Convergence of a constant step stochastic Douglas Rachford algorithm

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 writes $x_{n+1} = x_n - \gamma_n \nabla f(\xi_{n+1}, x_n)$ where (γ_n) is a positive sequence of step-size. In the context of online machine learning or adaptive signal processing, 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 [2]).

The aim of this work is to analyze a stochastic version of the well known Douglas-Rachford algorithm. Let $F : \mathbb{R}^N \to \mathbb{R}$ be a proper, convex, lower semi-continuous (lsc) function (notation : $F \in \Gamma_0(\mathbb{R}^N)$). We denote by ∂F the subdifferential of F. Let $G \in \Gamma_0(\mathbb{R}^N)$, assume that $F + G$ has a minimizer, *i.e* that the set $Z(\partial F + \partial G) = \{x \in \mathbb{R}^N \text{ such that } 0 \in \partial F(x) + \partial G(x)\}$ of zeroes of $\partial F + \partial G$ is not empty. The Douglas-Rachford algorithm is written

$$\begin{aligned} y_{n+1} &= \mathrm{prox}_{\gamma F}(x_n) \\ z_{n+1} &= \mathrm{prox}_{\gamma G}(2y_{n+1} - x_n) \\ x_{n+1} &= x_n + z_{n+1} - y_{n+1} \end{aligned} \tag{1}$$

where $\mathrm{prox}_{\gamma G}$ is the proximity operator of G and $\gamma > 0$ a step. If the standard qualification condition $0 \in \mathrm{ri}(\mathrm{dom}(F) - \mathrm{dom}(G))$ and assuming that $Z(\partial F + \partial G)$ is non empty, the sequence (y_n) converge to an element of $Z(\partial F + \partial G)$ as $n \to \infty$.

2 The constant step Douglas Rachford algorithm

Consider a probability space $(\Xi, \mathcal{F}, \rho)$. We say that a mapping $f : \mathbb{R}^N \times \Xi \to (-\infty, +\infty]$ is a normal convex integrand if $f(\,.\,, s) \in \Gamma_0(\mathbb{R}^N)$ for every $s \in \Xi$ and if $f(x, \,.\,)$ is measurable for every $x \in \mathbb{R}^N$.

From now on, assume that the mapping F and G are of the form $F(x) = \mathbb{E}(f(x, \xi))$ and $G(x) = \mathbb{E}(g(x, \xi))$ where f, g are normal convex integrands. The aim of the adaptive Douglas Rachford algorithm is to solve

$$\min_{x \in \mathbb{R}^N} F(x) + G(x). \tag{2}$$

Denote by $(\xi_n : n \in \mathbb{N})$ a sequence of iid copies of the r.v. ξ. In the sequel, we use the notation $f_n := f(\,.\,,\xi_n)$ and $g_n := g(\,.\,,\xi_n)$. The adaptive Douglas-Rachford algorithm is given by

$$\begin{aligned} y_{n+1} &= \mathrm{prox}_{\gamma, f_{n+1}}(x_n) \\ z_{n+1} &= \mathrm{prox}_{\gamma, g_{n+1}}(2y_{n+1} - x_n) \\ x_{n+1} &= x_n + z_{n+1} - y_{n+1}\,. \end{aligned} \tag{3}$$

This algorithm is of interest if the functions F, G are not available, or hard to compute, or the computation of their proximity operator is computationally demanding. In these context, we replace the knowledge of the functions F and G by noisy versions f_n and g_n. An applications of of this algorithm can be find in [**?**] to solve a tracking problem.

The problem (2) is equivalent to the problem of finding an element in $Z(\partial F + \partial G)$. Since the algorithm (3) is a constant step size algorithm, it is not expected to converge to $Z(\partial F + \partial G)$, but we will show in the sequel that when $n \to +\infty$ and $\gamma \to 0$ the iterates x_n stay close to $Z(\partial F + \partial G)$.

Theorem 1. [**?**] Assume that $F(x) + G(x) \longrightarrow_{\|x\| \to +\infty} +\infty$, that ρ-a.s $f(\cdot, s)$ is differentiable and that there exists $L > 0$ such that ρ-a.s $\nabla f(\cdot, s)$ is L-Lipschitz continuous.

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

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

The assumptions of the theorem deserve some comments.

3 Proof of the convergence

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 ([2]). 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 [1])

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

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 $\mathrm{Inv} = \cup_{\gamma \in (0, \gamma_0]} I_\gamma$ 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.

Références

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

[2] 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.