Convergence of a constant step stochastic proximal gradient algorithm with generalization to random monotone operators

A. Salim, P. Bianchi, W. Hachem

1 Introduction

The stochastic gradient algorithm aims to minimize a cost function that can be written as an expectation $x \mapsto \mathbb{E}(f(\xi, x))$, where ξ is a random variable and $f(\xi, .) : \mathbb{R}^N \to \mathbb{R}$ is a convex differentiable function. It can be used in the context where the expectation cannot be computed, but is revealed across time by the observation of i.i.d copies (ξ_n) of ξ. The stochastic gradient algorithm is written $x_{n+1} = x_n - \gamma_n \nabla f(\xi_{n+1}, x_n)$ where (γ_n) is a positive sequence of step-size. In the context of online machine learning, we often suppose that the step size is constant, *i.e* $\gamma_n \equiv \gamma$. In this case the process (x_n) generally doesn't almost surely converge as $n \to \infty$, but stay close with high probability to the set of minimizers (assumed to be not empty) in a double asymptotic regime : $n \to +\infty$ and $\gamma \to 0$ (see [4]).

The aim of this work is to analyze a generalization of the latter algorithm : the stochastic proximal gradient algorithm. In the deterministic case, this algorithm boils down to the standard proximal gradient algorithm, widely used in machine learning. Let $F : \mathbb{R}^N \to \mathbb{R}$ be a differentiable convex function and $G : \mathbb{R}^N \to (-\infty, +\infty]$ be a proper, convex, lower semi-continuous (lsc) function (notation : $G \in \Gamma_0$). Assume that $F + G$ has a minimizer, *i.e* that the set $Z(\nabla F + \partial G)$ of zeroes of $\nabla F + \partial G$ is not empty. The proximal gradient algorithm is written

$$x_{n+1} = \mathrm{prox}_{\gamma G}(x_n - \gamma \nabla F(x_n)) \tag{1}$$

where $\mathrm{prox}_{\gamma G}$ is the proximity operator of G and $\gamma > 0$ a step. If ∇F is Lipschitz continuous and if γ is enough small, then this algorithm converges to $Z(\nabla F + \partial G)$.

In the next section, we present the general problem and the constant step size stochastic proximal gradient algorithm that both generalizes the stochastic gradient algorithm and the proximal gradient algorithm.

2 The constant step proximal gradient algorithm

Consider two functions F and H that can be written as expectations $F(x) = \mathbb{E}(f(\xi, x))$, $H(x) = \mathbb{E}(h(\xi, x))$ where ξ is a random variable with distribution ρ over some measurable space Σ, $f(\xi, .) : \mathbb{R}^N \to \mathbb{R}$ is a convex differentiable function and $h(\xi, .) \in \Gamma_0$. Finally, consider $m \in \mathbb{N}^*$ and $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$ a family of closed convex sets of $\mathbb{R}^N$. Assume that the intersection $\bigcap_{i=1}^m \mathrm{ri}(\mathcal{C}_i)$ of relative interiors of $\mathcal{C}_i$ is not empty.

Our aim is to solve

$$\min_{x \in \mathcal{C}} F(x) + H(x), \quad \mathcal{C} := \bigcap_{i=1}^{m} \mathcal{C}_i \tag{2}$$

where we assume that the set of minimizers is not empty.

To solve this problem, we implement a stochastic version of the proximal gradient algorithm. Let (u_n) an i.i.d sequence with distribution ρ over Σ and (I_n) an i.i.d sequence with distribution α over the set $\{0, 1, \ldots, m\}$. We assume that $\alpha(k) = \mathbb{P}(I_1 = k) > 0$ for all k and (I_n) is independant of (u_n). In order to solve (2), the iterations write

$$x_{n+1} = \begin{cases} \operatorname{prox}_{\alpha(0)^{-1}\gamma h(u_{n+1},\cdot)}(x_n - \gamma \nabla f(u_{n+1}, x_n)) & \\ & \text{if } I_{n+1} = 0, \\ \Pi_{\mathcal{C}_{I_{n+1}}}(x_n - \gamma \nabla f(u_{n+1}, x_n)) & \text{else,} \end{cases} \tag{3}$$

where $\gamma > 0$ is a step size and $\Pi_{\mathcal{C}_i}$ is the projection onto $\mathcal{C}_i$.

This algorithm is of interest if the functions F (resp. H) is not available, or hard to compute, or the computation of its gradient (resp. its proximity operator) is computationally demanding. In these context, we replace the knowledge of the functions F and G by noisy versions $f(u_n, \cdot)$ and $h(u_n, \cdot)$. Moreover, if the number m is high, the projection onto $\mathcal{C}$ can be demanding whereas the projection onto the sets $\mathcal{C}_i$ is often easier. It is then useful to replace the projection onto $\mathcal{C}$ by projections onto $\mathcal{C}_i$. This algorithm generalizes the stochastic proximal gradient algorithm studied in [1] where the proximity step is deterministic. A lot of applications can be found in [1]. Applications of a slightly modified version of (3) can be found in [5] to solve learning problems over graphs.

The problem (2) is equivalent to the problem of finding an element in $Z(\nabla F + \partial G)$ where $G(x) := \sum_{k=1}^{m} \iota_{\mathcal{C}_k}(x) + H(x)$ and where ι_S is the indicator function of the set S in the sense of optimization theory. Since the algorithm (3) is a constant step size algorithm, it is not expected to converge to $Z(\nabla F + \partial G)$, but when $n \to +\infty$ and $\gamma \to 0$, we shall see that the iterates x_n stay close to $Z(\nabla F + \partial G)$.

Theorem 1. [2] Assume that $F(x) + G(x) \longrightarrow_{\|x\| \to +\infty} +\infty$ and that there exists $c > 0$ such that for all $x \in \mathbb{R}^N$,

$$\int \langle \nabla f(s, x) - \nabla f(s, x_\star), x - x_\star \rangle \, \rho(ds) \geq c \int \| f(s, x) - f(s, x_\star) \|^2 \, \rho(ds). \tag{4}$$

Then, under mild additional assumptions, for all r.v x_0 such that $\mathbb{E}[x_0^2] < \infty$,

$$\limsup_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}\left[d(x_k, Z(\nabla F + \partial G)) > \varepsilon \right] \xrightarrow[\gamma \to 0]{} 0 \, .$$

Using Baillon-Haddad theorem, the assumption (4) can be seen as a generalization of the standard assumption of uniform Lipschitz continuity of $\nabla f(\xi, \cdot)$.

3 Approach and general results

Our approach to prove this theorem is first to study the dynamical behavior of the iterates. Namely, we adapt the O.D.E method, well known in the literature of stochastic approximation ([4]). Consider x_γ the continuous time process obtained by linearly interpolating with time interval γ the iterates of the stochastic proximal gradient algorithm with step γ. We show that x_γ weakly converges to x as $\gamma \to 0$ over $\mathbb{R}_+$, where x is the unique solution to the Differential Inclusion (see [3])

$$\begin{cases} \dot{\mathsf{x}}(t) & \in & -(\nabla F + \partial G)(\mathsf{x}(t)) \\ \mathsf{x}(0) & = & x_0 \in \mathcal{D} \end{cases}$$

The latter Differential Inclusion induces a map $\Phi : \mathcal{D} \times \mathbb{R}_+ \to \mathcal{D}, (x_0, t) \mapsto \mathsf{x}(t)$ that can be extended to a semi-flow over $\overline{\mathcal{D}}$, still denoted by Φ.

The weak convergence is not enough to study the long term behavior of the iterates (x_n) : a stability result is needed. We then look at (x_n) as a Feller Markov chain with transition kernel Π_γ. The assumptions of the Theorem 1 ensures that the set I_γ of invariant measures of the Markov kernel Π_γ is not empty and that the set Inv $= \cup_{\gamma \in (0,\gamma_0]} I_\gamma$ is *tight* for all $\gamma_0 > 0$. Combined with the "dynamical behavior result" (the weak convergence of x_γ to x), this shows that all cluster point of Inv as $\gamma \to 0$ is an invariant measure for the semi-flow Φ. The conclusion of theorem 1, and other results, follow at once from this fact.

4 Generalization to random monotone operators

The stochastic proximal gradient algorithm aims to find an element in $Z(\nabla F + \partial G)$. The functions ∇F and ∂G are particular cases of functions called maximal monotone operators. A maximal monotone operator is a multivalued function $\mathbb{R}^N \to 2^{\mathbb{R}^N}$ that generalizes the subdifferential of a function in Γ_0. A general problem in optimization and more generally in applied mathematics is to find the set $Z(A + B) = \{x \in \mathbb{R}^N, \text{ such that } 0 \in (A + B)(x)\}$ of zeroes of the sum of two maximal monotone operators. This can be done iteratively by applying a generalized version of the proximal gradient algorithm called the forward backward algorithm. In the case where the monotone operators are revealed through realizations of i.i.d random monotone operators, a stochastic version of the forward backward algorithm that generalizes the stochastic proximal gradient algorithm can be implemented. Our study applies to the case of random maximal monotone operators, and we obtain similar results as in the previous section, by replacing $Z(\nabla F + \partial G)$ by $Z(A + B)$ (see [2]).

Références

[1] Yves F. Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10) :1–33, 2017.

[2] P. Bianchi, W. Hachem, and A. Salim. A constant step Forward-Backward algorithm involving random maximal monotone operators. *arXiv preprint arXiv :1702.04144*, 2017.

[3] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert.* North-Holland mathematics studies. Elsevier Science, Burlington, MA, 1973.

[4] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York).* Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[5] A. Salim, P. Bianchi, and W. Hachem. Snake : a stochastic proximal gradient algorithm for regularized problems over large graphs. in preparation, 2017.