Perturbed (accelerated) Proximal-Gradient algorithms

Gersende Fort

CNRS & Institut de Mathématiques de Toulouse
France

Works with Eric Moulines (Ecole Polytechnique, France); Yves Atchadé (Univ. Michigan, USA); J.F. Aujol (Univ. Bordeaux, France) and C. Dossal (INSA Toulouse, France)

$$\text{(arg)}\min_{\theta \in \mathbb{R}^p} \left(f(\theta) + g(\theta) \right)$$

with

- $g : \mathbb{R}^p \to [0, \infty]$ is convex, non smooth, not identically equal to $+\infty$, and lsc.
- $\mathrm{Prox}_{\gamma g}(\tau)$ is explicit
- f is smooth (gradient Lipschitz) with an untractable gradient

Algorithm: Perturbed Proximal-Gradient

$$\theta_{k+1} = \mathrm{Prox}_{\gamma_{k+1} g} \left(\theta_k - \gamma_{k+1} \widehat{\nabla f(\theta_k)} \right)$$

Questions: Conditions on γ_{k+1} and on $\widehat{\nabla f(\theta_k)} - \nabla f(\theta_k)$ to ensure the same limiting behavior as the Prox-Gdt algorithm ?

Furthermore, in the case

a) the gradient is an untractable expectation

$$\nabla f(\theta) = \int_{\mathsf{X}} \underbrace{H(\theta, x)}_{\text{explicit}} \ \underbrace{\pi_\theta(\mathrm{d}x)}_{\text{probability}}$$

b) Stochastic approximation to avoid curse of dimensionality

c) **i.i.d. Monte Carlo not possible/efficient** $\rightarrow$ Markov Chain MC (MCMC) sampling

Questions: Since MCMC provides a biased approximation

$$\nabla f(\theta_k) \approx \frac{1}{m_{k+1}} \sum_{j=1}^{m_{k+1}} H(\theta, X_{jk}) \qquad \mathbb{E}\left[\frac{1}{m_{k+1}} \sum_{j=1}^{m_{k+1}} H(\theta, X_{jk})\right] - \nabla f(\theta_k) \neq 0$$

where $\{X_{1k}, \cdots, X_{jk}, \cdots\}$ Markov chain with stationary distribution π_{θ_k}

- which conditions on γ_{k+1} and on the Monte Carlo batch size m_{k+1} ?
- is it possible to have a non vanishing bias i.e. $m_{k+1} = m$?

Perturbed Prox-Gdt + Acceleration:

$$\tau_k = \theta_k + \frac{t_{k-1} - 1}{t_k}(\theta_k - \theta_{k-1})$$
$$\theta_{k+1} = \mathrm{Prox}_{\gamma_{k+1} g}\left(\theta_k - \gamma_{k+1}\widehat{\nabla f(\tau_k)}\right)$$

Questions:

- Which sequences γ_k, t_k, among those satisfying

$$\gamma_{k+1} t_k (t_k - 1) \leq \gamma_k t_{k-1}^2$$

- When stochastic approx of the gradient: which Monte Carlo batch size m_k ?
- Is there a gain to consider $t_k = O(k^d)$ for some $0 \leq d \leq 1$?

Computational Statistics, Statistical Learning

- Online learning: here the "Monte Carlo points" are the examples/observations.
- Penalized Maximum Likelihood Estimation in a parametric model

$$\mathrm{argmin}_\theta \quad \underbrace{f(\theta)}_{\text{negative log-likelihood}} + \underbrace{g(\theta)}_{\text{penalty term}}$$

Example 1: Latent variable models

- The log-likelihood $\ell(\theta)$ of the n observations dependence upon the obs. is omitted

$$\ell(\theta) = \log \int_{\mathsf{X}} \underbrace{p(x,\theta)}_{\text{complete likelihood}} \mu(\mathrm{d}x)$$

Untractable integral

- Its gradient

$$\nabla \ell(\theta) = \int \partial_\theta \log p(x,\theta) \underbrace{\frac{p(x,\theta)}{\int p(u,\theta)\mu(\mathrm{d}u)} \mu(\mathrm{d}x)}_{\text{a posteriori distribution}}$$

Untractable integral since the normalizing constant unknown $\longrightarrow$ MCMC

Example 2: Binary graphical model

- N i.i.d. $\{0,1\}^p$ observations from the distribution

$$\pi_\theta(y_{1:p}) \propto \frac{1}{Z_\theta} \exp\left(\sum_{i=1}^p \theta_i y_i + \sum_{1 \leq i < j \leq p} \theta_{ij} \mathbb{1}_{y_i = y_j} \right)$$

- The log-likelihood of the obs. $Y^1, \cdots, Y^N$

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

- Its gradient

$$\nabla_{\theta_i} \ell(\theta) = \sum_{n=1}^N Y_i^n - \sum_{y_{1:p} \in \{0,1\}^p} y_i \pi_\theta(y)$$

$$\nabla_{\theta_{ij}} \ell(\theta) = \sum_{n=1}^N \mathbb{1}_{Y_i^n = Y_j^n} - \sum_{y_{1:p} \in \{0,1\}^p} \mathbb{1}_{y_i = y_j} \pi_\theta(y)$$

Results on Perturbed Prox-Gdt (1/2)

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

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

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*

$$\sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^2, \qquad \sum_n \gamma_{n+1}\eta_{n+1}, \qquad \sum_n \gamma_{n+1} \langle \mathbf{A}_n, \eta_{n+1} \rangle$$

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

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

It generalizes and improves on previous results. What can be said in the non-convex case (open question) and with non explicit “Prox” ?

Results on Perturbed Prox-Gdt (2/2)

Given non-negative weights $a_1, \cdots, a_n$, set $A_n \stackrel{\text{def}}{=} \sum_{k=1}^n a_k$

Theorem (Atchadé, F., Moulines)

For any $\theta_\star \in \operatorname{argmin}_\Theta(f+g)$,

$$\begin{aligned}(f+g)\left(\sum_{k=1}^n \frac{a_k}{A_n}\theta_k\right) - \min(f+g) &\leq \frac{a_0}{2\gamma_0 A_n}\|\theta_0 - \theta_\star\|^2 \\ &+ \frac{1}{2A_n}\sum_{k=1}^n \left(\frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}}\right)\|\theta_{k-1} - \theta_\star\|^2 \\ &+ \frac{1}{A_n}\sum_{k=1}^n a_k\gamma_k\|\eta_k\|^2 - \frac{1}{A_n}\sum_{k=1}^n a_k\left\langle \mathbf{A}_{k-1} - \theta_\star, \eta_k\right\rangle\end{aligned}$$

In the case of stochastic perturbation $\eta_k = \widehat{\nabla f(\theta_k)} - \nabla f(\theta_k)$: it yields bounds with high probability, in expectation, in L^q, $\cdots$

Stochastic Prox-Gdt, with (possibly) biased MC approximation

Under ergodic conditions on the MCMC samplers, we have

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

with

- Constant MC batch size $m_n = m$ (i.e. non vanishing approximation → technical proof)

$$u_n = \frac{1}{\sqrt{n}} \qquad \text{with}\, \gamma_n = \frac{\gamma_\star}{n^a}, a \in [1/2, 1]$$

- Increasing MC batch size

$$u_n = \frac{1}{n} \qquad \text{with}\, \gamma_n = \gamma_\star \quad m_n \propto n$$

Rate with a computational MC cost: $O(n^2)$.

Nesterov-based acceleration of the Stochastic Prox-Gdt alg

Convergence Choose γ_n, m_n, t_n s.t.

$$\gamma_n \in (0, 1/L], \qquad \gamma_{k+1} t_k (t_k - 1) \leq \gamma_k t_{k-1}^2$$
$$\lim_n \gamma_n t_n^2 = +\infty, \qquad \sum_n \gamma_n t_n (1 + \gamma_n t_n) \frac{1}{m_n} < \infty$$

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

Rate on F In addition

$$\mathbb{E}\left[F(\theta_{n+1}) - \min F\right] = O\left(u_n\right)$$

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

In all strategies: for a MC computational cost N, the rate is $1/\sqrt{N}$.

1. Variance reduction technique Here the variance of the MC approximation is $O(1/m_n)$. What happens when a "variance reduction" MC technique is used ?
2. Averaging Given non-negative weights $a_1, \cdots, a_n$, do γ_k, t_k, m_k exist such that

$$\sup_n a_n \ \left((f+g)(\theta_n) - \min(f+g)\right) < \infty$$

$$(f+g)\left(\sum_{k=1}^{n} \frac{a_k}{\sum_{j=1}^{n} a_j} \theta_k\right) - \min(f+g) = O\left(\frac{1}{\sum_{k=1}^{n} a_k}\right)$$

3. Maximal rate What is the maximal rate after n iterations ? after N Monte Carlo draws ?
4. (F)ISTA ? What about $t_n = O(n^d)$ for some $0 < d < 1$?

A first answer: With variance reduction MC techniques, Nesterov acceleration $(d = 1)$, $\gamma_k = \gamma$, $m_n = n^3$ and $a_n = n$: **after N MC draws, the rate is always better than $1/\sqrt{N}$**