Perturbed Proximal-Gradient Algorithms

Gersende Fort
LTCI, CNRS, Telecom ParisTech, Université Paris-Saclay, 75013 Paris France
Joint work with Y. Atchadé (Univ. Michigan, USA) and Eric Moulines (Ecole Polytechnique, France)

Problem

How to minimize / find the minimum

- on a convex subset Θ of some finite dimensional Euclidean space with norm $\|\cdot\|$
- of a convex function $f : \theta \mapsto f(\theta)$ from Θ to $\mathbb{R}$, which is smooth enough:

$$\exists L > 0 \text{ s.t. } \forall \theta, \theta' \in \Theta, \qquad \|\nabla f(\theta) - \nabla f(\theta')\| \leq L \|\theta - \theta'\|$$

- under non-smooth convex constraints $g : \theta \mapsto g(\theta)$ from Θ to $(-\infty, +\infty]$

when ∇f is not explicit ?

Problem 1: $\min_{\theta \in \Theta} \left(f(\theta) + g(\theta) \right)$ **Problem 1':** $\mathrm{argmin}_{\theta \in \Theta} \left(f(\theta) + g(\theta) \right)$

Example: Penalized Maximum Likelihood Inference in Latent Variable Models

The function f is *"– the log-likelihood of the observations* Y*"*: $f(\theta) = -\log \int p(\mathsf{Y}|\mathsf{x};\theta)\phi(\mathsf{x})\mu(\mathrm{d}\mathsf{x})$
The feasible set: $\theta \in \Theta \subseteq \mathbb{R}^d$
The penalty term is a sparsity constraint: $g(\theta) = \lambda \sum_{i=1}^d |\theta_i|$ which is not a differentiable function
Non explicit gradient of f**, but can be approximated:**

$$-\nabla f(\theta) = \int \nabla_\theta \Big(\log p(\mathsf{Y}|\mathsf{x};\theta) \Big) \; \pi_\theta(\mathsf{x}|\mathsf{Y}) \, \mu(\mathrm{d}\mathsf{x}) \approx \frac{1}{m} \sum_{k=1}^m \nabla_\theta \log p(\mathsf{Y}|\mathsf{X}_k;\theta)$$

where $(\mathsf{X}_k)_k$ is from a MCMC with target $\pi_\theta(\cdot|\mathsf{Y}) \, \mathrm{d}\mu$, the cond. dist. of the latent variables X given Y.

The Proximal-Gradient Algorithm

Iterative algorithm: see [1] for convergence results

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

Examples:

`Projection` on a closed convex set $\mathcal{K} \subseteq \Theta$

$$g(\theta) = \begin{cases} +\infty & \theta \notin \mathcal{K} \\ 0 & \theta \in \mathcal{K} \end{cases}$$

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

`Elastic net penalty`

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

$\theta_{n+1,i} = \text{shrinkage/thresholding of} \left(\theta_n - \gamma_{n+1} \nabla f(\theta_n) \right)_i$

$\hookrightarrow$ Unapplicable since $\nabla f(\theta_n)$ is not explicit in our framework
Questions: Can we replace $\nabla f(\theta_n)$ with an approximation while keeping the same asymptotic behavior ? How to choose the step-size γ_n ? In the Monte Carlo case, how to choose the (possibly) time-dependent batch-size m_n?

References

[1] A. Beck, and M. Teboulle. *Gradient-based algorithms with applications to signal-recovery problems.* Convex Optimization in Signal Processing and Communications, 2009.
[2] Y. Atchadé, G. Fort and E. Moulines. On Stochastic Proximal Gradient Algorithms *arXiv:1402:2365*, revised in Dec 2015.

The Perturbed Proximal Gradient Algorithm

Iterative algorithm:

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

where $\mathcal{K}$ is a convex closed subset of Θ and H_{n+1} is a (possibly deterministic) approximation of $\nabla f(\theta_n)$.
Monte Carlo case: when $\nabla f(\theta) = \mathbb{E}_\theta \left[H_\theta(X) \right]$ with $X \sim \pi_\theta$

$$H_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1,k}) \qquad \text{with Markov (or i.i.d) samples with inv. dist. } \pi_{\theta_n}$$

A General Convergence Result FROM [2, SECTION 3]

Set $\eta_{n+1} = H_{n+1} - \nabla f(\theta_n)$ $\qquad \mathcal{L} = \{\text{minimizers of } f + g\}$

Theorem. If $\gamma_n \in (0, 1/L]$, $\sum_n \gamma_n = +\infty$ and the following series converge

$$\sum_n \gamma_{n+1} \eta_{n+1}, \qquad \sum_n \gamma_{n+1} \left\langle T_{\gamma_{n+1},g}(\theta_n); \eta_{n+1} \right\rangle, \qquad \sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^2$$

then there exists $\theta_\infty \in \mathcal{L}$ such that $\lim_n \tilde{\theta}_n = \theta_\infty$.

Other results. Explicit expression for U_n s.t.

$$(f+g)\left(\bar{\theta}_n\right) - \min(f+g) \leq \sum_{k=1}^n \frac{a_k}{\sum_{t=1}^n a_t} (f+g)(\tilde{\theta}_k) - \min(f+g) \leq U_n \qquad \text{with } \bar{\theta}_n = \sum_{k=1}^n \frac{a_k}{\sum_{t=1}^n a_t} \tilde{\theta}_k$$

where $a_1, \cdots, a_n$ are non-negative real numbers

When applied to the Monte Carlo Proximal-Gradient Algorithm FROM [2, SECTION 4]

Under conditions on the Monte Carlo samples (geometric ergodicity, containment condition, $\cdots$):

$$\mathbb{E}\left[\|\eta_{n+1}\|^2 | \mathcal{F}_n \right] = O_{L^1}\left(\frac{1}{m_{n+1}} \right) \qquad \left\| \mathbb{E}\left[\eta_{n+1} | \mathcal{F}_n \right] \right\| = O_{L^1}\left(\frac{1}{m_{n+1}} \right)$$

with fixed batch-size $(m_n = m)$ but decreasing step-size γ_n s.t. $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < \infty$,

the above convergence result

U_n is $O(1/\sqrt{n})$ for different choices of (a_k, γ_k).

with increasing batch-size $(m_n \leq m_{n+1})$ at a linear rate $m_n \sim n$, and with constant step-size $\gamma_n = \gamma$

the above convergence result

U_n is $O(\ln /n)$ with a uniform weight $a_k = 1$; $\qquad$ rate after $O(n^2)$ MC samples.

Conclusions: We provided sufficient conditions for
(a) the same asymptotic behavior and the same rate of convergence as the exact algorithm,
(b) which hold for both the cases of a **biased and unbiased approximation** H_{n+1}