$
$$= \sum_n \gamma_{n+1} \left(H_{n+1} - \nabla f(\theta_n) \right)$$

where

$$X_{j+1,n} | \text{past} \sim P_{\theta_n}(X_{j,n}, \cdot) \qquad \pi_\theta P_\theta = \pi_\theta;$$

- The RHS

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

Convergence: when H_{n+1} is a Monte-Carlo approximation (1/3)

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

$$\sum_n \gamma_{n+1}\eta_{n+1} = \sum_n \gamma_{n+1} \left(\frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{j,n}) - \int H_{\theta_n}(x)\, \pi_{\theta_n}(\mathrm{d}x) \right)$$
$$= \sum_n \gamma_{n+1} \left(H_{n+1} - \nabla f(\theta_n) \right)$$

where

$$X_{j+1,n} | \text{past} \sim P_{\theta_n}(X_{j,n}, \cdot) \qquad \pi_\theta P_\theta = \pi_\theta;$$

- The RHS

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

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

  Solution $\widehat{H}_\theta$ to the Poisson equation: $H_\theta - \pi_\theta H_\theta = \widehat{H}_\theta - P_\theta \widehat{H}_\theta$

  $H_{n+1} - \nabla f(\theta_n) =$ martingale increment + remainder

  Regularity in θ of $t \mapsto \widehat{H}_t$.

Convergence: when H_{n+1} is a Monte-Carlo approximation (2/3)

Increasing batch size: $\lim_n m_n = +\infty$

Conditions on the step sizes and batch sizes

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

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* $x \in \mathsf{X}$,

$$\sup_{\theta \in \Theta} \|P_\theta^n(x, \cdot) - \pi_\theta\|_{W^\ell} \leq C \rho^n W^\ell(x). \tag{1}$$

Condition on Θ: Θ *is bounded.*

Convergence: when H_{n+1} is a Monte-Carlo approximation (3/3)

Fixed batch size: $m_n = m$

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.

Rates of convergence (1/3) : the problem

For non negative weights a_k, find an upper bound of

$$\sum_{k=1}^{n} \frac{a_k}{\sum_{\ell=1}^{n} a_\ell} F(\theta_k) - \min F$$

It provides

- an upper bound for the cumulative regret ($a_k = 1$)
- an upper bound for an averaging strategy when F is convex since

$$F\left(\sum_{k=1}^{n} \frac{a_k}{\sum_{\ell=1}^{n} a_\ell} \theta_k\right) - \min F \leq \sum_{k=1}^{n} \frac{a_k}{\sum_{\ell=1}^{n} a_\ell} F(\theta_k) - \min F.$$

Rates of convergence (2/3): a deterministic control

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

For any $\theta_\star \in \operatorname{argmin}_\Theta F$,

$$\sum_{k=1}^{n} \frac{a_k}{A_n} F(\theta_k) - \min F \leq \frac{a_0}{2\gamma_0 A_n} \|\theta_0 - \theta_\star\|^2 + \frac{1}{2A_n} \sum_{k=1}^{n} \left(\frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right) \|\theta_{k-1} - \theta_\star\|^2 + \frac{1}{A_n} \sum_{k=1}^{n} a_k \gamma_k \|\eta_k\|^2 - \frac{1}{A_n} \sum_{k=1}^{n} a_k \langle \mathsf{T}_{k-1} - \theta_\star, \eta_k \rangle$$

where

$$A_n = \sum_{\ell=1}^{n} a_\ell, \qquad \eta_k = H_k - \nabla f(\theta_{k-1}), \qquad \mathsf{T}_k = \operatorname{Prox}_{\gamma_k, g}(\theta_{k-1} - \gamma_k \nabla f(\theta_{k-1})).$$

Rates (3/3): when H_{n+1} is a Monte Carlo approximation, bound in L^q

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

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

with fixed size of the batch and (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 $\gamma_n \sim 1/\sqrt{n}$.

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

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 !*

Acceleration (1)

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

Acceleration (2) Aujol-Dossal-F.-Moulines, work in progress

Perturbed Nesterov acceleration: some convergence results

Choose γ_n, m_n, t_n *s.t.*

$$\gamma_n \in (0, 1/L]\,, \qquad \lim_n \gamma_n t_n^2 = +\infty, \qquad \sum_n \gamma_n t_n (1 + \gamma_n t_n) \frac{1}{m_n} < \infty$$

Then there exists $\theta_\star \in \operatorname{argmin}_\Theta F$ *s.t* $\lim_n \theta_n = \theta_\star$.
In addition

$$F(\theta_{n+1}) - \min F = O\left(\frac{1}{\gamma_{n+1} t_n^2}\right)$$

Schmidt-Le Roux-Bach (2011); Dossal-Chambolle(2014); Aujol-Dossal(2015)

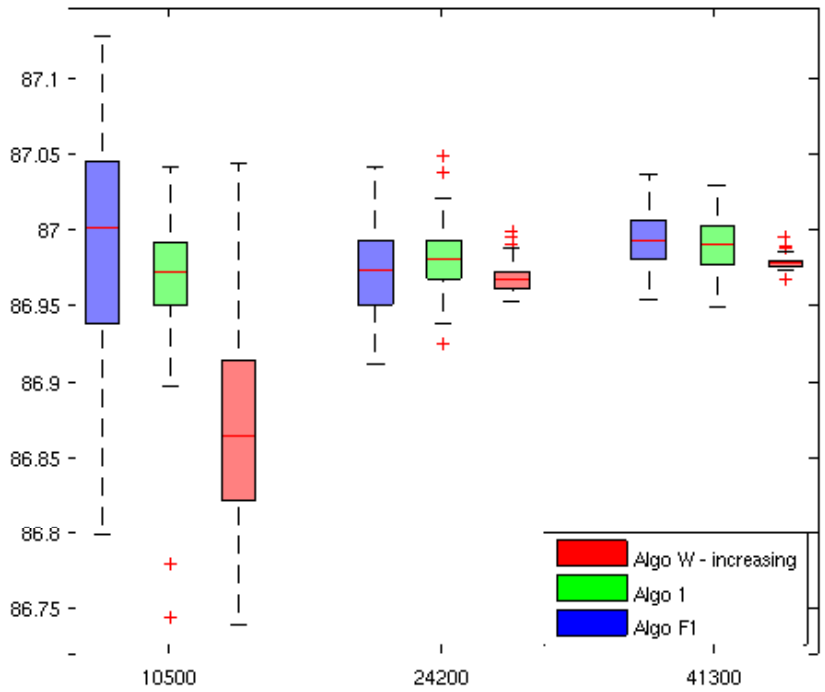
γ_n	m_n	t_n	rate	NbrMC
γ	n^3	n	n^{-2}	n^4
$\gamma/\sqrt{n}$	n^2	n	$n^{-3/2}$	n^3

Table: Control of $F(\theta_n) - \min F$

Outline

Conclusion (1/2): acceleration ?

- with or without the acceleration: complexity $O(1/\sqrt{n})$.
- acceleration: longer Markov chains, few iterations.



Conclusion (2/2): weaken the assumptions

- $\theta \in \mathbb{R}^d \to \theta$ in a Hilbert space
- Θ bounded $\to$ no boundedness condition on Θ
- f convex $\to$ f non convex