Convergence of Perturbed Gradient-based methods for non-smooth convex optimization

Gersende Fort

LTCI, CNRS and Telecom ParisTech
Paris, France

Based on joint works with

- Yves Atchadé (Univ. Michigan, USA)
- Jean-François Aujol (Univ. Bordeaux, France)
- Charles Dossal (Univ. Bordeaux, France)
- Eric Moulines (Ecole Polytechnique, France)

↪ `On Perturbed Proximal-Gradient algorithms (2015, arXiv)`

Problem:

Convergence of an algorithm designed for solving

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

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 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: $\Theta = \{\theta : g(\theta) < \infty\}$.

when f and $\nabla f(\theta)$ are not explicit

Outline

The function g

- Can be evaluated, is convex but is not smooth
- Typically: a constraint in the optimization problem
  - ⋆ optimization restricted to a convex set $\mathcal{K}$

$$g(\theta) \in \{0, +\infty\} = \begin{cases} 0 & \text{if } \theta \in \mathcal{K} \\ +\infty & \text{otherwise} \end{cases}$$

  - ⋆ Sparsity constraints

$$g(\theta) \propto \|\theta\|_1 = \sum_{i=1}^{d} |\theta_i|$$

$$g(\theta) \propto \alpha \sum_{i=1}^{d} |\theta_i| + \frac{(1-\alpha)}{2} \sum_{i=1}^{d} \theta_i^2$$

The function f: Ex. 1, Inference in Latent variable models

- A vector of observations: Y
- A vector of latent variables: U
- A parametric model indexed by $\theta \in \Theta$

Minimize the negative log-likelihood:

$$f(\theta) = -\log \int p(\mathrm{Y}|\mathrm{u};\theta)\, \phi(\mathrm{u}) \mu(\mathsf{d}u)$$

which is (usually) intractable; same thing for the gradient

$$\nabla f(\theta) = -\int \nabla \log p(\mathrm{Y}|\mathrm{u};\theta) \; \frac{p(\mathrm{Y},\mathrm{u};\theta)}{\int p(\mathrm{Y},\mathrm{x};\theta)\mu(\mathsf{d}x)} \mu(\mathsf{d}u)$$

Example 1': (logistic regression, with random effects)

For example, logistic regression with random effects, under sparsity constraints

$$\mathbf{U} \sim \mathcal{N}_q(0, I)$$

$$Y_i | \mathbf{U} \overset{ind.}{\sim} \text{Ber}\left(\frac{\exp\left(x_i' \beta + \sigma\, z_i' \mathbf{U}\right)}{1 + \exp\left(x_i' \beta + \sigma\, z_i' \mathbf{U}\right)} \right)$$

$$\theta = (\beta, \sigma) \in \mathbb{R}^p \times \mathbb{R}_+$$

$$g(\theta) = \lambda \sum_{i=1}^{p} |\beta_i|$$

The function f: Ex. 2, Inference in Markov Random Fields

- Observations: i.i.d. $\mathbb{R}^p$-valued samples $Y_1, \cdots, Y_N$ from the distribution

$$\pi_\theta(y) = \frac{\gamma(y;\theta)}{Z_\theta}$$

  with an intractable normalizing constant Z_θ.
- A parametric model indexed by $\theta \in \mathbb{R}^d$.

Minimize the negative log-likelihood, which is intractable

$$f(\theta) = -\sum_{i=1}^{N} \log \gamma(Y_i;\theta) + N \log Z_\theta$$

and with intractable gradient

$$\nabla f(\theta) = -\sum_{i=1}^{N} \nabla_\theta \log \gamma(Y_i;\theta) + N \int \{\nabla_\theta \log \gamma(\mathrm{u},\theta)\} \, \pi_\theta(\mathrm{du})$$

The function f: Ex.3, Learning on huge data set

- Many component functions (ex. a cost function associated to each observation)

Minimize an additive cost function

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

intractable since N is large, and same thing for its gradient

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

The function f: Ex.4, Online learning

Minimize a mean value

$$f(\theta) = \int \bar{f}(\theta; \mathrm{u}) \pi(\mathrm{du})$$

when the distribution π is unknown, and only examples/samples from π are available online:

$$\nabla f(\theta) = \int \left\{ \nabla_\theta \bar{f}(\theta; \mathrm{u}) \right\} \; \pi(\mathrm{du})$$

Outline

Examples of problems of the form: $\text{argmin}_\theta\{f(\theta) + g(\theta)\}$

A first order method: the proximal gradient algorithm

Convergence of the Perturbed Proximal Gradient algorithm

Rates of convergence

Acceleration

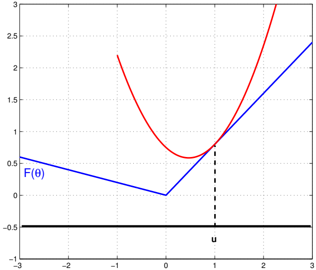
References

The proximal-gradient algorithm (1/2)

$$\mathrm{argmin}_{\theta \in \Theta} \left(\underbrace{f(\theta)}_{C^1 \text{ with Lipschitz gradient}} + \underbrace{g(\theta)}_{\text{not differentiable, convex}} \right)$$

Idea: majorization-minimization iterative method

Nesterov (2004)



We have:

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

The proximal-gradient algorithm (2/2)

- Since f is smooth: define a majorizing function $\theta \mapsto Q_\gamma(\theta; \theta_n)$

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

  where $L \leq 1/\gamma$.

- A family of majorizing functions - indexed by γ; all of them are equal to $F(\theta_n)$ at $\theta = \theta_n$
- Update the current solution

$$\theta_{n+1} = \mathrm{argmin}_\theta Q_\gamma(\theta; \theta_n) = \mathrm{argmin}_\theta \left(g(\theta) + \frac{1}{2\gamma} \|\theta - \{\theta_n - \gamma \nabla f(\theta_n)\}\|^2 \right)$$

$$= \mathrm{Prox}_{\gamma, g} \left(\theta_n - \gamma \nabla f(\theta_n)\right)$$

The proximal-gradient algorithm (2/2)

- Since f is smooth: define a majorizing function $\theta \mapsto Q_\gamma(\theta; \theta_n)$

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

  where $L \leq 1/\gamma$.
- A family of majorizing functions - indexed by γ; all of them are equal to $F(\theta_n)$ at $\theta = \theta_n$
- Update the current solution

$$\theta_{n+1} = \mathrm{argmin}_\theta Q_\gamma(\theta; \theta_n) = \mathrm{argmin}_\theta \left(g(\theta) + \frac{1}{2\gamma} \|\theta - \{\theta_n - \gamma \nabla f(\theta_n)\}\|^2 \right)$$
$$= \mathrm{Prox}_{\gamma_{n+1}, g} \left(\theta_n - \gamma_{n+1} \nabla f(\theta_n)\right)$$

In practice

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

- A gradient step w.r.t. the smooth part of $f+g$
- A correction mecanism through the "prox" operator
  - ⋆ when g is the indicator of $\mathcal{K}$

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

  - ⋆ when g is the elastic net penalty

$$\theta_{n+1} = (\text{componentwise soft-thresholding of } \theta_n - \gamma_{n+1}\nabla f(\theta_n))$$

- In practice, it may happen that
  - ⋆ the gradient is not explicit but an approximation is available.
  - ⋆ the Prox operator is not explicit.

When the gradient can not computed

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

Run a Perturbed Proximal-Gradient Algorithm

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

When the gradient can not computed

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

Run a Perturbed Proximal-Gradient Algorithm

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

Questions:

- Conditions on η_{n+1}, γ_n for the convergence of the algorithm
- Rates of convergence

When the gradient can not computed

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

Run a Perturbed Proximal-Gradient Algorithm

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

Questions:

- Conditions on η_{n+1}, γ_n for the convergence of the algorithm
- Rates of convergence
- When H_{n+1} is a Monte Carlo sum
  - ⋆ how many points at each iteration (fixed/increasing batch size; constant/increasing number of draws)
  - ⋆ how to choose the stepsize γ_n : constant or decreasing ?

Outline

A deterministic result for the convergence of $\{\theta_n, n \geq 0\}$

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

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

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

Assume

1. g *convex, lower semi-continuous.*
2. f **convex**, *gradient is lipschitz.*
3. $\sum_n \gamma_n = +\infty$
4. *Convergence of the series*

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

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

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

- Result available for both deterministic and stochastic perturbations
- If stochastic perturbations: both biased and unbiased approximation of $\nabla f(\theta_n)$

Sketch of proof:

1. For any minimizer $\theta_\star$ of F

$$\|\theta_{n+1}-\theta_\star\|^2 \leq \|\theta_n-\theta_\star\|^2-\gamma_{n+1}\left(F(\theta_{n+1}) - \min F\right)+\gamma_{n+1}\text{noise}_{n+1} \quad (1)$$

2. Use a (deterministic) Siegmund-Robbins lemma:
   If
$$\sum_n \gamma_n = \infty, \qquad \sum_n \gamma_{n+1}\,\text{noise}_{n+1} < \infty$$
   then the limiting points of $\{\theta_n, n \geq 0\}$ are minimizers of F.

3. Use again (1) to show the convergence of $\{\theta_n\}_n$ to a minimizer of F.

Case of a Monte Carlo approximation

When

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

- replace $\nabla f(\theta_n)$ by a Monte Carlo approximation

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

  where $\{X_{n+1,j}, j \geq 0\}$ is a Markov chain with inv. dist. π_{θ_n}

- with an increasing number of samples m_{n+1} and a step-size γ_n s.t.

$$\sum_n \gamma_{n+1} = +\infty, \qquad \sum_n \frac{\gamma_{n+1}^2}{m_{n+1}} < \infty, \qquad \sum_n \frac{\gamma_{n+1}}{m_{n+1}} < \infty \text{ when biased approx.}$$

- or with a constant number of samples $m_{n+1} = m$ and a decreasing step-size γ_n s.t.

$$\sum_n \gamma_{n+1} = +\infty \qquad \sum_n \gamma_{n+1}^2 < \infty, \qquad + \text{ ergodicity cond. on the chains}$$

Outline

Examples of problems of the form: $\mathrm{argmin}_{\theta}\{f(\theta)+g(\theta)\}$

A first order method: the proximal gradient algorithm

Convergence of the Perturbed Proximal Gradient algorithm

Rates of convergence

Acceleration

References

A deterministic result

For non negative weights a_k

$$\sum_{k=1}^{n} a_k \{F(\theta_k) - \min F\} \leq U_n(\theta_\star)$$

Theorem (Atchadé, F., Moulines (2016))

For any $\theta_\star \in \mathcal{L}$,

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

Case of a Monte Carlo approximation

When

$$\nabla f(\theta_n) = \int H_{\theta_n}(x)\,\pi_{\theta_n}(\mathsf{d}x) \approx \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{n+1,j})$$

From the previous result, convergence rates in expectation, in L^q, $\cdots$: e.g.

- with $m_n = m$ and $\gamma_n = O(1/\sqrt{n})$

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

Case of a Monte Carlo approximation

When

$$\nabla f(\theta_n) = \int H_{\theta_n}(x)\, \pi_{\theta_n}(\mathsf{d}x) \approx \frac{1}{m_{n+1}} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{n+1,j})$$

From the previous result, convergence rates in expectation, in L^q, $\cdots$: e.g.

- with $m_n = m$ and $\gamma_n = O(1/\sqrt{n})$

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

- with $m_n \sim n$ and $\gamma_n = \gamma$

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

but $\cdots$ with $O(n^2)$ Monte Carlo samples.

Outline

Accelerated Proximal Gradient algorithm

Similarly to the Nesterov acceleration of the gradient algorithm (1983),

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

where t_n is a positive sequence s.t.

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

- The rate of convergence of the exact Proximal-Gradient algorithm:

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

- For the Accelerated Proximal-Gradient algorithm Beck and Teboulle, 2009, the rate is

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

Perturbed Accelerated Proximal Gradient Algorithm

- Sufficient conditions on the stepsizes γ_n, the coefficient t_n and on the perturbation
$$\eta_{n+1} = H_{n+1} - \nabla f(\tau_n)$$
for the convergence of $\{\theta_n, n \geq 0\}$

- Rate of convergence:
  - ⋆ deterministic case: $F(\theta_{n+1}) - \min F = O(t_n^{-2}\gamma_{n+1}^{-1})$
  - ⋆ Monte Carlo case, with $\gamma_n = \gamma$, $t_n = O(n)$, $m_n \sim n^3$:
  $$\mathbb{E}\left[F(\theta_n)\right] - \min F = O\left(\frac{1}{n^2}\right)$$
  but $\cdots$ after n^4 Monte Carlo samples
  - ⋆ (works in progress)

Outline

Convergence of Unbiased Stochastic Proximal-Gradient Algorithm

Combettes and Pesquet (2015, SIAM J. Optim) Stochastic Quasi-Fejer block-coordinate fixed point iterations with random sweeping.

Combettes and Pesquet (2015, arXiv) Stochastic Approximations and Perturbations in Forward-Backward Splitting for Monotone Operators.

Ghadimi and Lan (2015, Mathematical Programming) Accelerated Gradient methods for Nonconvex Nonlinear and Stochastic Programming.

Lin, Rosasco, Villa and Zhou (2015, arXiv) Modified Fejer Sequences and Applications

Nitanda (2014, NIPS) Stochastic Proximal Gradient descent with Acceleration Techniques.

Rosasco, Villa and Vu (2014, arXiv) Convergence of a Stochastic Proximal Gradient Algorithm.

Rosasco, Villa and Vu (2015, arXiv) A Stochastic Inertial Forward-Backward Splitting Algorithm for multivariate monotone inclusions.

Xiao and Zhang (2014, SIAM J. Optim) A Proximal Stochastic Gradient Method with Progressive Variance Reduction.

(Perturbed) Nesterov Acceleration of Proximal Gradient Algorithm

Attouch and Peypouquet (2015, arXiv) The rate of convergence of Nesterov's accelerated forward-backward method is actually $o(k^{-2})$.

Aujol and Dossal (2015, SIAM J. on Optim.) Stability of over-relaxations for the Forward-Backward algorithm, application to FISTA.

Chambolle and Dossal (2015, Journal of Optimization Theory and Applications) On the convergence of the iterates of the Fast Iterate Schrinkage Thresholding Algorithm.