Perturbed Proximal Gradient Algorithm

Gersende FORT

LTCI, CNRS, Telecom ParisTech
Université Paris-Saclay, 75013, Paris, France

Large-scale inverse problems and optimization
Applications to image processing and astrophysics
Grenoble, November 2015

Works in collaboration with

- Eric Moulines (Professor, Ecole Polytechnique)
- Yves Atchadé (Assistant Professor, Univ. Michigan, USA)

and also Jean-Francois Aujol (IMB, Univ. Bordeaux), Charles Dossal (IMB, Univ. Bordeaux) and Soukaina Douissi.

↪ Y. Atchadé, G. Fort and E. Moulines. On Stochastic Proximal Gradient Algorithms. arXiv:1402:2365 math.ST.

Outline

Problem

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

where

- Θ finite-dimensional Euclidean space with scalar product $\langle , \rangle$ and norm $\| \cdot \|$
- the function f:$\Theta \to \mathbb{R}$ is a smooth function

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

  $$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L \, \|\theta - \theta'\|$$

- the function g: $\Theta \to (-\infty, \infty]$ is convex, not identically equal to $+\infty$, and lower semi-continuous

in the case $\nabla f(\theta)$ and f are intractable.

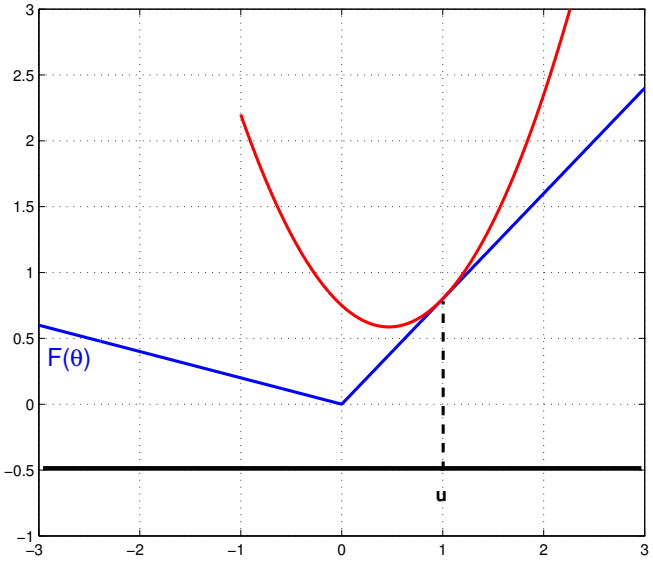
Classical algorithm when ∇f tractable (1/3)

Since ∇f is Lipschitz, for any u, θ,

$$f(\theta) \leq f(u) + \langle \nabla f(u), \theta - u \rangle + \frac{L}{2} \|\theta - u\|^2$$

which yields for any $L \leq \gamma^{-1}$

$$F(\theta) \leq f(u) + \langle \nabla f(u), \theta - u \rangle + \frac{1}{2\gamma} \|\theta - u\|^2 + g(\theta)$$



The RHS satisfies

- for fixed u, an upper bound of $\theta \mapsto F(\theta)$
- for $\theta = u$, this upper bound is equal to $F(u)$.
- for fixd u, it is convex (in θ)

$$C(u) + \frac{1}{2\gamma} \|\theta - \{u - \gamma \nabla f(u)\}\|^2 + g(\theta)$$

Classical algorithm when ∇f tractable (2/3)

Denote the upper bound by

$$Q_\gamma(\theta|u) \stackrel{\text{def}}{=} C(u) + \frac{1}{2\gamma}\|\theta - \{u - \gamma\nabla f(u)\}\|^2 + g(\theta)$$

$\hookrightarrow$ Majorization-Minimization (MM) algorithm

Define $\{\theta_n, n \geq 0\}$ iteratively by

$$\theta_{n+1} = \text{argmin}_\theta Q_\gamma(\theta|\theta_n)$$

or equivalently

$$\theta_{n+1} = \text{Prox}_\gamma(\theta_n - \gamma\nabla f(\theta_n))$$

with

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

also called Proximal-Gradient algorithm

Classical algorithm when ∇f tractable (3/3)

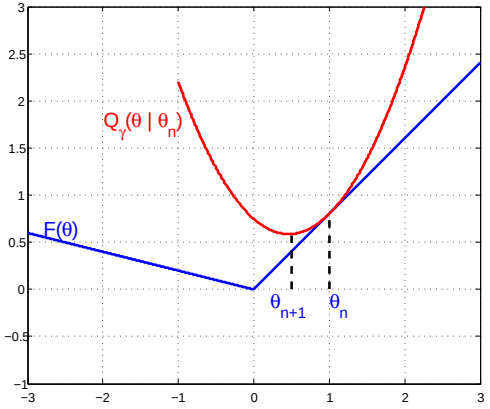
The sequence $\{\theta_n, n \geq 0\}$ is given by

$$\theta_{n+1} = \text{argmin}_\theta Q_\gamma(\theta|\theta_n)$$

where the upper bound $\theta \mapsto Q_\gamma(\theta|u)$ satisfies

$$F(\theta) \leq Q_\gamma(\theta|u) \qquad F(u) = Q_\gamma(u|u)$$

$\hookrightarrow$ Lyapunov function



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

since

$$\begin{aligned} F(\theta_{n+1}) &\leq Q_\gamma(\theta_{n+1}|\theta_n) \\ &\leq Q_\gamma(\theta_n|\theta_n) = F(\theta_n) \end{aligned}$$

Untractable proximal-gradient iteration

The exact *proximal-gradient* algorithm:

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

where $\{\gamma_n, n \geq 0\}$ is a step-size sequence in $(0, 1/L]$.

1. $\mathrm{Prox}_{\gamma}(u)$ can be untractable (not in this talk)
2. ∇f can be untractable (in this talk)

Untractable proximal-gradient iteration: explicit proximal operator

- (Projection on C)

  When $g(\theta) = \begin{cases} 0 & \text{if } \theta \in \mathsf{C} \\ +\infty & \text{otherwise} \end{cases}$ where C is closed convex,

$$\mathrm{Prox}_{\gamma}(\tau) = \min_{\theta \in \mathsf{C}} \|\tau - \theta\|^2$$

- (Elastic net penalty) $g(\theta) = \lambda \left(\frac{1-\alpha}{2} \|\theta\|_2^2 + \alpha \|\theta\|_1 \right)$

$$\left(\mathrm{Prox}_{\gamma}(\tau)\right)_i = \frac{1}{1+\gamma\lambda(1-\alpha)} \begin{cases} \tau_i - \gamma\lambda\alpha & \text{si } \tau_i \geq \gamma\lambda\alpha \\ \tau_i + \gamma\lambda\alpha & \text{si } \tau_i \leq -\gamma\lambda\alpha \\ 0 & \text{sinon} \end{cases}$$

  $\hookrightarrow$ proximal gradient algorithm = thresholded gradient algorithm

Untractable proximal-gradient iteration: untractable gradient ∇f

1. Unknown function f and its gradient is of the form

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

In this case

$$\nabla f(\theta) \approx \frac{1}{m} \sum_{k=1}^{m} H_\theta(X_k)$$

$\{X_k, k \geq 1\}$: (Online) Learning, Markov chain Monte Carlo.

Untractable proximal-gradient iteration: untractable gradient ∇f

1. Unknown function f and its gradient is of the form

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

In this case

$$\nabla f(\theta) \approx \frac{1}{m} \sum_{k=1}^{m} H_\theta(X_k)$$

$\{X_k, k \geq 1\}$: (Online) Learning, Markov chain Monte Carlo.

2. Large scale optimization

$$f(\theta) = \frac{1}{N} \sum_{k=1}^{N} f_k(\theta), \qquad \text{large } N$$

In this case

$$\nabla f(\theta) = \frac{1}{N} \sum_{k=1}^{N} \nabla f_k(\theta) \approx \frac{1}{m} \sum_{k=1}^{m} \nabla f_{I_k}(\theta)$$

In this talk

The exact *proximal-gradient* algorithm

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

The *perturbed proximal-gradient* algorithm

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

1. Which conditions on γ_n, η_n to ensure the convergence to the same limiting set as for the exact algorithm ?
2. When η_n is a (random) Monte Carlo approximation, which conditions on γ_n, m_n ?

Outline

Assumptions

$$(\arg)\min_{\theta\in\Theta} F(\theta) \qquad F(\theta) = f(\theta) + g(\theta)$$

1. the function g: $\Theta \to (-\infty, \infty]$ is convex, not identically equal to $+\infty$, and lower semi-continuous.
2. the function f: $\Theta \to \mathbb{R}$ is continuously differentiable and there exists $L > 0$ such that

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L \, \|\theta - \theta'\|$$

3. the function f is convex and the set $\mathcal{L} \stackrel{\text{def}}{=} \operatorname{argmin}_\theta F(\theta)$ is not empty.
4. the stepsize $\{\gamma_n, n \geq 0\}$ is positive and $\gamma_n \in (0, 1/L]$.

The algorithm

Stable sequence

Let $\mathcal{K} \subset \text{int}(\text{Dom}(\text{g}))$ be a compact subset of Θ such that $\mathcal{K} \cap \mathcal{L} \neq \emptyset$.

Algorithm:

$$\tilde{\theta}_{n+1} = \text{Prox}_{\gamma_{n+1}} \left(\theta_n - \gamma_{n+1} \nabla f(\theta_n) - \gamma_{n+1} \eta_{n+1}\right)$$
$$\theta_{n+1} = \text{Proj}_{\mathcal{K}}(\tilde{\theta}_{n+1})$$

Weighted average sequence

Let $\{a_n, n \geq 0\}$ be a non-negative sequence.

$$\bar{\theta}_n = \frac{1}{\sum_{k=1}^{n} a_k} \sum_{k=1}^{n} a_k \tilde{\theta}_k$$

Convergence of $\{\theta_n, n \geq 0\}$

$$\theta_{n+1} = \mathrm{Proj}_{\mathcal{K}}(\tilde{\theta}_{n+1}) \qquad \tilde{\theta}_{n+1} = \mathrm{Prox}_{\gamma_{n+1}}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n) - \gamma_{n+1}\eta_{n+1}\right)$$

Theorem (Atchadé, F., Moulines (2015))

If assumptions 1 *to* 4, $\sum_n \gamma_n = +\infty$ *and*

$$\sum_n \gamma_{n+1}\eta_{n+1} < \infty$$

$$\sum_n \gamma_{n+1}\left\langle T_{\gamma_{n+1}}(\theta_n), \eta_{n+1}\right\rangle < \infty \qquad \text{where } T_\gamma(\theta) = \mathrm{Prox}_\gamma(\theta - \gamma\nabla f(\theta))$$

$$\sum_n \gamma_{n+1}^2 \left\|\eta_{n+1}\right\|^2 < \infty$$

there exists $\theta_\star \in \mathcal{L} \cap \mathcal{K}$ *such that*

$$\lim_n \theta_n = \lim_n \tilde{\theta}_n = \theta_\star$$

Includes the convergence analysis for the exact algorithm ($\eta_n = 0$) Beck and Teboulle (2009); improves previous results Combettes and Wajs (2005); Combettes and Pesquet (2014).

Rates of convergence for $\{F(\theta_n), n \geq 0\}$

$$\theta_{n+1} = \mathrm{Proj}_{\mathcal{K}}(\tilde{\theta}_{n+1}) \qquad \tilde{\theta}_{n+1} = \mathrm{Prox}_{\gamma_{n+1}}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n) - \gamma_{n+1}\eta_{n+1}\right)$$

Theorem (Atchadé, F., Moulines (2015))

If assumptions 1 *to* 4*, for any* $a_k \geq 0$

$$\sum_{k=1}^{n} a_k \left\{F(\tilde{\theta}_k) - \min F\right\} \leq U_n$$

with

$$U_n \stackrel{\mathrm{def}}{=} \frac{1}{2}\sum_{k=1}^{n}\left(\frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}}\right)\|\theta_{k-1} - \theta_\star\|^2 + \frac{a_0}{2\gamma_0}\|\theta_0 - \theta_\star\|^2$$
$$- \sum_{k=1}^{n} a_k \langle T_{\gamma_k}(\theta_{k-1}) - \theta_\star, \eta_k\rangle + \sum_{k=1}^{n} a_k\gamma_k\|\eta_k\|^2.$$

Includes the convergence analysis for the exact algorithm ($\eta_{n+1} = 0$); Extends previous results in the case $\gamma_n = \gamma$, $a_n = 1$ Schmidt, Le Roux, Bach (2011) where it is assumed $\sum_n \|\eta_n\| < \infty$.

Outline

Monte Carlo approximation of the gradient

Assume that $\nabla f(\theta)$ is of the form

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

Consider a Monte Carlo perturbation

$$\eta_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1}^{(k)}) - \nabla f(\theta_n)$$

Monte Carlo approximation of the gradient

Assume that $\nabla f(\theta)$ is of the form

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

Consider a Monte Carlo perturbation

$$\eta_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1}^{(k)}) - \nabla f(\theta_n)$$

which includes the cases

1. $\{X_{n+1}^{(1)}, \cdots, X_{n+1}^{(m_{n+1})}\}$ are i.i.d. with distribution π_{θ_n}:

$$\mathbb{E}\left[\eta_{n+1} | \mathrm{Past}_n\right] = 0 \qquad \text{(unbiased approximation)}$$

Monte Carlo approximation of the gradient

Assume that $\nabla f(\theta)$ is of the form

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

Consider a Monte Carlo perturbation

$$\eta_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1}^{(k)}) - \nabla f(\theta_n)$$

which includes the cases

1. $\{X_{n+1}^{(1)}, \cdots, X_{n+1}^{(m_{n+1})}\}$ are i.i.d. with distribution π_{θ_n}:

$$\mathbb{E}\left[\eta_{n+1} | \mathrm{Past}_n\right] = 0 \qquad \text{(unbiased approximation)}$$

2. $\{X_{n+1}^{(1)}, \cdots, X_{n+1}^{(m_{n+1})}\}$ is a non-stationary Markov chain (e.g. MCMC path) with invariant distribution π_{θ_n}:

$$\mathbb{E}\left[\eta_{n+1} | \mathrm{Past}_n\right] \neq 0 \qquad \text{(biased approximation)}$$

Additional assumptions

5. the error is of the form

$$\eta_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1}^{(k)}) - \nabla f(\theta_n) \qquad \text{where } \nabla f(\theta) = \int H_\theta(x)\, \pi_\theta(\mathrm{d}x)$$

6. $\{X_{n+1}^{(k)}, k \geq 0\}$ is a Markov chain with transition kernel P_{θ_n}. For all θ, π_θ is invariant for P_θ.

7. The kernels $\{P_\theta, \theta \in \Theta\}$ are geometrically ergodic uniformly-in-θ (aperiodic, phi-irreducible, uniform-in-θ geometric drift inequalities w.r.t. W^p where $p \geq 2$, level sets of W^p are small): there exists $p \geq 2$ and for any $\ell \in (0, p]$, there exist $C > 0$, $\rho \in (0, 1)$ s.t.

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

↪ Trivial condition in the i.i.d. case
↪ There exist many sufficient conditions for the Markov case when samples are drawn from MCMC samplers.

Convergence of θ_n when $m_n \to \infty$

$$\eta_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1}^{(k)}) - \nabla f(\theta_n)$$

Theorem (Atchadé, F., Moulines (2015))

Assume Assumption 1 *to* 7 *and* $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_{n+1}^2 m_{n+1}^{-1} < \infty$.
If the approximation is biased, assume also: $\sum_n \gamma_{n+1} m_{n+1}^{-1} < \infty$.

With probability one, there exists $\theta_\star \in \mathcal{L} \cap \mathcal{K}$ *such that*

$$\lim_n \theta_n = \lim_n \tilde{\theta}_n = \theta_\star.$$

Convergence of θ_n when $m_n \to \infty$

$$\eta_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1}^{(k)}) - \nabla f(\theta_n)$$

Theorem (Atchadé, F., Moulines (2015))

Assume Assumption 1 *to* 7 *and* $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_{n+1}^2 m_{n+1}^{-1} < \infty$.
If the approximation is biased, assume also: $\sum_n \gamma_{n+1} m_{n+1}^{-1} < \infty$.

With probability one, there exists $\theta_\star \in \mathcal{L} \cap \mathcal{K}$ *such that*

$$\lim_n \theta_n = \lim_n \tilde{\theta}_n = \theta_\star.$$

The key ingredient for the proof is the control F. and Moulines (2003) for $p \geq 2$: w.p.1

$$\|\mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right]\| \leq C\, m_{n+1}^{-1} W(X_n^{(m_n)}),$$
$$\mathbb{E}\left[\|\eta_{n+1}\|^p|\mathcal{F}_n\right] \leq C\, m_{n+1}^{-p/2} W^p(X_n^{(m_n)}).$$

and the decomposition

$$\eta_{n+1} = \eta_{n+1} - \mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right] + \mathbb{E}\left[\eta_{n+1}|\mathcal{F}_n\right] = \text{Martingale Increment} + \text{Bias}$$

Convergence of θ_n when $m_n = m$

$$\eta_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1}^{(k)}) - \nabla f(\theta_n)$$

Theorem (Atchadé, F., Moulines (2015))

Assume Assumption 1 *to* 7 *and* $\sum_n \gamma_{n+1} = +\infty$, $\sum_n \gamma_{n+1}^2 < \infty$.
If the approximation is biased, assume also:

- *there exists a constant* C *such that for any* $\theta, \theta' \in \mathcal{K}$

$$\|H_\theta - H_{\theta'}\|_W + |||P_\theta - P_{\theta'}|||_W + |||\pi_\theta - \pi_{\theta'}|||_W \leq C \|\theta - \theta'\|.$$

- $\sup_{\gamma \in (0, 1/L]} \sup_{\theta \in \mathcal{K}} \gamma^{-1} \|\mathrm{Prox}_\gamma(\theta) - \theta\| < \infty$.
- $\sum_n |\gamma_{n+1} - \gamma_n| < \infty$.

With probability one, there exists $\theta_\star \in \mathcal{L} \cap \mathcal{K}$ *such that*

$$\lim_n \theta_n = \lim_n \tilde{\theta}_n = \theta_\star.$$

Convergence of $F(\tilde{\theta}_n)$ when $m_n \to \infty$

Theorem (Atchadé, F., Moulines (2015))

Assume Assumption 1 *to* 7. *For any* $q \in (1, p/2]$, *there exists* $C > 0$ *s.t.*

$$\left\| \sum_{k=1}^{n} a_k \left\{ F(\tilde{\theta}_k) - \min F \right\} \right\|_{L^q}$$

$$\leq C \left(\frac{a_0}{\gamma_0} + \sum_{k=1}^{n} \left| \frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right| + \left(\sum_{k=1}^{n} a_k^2 m_{k+1}^{-1} \right)^{1/2} + \sum_{k=1}^{n} a_k (\gamma_k + \upsilon) m_{k+1}^{-1} \right)$$

and

$$\sum_{k=1}^{n} a_k \{ \mathbb{E}[F(\tilde{\theta}_k)] - \min F \}$$

$$\leq C \left(\frac{a_0}{\gamma_0} + \sum_{k=1}^{n} \left| \frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right| + \sum_{k=1}^{n} a_k (\gamma_k + \upsilon) m_k^{-1} \right),$$

where $\upsilon = 0$ *if the Monte-Carlo approximation is unbiased and* $\upsilon = 1$ *otherwise.*

Convergence of $F(\tilde{\theta}_n)$ when $m_n = m$

Theorem (Atchadé, F., Moulines (2015))

Assume Assumption 1 *to* 7. *For any* $q \in (1, p/2]$, *there exists* $C > 0$ *s.t.*

$$\left\| \sum_{k=1}^{n} a_k \left\{ F(\tilde{\theta}_k) - \min F \right\} \right\|_{L^q}$$

$$\leq C \left(\frac{a_0}{\gamma_0} + \sum_{k=1}^{n} \left| \frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right| + \left(\sum_{k=1}^{n} a_k^2 \right)^{1/2} + \sum_{k=1}^{n} a_k \gamma_k + \upsilon \sum_{k=1}^{n} |a_{k+1} - a_k| \right)$$

and

$$\sum_{k=1}^{n} a_k \{ \mathbb{E}[F(\tilde{\theta}_k)] - \min F \}$$

$$\leq C \left(\frac{a_0}{\gamma_0} + \sum_{k=1}^{n} \left| \frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right| + \sum_{k=1}^{n} a_k \gamma_k + \upsilon \sum_{k=1}^{n} |a_{k+1} - a_k| \right)$$

where $\upsilon = 0$ *if the Monte-Carlo approximation is unbiased and* $\upsilon = 1$ *otherwise.*

Fixed or Increasing batch-size m_n ? Fixed or Decreasing step-size γ_n ?

Consider the L^q-convergence rate:

$$\left\| \left(\sum_{k=1}^{n} a_k \right)^{-1} \sum_{k=1}^{n} a_k F(\tilde{\theta}_k) - F(\theta_\star) \right\|_{L^q}$$

Increasing batch size $m_n \to \infty$: With $\gamma_n = \gamma$ $m_n \sim n$ $a_n = 1$,

Rate: $O(\ln n / n)$

Complexity: $O(\ln n / \sqrt{n})$.

Fixed batch size $m_n = m$ With $\gamma_n \sim \gamma_\star / \sqrt{n}$ $a_n = 1$ or $a_n = \gamma_n$,

Rate: $O(1/\sqrt{n})$

Complexity: $O(1/\sqrt{n})$.

Outline

Conclusion

- Contributions:
  a) NOT in the "strongly convex" case.
  b) Sufficient conditions for the convergence of perturbed Proximal-Gradient algorithms.
  c) Case of Monte Carlo approximations, biased or unbiased, increasing or fixed batch-size.

- Major contributions
  a) for Monte Carlo approximations
  b) biased approximations
  c) fixed batch-size

Other results, Works in progress and Future works

a) When f is not convex.

b) Accelerations (Nesterov, $\cdots$)

c) Convergence of the **Prox**imal **S**tochastic **A**pproximation **E**xpectation **M**aximization algorithm : for the maximization of a penalized likelihood in latent models by using a generalization of the SAEM algorithm.

d) Rates of convergence $\rightarrow$ explicit controls.