When Monte Carlo and Optimization met in a Markovian dance

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse, France

ICTS "Advances in Applied Probability", Bengaluru, August 2019.

Part IV - talk:
Stochastic Proximal-Gradient based algorithms: is Nesterov acceleration efficient ?

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)
- Jean-François Aujol (Univ. de Bordeaux, France)
- Charles Dossal (Univ. de Toulouse, 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 (Stat. and Computing, 2018)
- Stochastic FISTA algorithms : so fast ? (IEEE workshop SSP, 2018)
- Rates of convergence of perturbed FISTA-based algorithms (arXiv 2019)

This talk : solve a computational issue

- Find

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

where
- $g : \mathbb{R}^p \to (0, +\infty]$ is **not smooth**, but is **convex** and proper, lower semi-continuous
- the set $\Theta \subseteq \mathbb{R}^p$ *(extension to any Hilbert possible; not done)* is defined by: $\Theta = \{g < \infty\}$
- $f : \Theta \to \mathbb{R}$ is **not explicit / intractable**, ∇f exists but is **not explicit / intractable**

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

Motivations: example 1

$$\theta_* \in \text{argmin}_{\theta \in \Theta} \left(f(\theta) + g(\theta)\right)$$

- Large scale learning

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

and g is a regularization on the parameter θ.

- Intractability comes from the large value of N.

- Key:

$$\nabla f(\theta) = N \mathbb{E}\left[f_I(\theta)\right] \qquad I \text{ unif. on } \{1, \cdots, N\},$$

$\rightarrow$ Monte Carlo approximation $\rightarrow$ sampling distribution indep. of θ.

Motivations: example 2

$$\theta_* \in \text{argmin}_{\theta \in \Theta} \left(f(\theta) + g(\theta) \right)$$

- Inference in latent variable model <see Lecture 1>

$$f(\theta) = -\log \int p(Y_{1:N}, x; \theta) \; \mathrm{d}\nu(x)$$

and g is a regularization on the parameter θ.

- Intractability comes from the non explicit integral.

- Key:

$$\nabla f(\theta) = -\int \partial_\theta \left(\log p(Y_{1:N}, x; \theta) \right) \; \mathrm{d}\pi_\theta(x)$$

$\rightarrow$ Monte Carlo approximation $\rightarrow$ sampling distribution is the a posteriori distribution of x given $Y_{1:N}$ and depends on θ.

- Generally, f is not convex.

Motivations: example 3

$$\theta_* \in \text{argmin}_{\theta \in \Theta} \left(f(\theta) + g(\theta) \right)$$

- Binary graphical models: $Y^{(n)} \in \{0,1\}^p$ i.i.d. so that the negative log-likelihood

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

and g is a regularization on the parameter θ.

- Intractability comes from the non explicit normalizing constant Z_θ.

- Key:

$$\nabla f(\theta) = \sum_{x \in \{0,1\}^p} H(\theta, x) \; \pi_\theta(x) \qquad \pi_\theta(x) = \frac{1}{Z_\theta} \exp\left(\sum_{i=1}^{p} \theta_i x_i + \sum_{1 \leq i < j \leq p} \theta_{ij} 1_{x_i = x_j} \right)$$

$\to$ Monte Carlo approximation $\to$ sampling distribution depends on θ and is known up to a normalization constant.

- Here, f is convex.

If ∇f were available: a numerical solution (1/2)

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

- Assumptions:

- the function $g : \mathbb{R}^p \to (0, +\infty]$ is convex, proper, lower semi-continuous
- set $\Theta = \{g < \infty\}$
- the function $f : \Theta \to \mathbb{R}$ is C^1, with **Lipschitz gradient** (of constant L)

- The proximal operator (Moreau, 1962) : given $\gamma > 0$:

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

- well defined under the assumptions on g
- when $g = 0$, $\mathrm{Prox}_{\gamma,g}(\tau) = \tau$
- when g is the indicator function of a closed set, it is the projection
- computation explicit, or not. In this talk: assumed explicit.

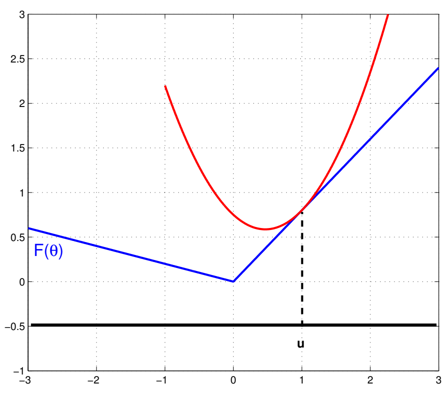
If ∇f were available: a numerical solution (2/2)

- **The proximal-gradient (PG) algorithm** Given a sequence of positive step sizes $\{\gamma_t\}_t$, it is defined by

$$\theta_{t+1} = \text{Prox}_{\gamma_{t+1},g}\left(\theta_t - \gamma_{t+1}\nabla f(\theta_t)\right)$$

Beck-Teboulle, 2010; Combettes-Pesquet, 2011; Parikh-Boyd,2013

- It is a Majorize-Minorization algorithm:



For any $\gamma \in (0, 1/L)$,

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

the minimization of the RHS is the computation of $\text{Prox}_{\gamma,g}(\theta_t - \gamma\nabla f(\theta_t))$

It holds: $(f+g)(\theta_{t+1}) \leq (f+g)(\theta_t)$

Perturbed PG

Prox-Gdt: $\theta_{t+1} = \mathsf{Prox}_{\gamma_{t+1},g}\left(\theta_t - \gamma_{t+1}\nabla f(\theta_t)\right)$

- When the gradient is intractable, a natural idea

$$\theta_{t+1} = \mathsf{Prox}_{\gamma_{t+1},g}\left(\theta_t - \gamma_{t+1}\widehat{\nabla f(\theta_t)}\right)$$

- When the gradient is an expectation: $\widehat{\nabla f(\theta_t)}$ can rely on a Monte Carlo approximation

- Questions:

- Suff cond on the approximation so that this perturbed algorithm inherits the behavior of the (exact) PG.
- Rate of convergence
- Implementation issues in the Monte Carlo case.

Stability result

$$\theta_{t+1} = \mathsf{Prox}_{\gamma_{t+1},g}\left(\theta_t - \gamma_{t+1}\widehat{\nabla f(\theta_t)}\right)$$

- (F.-Moulines, 2020; work in progress)

Under conditions essentially of the form of those on the following slide, it can be proved that the Chen's technique provides a self-stabilized perturbed proximal-gradient algorithm.

Convergence result

$$\theta_{t+1} = \mathsf{Prox}_{\gamma_{t+1},g}\left(\theta_t - \gamma_{t+1}\widehat{\nabla f(\theta_t)}\right)$$

Set $\mathcal{L} := \mathrm{argmin}_\Theta(f+g)$ $\qquad \eta_{t+1} := \widehat{\nabla f(\theta_t)} - \nabla f(\theta_t).$

●(Atchadé-F.-Moulines, 2017) Assume

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

$$\sum_t \gamma_{t+1}^2 \|\eta_{t+1}\|^2, \qquad \sum_t \gamma_{t+1}\eta_{t+1}, \qquad \sum_t \gamma_{t+1}\left\langle \mathsf{A}_t, \eta_{t+1}\right\rangle$$

where $\mathsf{A}_t = \mathsf{Prox}_{\gamma_{t+1},g}(\theta_t - \gamma_{t+1}\nabla f(\theta_t))$.

Then there exists $\theta_\star \in \mathcal{L}$ such that $\lim_t \theta_t = \theta_\star$.

● It is a deterministic result. Holds also "a.s." in the case of stochastic approximations of the gradient.

Sketch of proof

The proof relies on

- a Lyapunov inequality - which uses the convexity of f and g

$$\|\theta_{t+1}-\theta_\star\|^2 \leq \|\theta_t-\theta_\star\|^2 - \underbrace{2\gamma_{t+1}\left((f+g)(\theta_{t+1}) - \min(f+g)\right)}_{\text{non-negative}} \underbrace{-2\gamma_{t+1}\left\langle \mathsf{A}_t-\theta_\star ; \eta_{t+1}\right\rangle + 2\gamma_{t+1}^2\|\eta_{t+1}\|^2}_{\text{signed noise}}$$

- (an extension of) the Robbins-Siegmund lemma:

Let $\{v_t\}_t$ and $\{\chi_t\}_t$ be non-negative sequences and $\{\xi_t\}_t$ be such that $\sum_t \xi_t$ exists. If for any $t \geq 0$,

$$v_{t+1} \leq v_t - \chi_{t+1} + \xi_{t+1}$$

then $\sum_t \chi_t < \infty$ and $\lim_t v_t$ exists.

Note: deterministic lemma, signed noise.

What about Nesterov-based acceleration ?

Let $\{\lambda_t\}_t$ be a positive sequence s.t. $\gamma_{t+1}\lambda_t(\lambda_t - 1) \leq \gamma_t \lambda_{t-1}^2$.
Ex. $\gamma_t = \gamma$ and $\lambda_t = O(t)$.

- The algorithm: define the sequence $\{\theta_t\}_t$ by

$$\theta_{t+1} = \mathrm{Prox}_{\gamma_{t+1},g}\left(\tau_t - \gamma_{t+1}\nabla f(\tau_t)\right), \qquad \tau_{t+1} = \theta_{t+1} + \frac{\lambda_t - 1}{\lambda_{t+1}}\left(\theta_{t+1} - \theta_t\right)$$

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

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

- Known:

Proximal-gradient $\quad (f+g)(\theta_t) - \min(f+g) = O\left(\frac{1}{t}\right)$

Accelerated PG $\quad (f+g)(\theta_t) - \min(f+g) = O\left(\frac{1}{t^2}\right)$

- Do we have the same acceleration when replacing the gradient with an approximation ?

Convergence results for the perturbed Accelerated PG

●(F.-Risser-Atchadé-Moulines, 2018) Sufficient conditions on λ_t, γ_t and on the errors

$$\tilde{\eta}_{t+1} := \widehat{\nabla f(\tau_t)} - \nabla f(\tau_t)$$

so that:
- the limit $\lim_t \gamma_t \ \lambda_t^2 \ \left((f+g)(\theta_t) - \min(f+g)\right)$ exists.
- explicit upper bound for this quantity.

●(Aujol-Dollal-F.-Moulines, 2019) Sufficient conditions for the case

$$\gamma_t = \gamma, \qquad \qquad \lambda_t = O(t^d), d \in (0,1).$$

implying
- the limit $\lim_t \gamma_t \ \lambda_t^2 \ \left((f+g)(\theta_t) - \min(f+g)\right)$ exists.
- explicit upper bound for this quantity.
- convergence of the parameters $\{\theta_t\}_t$.

Case of Monte Carlo approximations of the gradient (1/6)

$$\nabla f(\theta_t) = \int H(\theta_t, x)\, \mathrm{d}\pi_{\theta_t}(x),$$

- Idea 1: sample points $X_{1,t+1}, \cdots, X_{m_{t+1},t+1}$ approximating $\mathrm{d}\pi_{\theta_t}$

$$\widehat{\nabla f(\theta_t)} := \frac{1}{m_{t+1}} \sum_{k=1}^{m_{t+1}} H(\theta_t, X_{k,t+1})$$

- Idea 2 when $H(\theta, x) = \phi(\theta) + \langle S(x); \psi(\theta) \rangle$, $\nabla f(\theta) = \phi(\theta) + \langle \int S \mathrm{d}\pi_\theta; \psi(\theta) \rangle$

$$\widehat{\nabla f(\theta_t)} := \phi(\theta_t) + \left\langle \tilde{S}_{t+1}; \psi(\theta_t) \right\rangle$$

where

$$\tilde{S}_{t+1} = \tilde{S}_t + \delta_{t+1} \left(\frac{1}{m_{t+1}} \sum_{k=1}^{m_{t+1}} H(\theta_t, X_{k,t+1}) - \tilde{S}_t \right)$$

for some positive "step size" δ_{t+1}. (see F.-Ollier-Samson, 2018)

- Hereafter: case of "idea 1".

Case of Monte Carlo approximation of the gradient (2/6)

• This is a again an intertwinings of Monte Carlo and Optimization: at each iteration

- sample points $X_{1,t+1}, \cdots, X_{m_{t+1},t+1}$ from a Markov chain converging to $\mathrm{d}\pi_{\theta_t}$.

- update the parameter

$$\theta_{t+1} = \mathrm{Prox}_{\gamma_{t+1},g}\left(\theta_t - \gamma_{t+1} \frac{1}{m_{t+1}} \sum_{j=1}^{m_{t+1}} H(\theta_t, X_{j,t+1})\right)$$

• We will see that we can have $m_t = m(= 1)$ ("SA rule") or $m_t \to \infty$ ("mini-batch rule").

Case of Monte Carlo approximation of the gradient (3/6)

- Conditions on the *design parameters* γ_t, m_t, λ_t, on the sampling mecanism, in order to observe, w.p.1., the convergence to a minimizer ?

- Is there a choice of the *design parameters* γ_t, m_t, λ_t to reach the same rate of convergence as the exact PG (and observe the benefit of the Nesterov acceleration ?) What about averaging strategy ?

- The answers will use:

$$\left| \mathbb{E}\left[\frac{1}{m_{t+1}} \sum_{i=1}^{m_{t+1}} H(\theta_t, X_{i,t+1}) \Big| \mathcal{F}_t \right] - \int H(x, \theta_t) \ \pi_{\theta_t}(\mathrm{d}x) \right| \leq \frac{C(\theta_t, X_{m_t,t})}{m_{t+1}}$$

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

These results depend on **ergodic properties** of the MCMC sampler at iteration t; and it is easier when the controls can be indep of θ_t (stability !!)

Case of Monte Carlo approximation of the gradient (4/6) - with $m_t \to \infty$

• For the almost-sure convergence of $\{\theta_t\}_t$ given by Perturbed-PG

- Conditions on m_t, γ_t:

$$\sum_t \gamma_t = +\infty, \qquad \sum_t \frac{\gamma_t^2}{m_t} < \infty; \qquad \sum_t \frac{\gamma_t}{m_t} < \infty$$

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

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

• Rate of cvg of the functional in L^q for the averaged sequence $\bar{\theta}_t := t^{-1} \sum_{k=1}^t \theta_k$:

$$\gamma_t = \gamma_\star, \quad m_t = O(t) \Rightarrow \text{rate of cvge } O(1/t)$$

Beware ! Rate after $O(t^2)$ Monte Carlo samples. Given a MC budget of $O(t)$, the rate is $O(1/\sqrt{t})$.

Case of Monte Carlo approximation of the gradient (5/6) - with $m_t = m$

- For the almost-sure convergence of $\{\theta_t\}_t$ given by Perturbed-PG

- Condition on the step size:

$$\sum_t \gamma_t = +\infty \qquad \sum_t \gamma_t^2 < \infty \qquad \sum_t |\gamma_{t+1} - \gamma_t| < \infty$$

- Condition on the Markov chain

same as in the case "increasing batch size" $+$ regularity-in-θ of the Poisson equation

- Condition on the Prox:

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

- Rate of cvg of the functional in L^q for the averaged sequence $\bar{\theta}_t := t^{-1} \sum_{k=1}^t \theta_k$:

$$\gamma_t = \gamma_\star t^{-a}, \ a \in [1/2, 1], \qquad m_t = m_\star \Longrightarrow \text{rate of cvge } O(1/\sqrt{t})$$

Rate after $O(t)$ Monte Carlo samples.

Case of Monte Carlo approximation of the gradient (6/6) - what about acceleration strategies ?

- F.-Risser-Atchadé-Moulines, 2018

$$\lim_t t^2 \quad ((f+g)(\theta_t) - \min(f+g)) < \infty \quad \text{a.s.}$$
$$\sup_t t^2 \;\; \mathbb{E}\left[(f+g)(\theta_t) - \min(f+g)\right] < \infty$$

with

$$\lambda_t = O(t), \qquad \gamma_t = \gamma \qquad m_t = O(t^3)$$

- Given a MC budget of $O(t)$:

- the rate is $O(1/\sqrt{t})$
- the same rate as the (perturbed) Proximal-Gradient with an averaging strategy.

- Other strategies $\lambda_t = O(t^d)$ for some $d \in (0,1)$: no improvements, still this "$O(1/\sqrt{t})$"

Conclusion

• the design paraemeters (+ the sampling mecanism of the Monte Carlo approx of the gradient) can be chosen in such a way that the stochastically perturbed algorithm inherits the same limiting behavior (convergence) as the exact algorithm.

• the design parameters can be chosen in such a way that the stochastically perturbed algorithm inherits the same rates of convergence as the exact algorithms (PG, accelerated PG).

• nevertheless, when taking into account the Monte Carlo computational cost: the stochastic algorithms **can not go** beyond the "$1/\sqrt{t}$" rate. All these results are obtained with Monte Carlo strategies:

m points in the Monte Carlo sum $\Rightarrow$ variance $O(1/m)$.

• Conclusions based on the asymptotic rate of cvg. What is the verdict of numerical analyses ?