A CONSTANT STEP STOCHASTIC DOUGLAS-RACHFORD ALGORITHM WITH APPLICATION TO NON SEPARABLE REGULARIZATIONS

Adil Salim*, Pascal Bianchi* and Walid Hachem†

$\star$ LTCI, Télécom ParisTech, Université Paris-Saclay.
46, rue Barrault, 75634 Paris Cedex 13, France.
$\dagger$ CNRS / LIGM (UMR 8049), Université Paris-Est Marne-la-Vallée.
5, boulevard Descartes, Champs-sur-Marne, 77454, Marne-la-Vallée Cedex 2, France.

ABSTRACT

The Douglas Rachford algorithm is an algorithm that converges to a minimizer of a sum of two convex functions. The algorithm consists in fixed point iterations involving computations of the proximity operators of the two functions separately. The paper investigates a stochastic version of the algorithm where both functions are random and the step size is constant. We establish that the iterates of the algorithm stay close to the set of solution with high probability when the step size is small enough. Application to structured regularization is considered.

Index Terms— Stochastic optimization, proximal methods, Douglas Rachford algorithm, structured regularizations

1. INTRODUCTION

Many applications in the fields of machine learning [1] and signal processing [2] require the solution of the programming problem

$$\min_{x \in \mathsf{X}} F(x) + G(x) \qquad (1)$$

where X is an Euclidean space, F and G are elements of the set $\Gamma_0(\mathsf{X})$ of convex, lower semi-continuous and proper functions. In these contexts, F often represents a cost function and G a regularization term. The Douglas-Rachford algorithm is one of the most popular approach towards solving Problem (1). Given $\gamma > 0$, the algorithm is written

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

where $\mathrm{prox}_{\gamma F}$ denotes the proximity operator of F, defined for every $x \in \mathsf{X}$ by the equation

$$\mathrm{prox}_{\gamma F}(x) = \arg\min_{y \in \mathsf{X}} \frac{1}{2}\|x - y\|^2 + \gamma F(y).$$

Assuming that a standard qualification condition holds and that the set of solutions $\arg\min F + G$ of (1) is not empty, the sequence $(y_n)_n$ converges to an element in $\arg\min F + G$ as $n \to +\infty$ ([3, 4]).

In this paper, we study the case where F and G are integral functionals of the form

$$F(x) = \mathbb{E}_\xi(f(x, \xi)), \qquad G(x) = \mathbb{E}_\xi(g(x, \xi))$$

where ξ is a random variable (r.v) from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into a measurable space $(\Xi, \mathcal{G})$, with distribution μ, and where $\{f(\cdot, s), s \in \Xi\}$ and $\{g(\cdot, s), s \in \Xi\}$ are subsets of $\Gamma_0(\mathsf{X})$. In this context, the stochastic Douglas Rachford algorithm aims to solve Problem (1) by iterating

$$\begin{aligned} y_{n+1} &= \mathrm{prox}_{\gamma f(\cdot, \xi_{n+1})}(x_n) \\ z_{n+1} &= \mathrm{prox}_{\gamma g(\cdot, \xi_{n+1})}(2y_{n+1} - x_n) \\ x_{n+1} &= x_n + z_{n+1} - y_{n+1}, \end{aligned} \qquad (3)$$

where $(\xi_n)_n$ is a sequence of i.i.d copies of the random variable ξ and $\gamma > 0$ is the constant step size. Compared to the "deterministic" Douglas Rachford algorithm (2), the stochastic Douglas Rachford algorithm (3) is an online method. The constant step size used make it implementable in adaptive signal processing or online machine learning contexts. In this algorithm, the function F (resp. G) is replaced at each iteration n by a random realization $f(\cdot, \xi_n)$ (resp. $g(\cdot, \xi_n)$). It can be implemented in the case where F (resp. G) cannot be computed in its closed form [5, 6] or in the case where the computation of its proximity operator is demanding [7]. Compared to other online optimization algorithm like the stochastic subgradient algorithm, the algorithm (3) benefits from the numerical stability of stochastic proximal methods.

Stochastic version of the Douglas Rachford algorithm have been considered in [2, 8]. These papers consider the case where G is deterministic, *i.e* is not written as an expectation and F is written as an expectation that reduces to a sum. The latter case is also contained as a particular case of the algorithm [9]. The algorithms [10, 11] are generalizations of a partially stochastic Douglas Rachford algorithm where

G is deterministic. The convergence of these algorithms is obtained under a summability assumption of the noise over the iterations. The stochastic Douglas Rachford studied in this paper was implemented in an adaptive signal processing context [12] to solve a target tracking problem.

Whereas the paper [12] is mainly focused on the application to target tracking, in this work we provide theoretical basis for the algorithm (3) and convergence results. Moreover, a novel application to solve a programming problem regularized with the overlapping group lasso online is provided.

The next section introduces some notations. Section 3 is devoted to the statement of the main convergence result. In Section 4, an outline of the proof of the result in Section 3 is provided. Finally, the algorithm (3) is implemented to solve a regularized problem in Section 5.

2. NOTATIONS

For every function $g \in \Gamma_0(\mathsf{X})$, $\partial g(x)$ denotes the subdifferential of g at the point $x \in \mathsf{X}$ and $\partial g_0(x)$ the least norm element in $\partial g(x)$. The domain of g is denoted as $\mathrm{dom}(g)$. It is a known fact that the closure of $\mathrm{dom}(g)$, denoted as $\mathrm{cl}(\mathrm{dom}(g))$, is convex. For every closed convex set $\mathcal{C}$, we denote by Π_C the projection operator onto $\mathcal{C}$. The indicator function of the set $\mathcal{C}$ is defined by $\iota_{\mathcal{C}}(x) = 0$ if $x \in \mathcal{C}$, and $\iota_{\mathcal{C}}(x) = +\infty$ elsewhere. It is easy to see that $\iota_{\mathcal{C}} \in \Gamma_0(\mathsf{X})$ and that $\mathrm{prox}_{\iota_{\mathcal{C}}} = \Pi_{\mathcal{C}}$.

The Moreau envelope of $g \in \Gamma_0(\mathsf{X})$ is equal to

$$g_\gamma(x) = \min_{y \in \mathsf{X}} g(y) + \frac{\|y - x\|^2}{2\gamma}$$

for every $x \in \mathsf{X}$. Recall that g_γ is differentiable and $\nabla g_\gamma(x) = \frac{1}{\gamma}(x - \mathrm{prox}_{\gamma g}(x))$. If $f \in \Gamma_0(\mathsf{X})$ is differentiable, then, $\partial f(x) = \{\nabla f(x)\}$ and $\nabla f(\mathrm{prox}_{\gamma f}(x)) = \nabla f_\gamma(x)$, for every $x \in \mathsf{X}$.

When $S \subset \mathsf{X}$, $d(x, S)$ denote the distance from the point $x \in \mathsf{X}$ to the set S. In the context of algorithm (3) we shall denote $D(s) = \mathrm{dom}(g(\cdot, s))$ and $\mathcal{D} = \mathrm{dom}(G)$. Denote $\mathscr{B}(\mathsf{X})$ the Borel sigma field over X. For every $p \geq 1$, $L^p(\Xi, \mathsf{X})$ is the set of all r.v φ from the probability space $(\Xi, \mathcal{G}, \mu)$ into the measurable space $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$, such that $\|\varphi\|^p$ is integrable.

From now on, we shall state explicitly the dependence of the iterates of the algorithm in the step size and the starting point. Namely, we shall denote $(x_n^{\gamma,\nu})_n$ the sequence $(x_n)_n$ generated by the stochastic Douglas Rachford algorithm (3) with step γ, such that the distribution of $x_0^{\gamma,\nu}$ over X is ν. If $\nu = \delta_a$, where δ_a is the Dirac measure at the point $a \in \mathsf{X}$, we shall prefer the notation $x_n^{\gamma,a}$.

3. MAIN CONVERGENCE THEOREM

Consider the following assumptions.

Assumption 1. For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x \in \mathcal{K} \cap \mathcal{D}} \int \|\partial g_0(x, s)\|^{1+\varepsilon} \mu(ds) < \infty.$$

Assumption 2. For μ-a.e $s \in \Xi$, $f(\cdot, s)$ is differentiable and there exists a closed ball in X such that $\|\nabla f(x, s)\| \leq M(s)$ for all x in this ball, where $M(s)$ is μ-integrable. Moreover, for every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon > 0$ such that

$$\sup_{x \in \mathcal{K}} \int \|\nabla f(x, s)\|^{1+\varepsilon} \mu(ds) < \infty.$$

Assumption 3. $\forall x \in \mathsf{X}$, $\int d(x, D(s))^2 \mu(ds) \geq C \boldsymbol{d}(x)^2$.

Assumption 4. For every compact set $\mathcal{K} \subset \mathsf{X}$, there exists $\varepsilon, C, \gamma_0 > 0$ such that for all $\gamma \in (0, \gamma_0]$ and all $x \in \mathcal{K}$,

$$\frac{1}{\gamma^{1+\varepsilon}} \int \| \mathrm{prox}_{\gamma g(\cdot, s)}(x) - \Pi_{\mathrm{cl}(D(s))}(x)\|^{1+\varepsilon} \mu(ds) < C.$$

Assumption 5. There exists $L > 0$ such that $\nabla f(\cdot, s)$ is μ-a.e, a L-Lipschitz continuous function.

Assumption 6. There exists $x_\star \in \arg\min F + G$ and $\varphi \in L^2(\Xi, \mathsf{X})$ such that $\varphi(s) \in \partial g(x_\star, s)$ μ-a.s, $\nabla f(x_\star, \cdot) \in L^2(\Xi, \mathsf{X})$ and $\int \nabla f(x_\star, s) \mu(ds) + \int \varphi(s) \mu(ds) = 0$.

Assumption 7. The function $F + G$ satisfies one of the following properties :

(a) $F + G$ is coercive *i.e* $F(x) + G(x) \longrightarrow_{\|x\| \to +\infty} +\infty$

(b) $F + G$ is supercoercive *i.e* $\frac{F(x) + G(x)}{\|x\|} \longrightarrow_{\|x\| \to +\infty} +\infty$.

Assumption 8. There exists $\gamma_0 > 0$, such that for all $\gamma \in (0, \gamma_0]$ and all $x \in \mathsf{X}$,

$$\int \|\nabla f_\gamma(x, s)\| + \frac{1}{\gamma}\| \mathrm{prox}_{\gamma g(\cdot, s)}(x) - \Pi_{\mathrm{cl}(D(s))}(x)\|\mu(ds)$$
$$\leq C(1 + |F^\gamma(x) + G^\gamma(x)|).$$

Theorem 1. Let Assumptions 1– 8 hold true. Then, for each probability measure ν over X having a finite second moment, for any $\varepsilon > 0$,

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

Moreover, if Assumption 7–(b) holds true, then

$$\limsup_{n \to \infty} \mathbb{P}\left(d\left(\bar{x}_n^{\gamma,\nu}, \arg\min(F + G)\right) \geq \varepsilon\right) \xrightarrow{\gamma \to 0} 0, \text{ and}$$

$$\limsup_{n \to \infty} d\left(\mathbb{E}(\bar{x}_n^{\gamma,\nu}), \arg\min(F + G)\right) \xrightarrow{\gamma \to 0} 0.$$

where $\bar{x}_n^{\gamma,\nu} = \frac{1}{n} \sum_{k=1}^{n} x_k^{\gamma,\nu}$.

Loosely speaking, the theorem states that, with high probability, the iterates $(x_n^{\gamma,\nu})_n$ stay close to the set of solutions $\arg\min F + G$ as $n \to \infty$ and $\gamma \to 0$.

Some Assumptions deserve comments.

Following [13], we say that a finite collection of subsets $\mathcal{C}_1, \ldots, \mathcal{C}_m$ of X is *linearly regular* if

$$\exists \kappa > 0, \forall x \in \mathsf{X}, \max_{s \in \{1,\ldots,m\}} d(x, \mathcal{C}_s) \geq \kappa d(x, \cap_{s=1}^m \mathcal{C}_s)$$

In the case where there exists a μ-probability one set $\tilde{\Xi}$ such that the set $\{D(s), s \in \tilde{\Xi}\} = \{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$ is finite, it is routine to check that Assumption 3 holds if and only if the domains $\mathcal{C}_1, \ldots, \mathcal{C}_m$ are linearly regular. See [12] for an applicative context of the algorithm (3) in the latter case.

It is a known fact that

$$\text{prox}_{\gamma g(\cdot,s)}(x) \longrightarrow_{\gamma \to 0} \Pi_{\text{cl}(\text{dom}(g(\cdot,s)))}(x),$$

for each (x,s). Assumptions 4 and 8 add controls on the convergence rate.

Since $f(\cdot,s), g(\cdot,s) \in \Gamma_0(\mathsf{X})$, and $f(\cdot,s)$ is differentiable, $\partial(F + G)(x) = \nabla F(x) + \partial G(x) = \mathbb{E}(\nabla f(x,\xi)) + \mathbb{E}(\partial g(x,\xi))$ [14], where the set $\mathbb{E}(\partial g(x,\xi))$ is defined by its Aumann integral

$$\left\{ \int \varphi(s)\,\mu(ds), \varphi \in L^1(\Xi, \mathsf{X}), \text{ s.t. } \varphi(s) \in \partial g(x,s), \mu\text{-a.s.} \right\}$$

Therefore, using Fermat's rule, if $x \in \arg\min F + G$, then there exists $\varphi \in L^1(\Xi, \mathsf{X})$, such that $\varphi(s) \in \partial g(x,s)$ μ-a.s, and $\int \nabla f(x,s)\,\mu(ds) + \int \varphi(s)\,\mu(ds) = 0$. We refer to $(\nabla f(x,\cdot), \varphi)$ as a *representation* of the solution x. Assumption 6 ensures the existence of $x_\star \in \arg\min F + G$ with a representation $\nabla f(x,\cdot), \varphi \in L^2(\Xi, \mathsf{X})$.

4. OUTLINE OF THE CONVERGENCE PROOF

This section is devoted to sketching the proof of the convergence of the stochastic Douglas Rachford algorithm. The approach follows the same steps as [6] and is detailed in [15]. The first step of the proof is to study the dynamical behavior of the iterates $(x_n^{\gamma,a})_n$ where $a \in \mathcal{D}$. The Ordinary Differential Equation (ODE) method, well known in the literature of stochastic approximation ([16]), is applied. Consider the continuous time stochastic process $\mathsf{x}_{\gamma,a}$ obtained by linearly interpolating with time interval γ the iterates $(x_n^{\gamma,a})$:

$$\mathsf{x}_{\gamma,a}(t) = x_n^{\gamma,a} + (t - n\gamma)\frac{x_{n+1}^{\gamma,a} - x_n^{\gamma,a}}{\gamma}, \qquad (4)$$

for all $t \geq 0$ such that $n\gamma \leq t < (n+1)\gamma$, for all $n \in \mathbb{N}$. Let Assumptions 1–4 [1] hold true. Consider the set $C(\mathbb{R}_+, \mathsf{X})$

of continuous functions from $\mathbb{R}_+$ to X equipped with the topology of uniform convergence on the compact intervals. It is shown that the continuous time stochastic process $\mathsf{x}_{\gamma,a}$ converges weakly over $\mathbb{R}_+$ (*i.e* in distribution in $C(\mathbb{R}_+, \mathsf{X})$) as $\gamma \to 0$. Moreover, the limit is proven to be the unique absolutely continuous function x over $\mathbb{R}_+$ satisfying $\mathsf{x}(0) = a$ and for almost every $t \geq 0$, the Differential Inclusion (DI),

$$\dot{\mathsf{x}}(t) \in -(\nabla F + \partial G)(\mathsf{x}(t)), \qquad (5)$$

(see [17]). Differential inclusions like (5) generalize ODE to set-valued mappings. The DI (5) 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 $\text{cl}(\mathcal{D})$, still denoted by Φ.

The weak convergence of $(\mathsf{x}_{\gamma,a})$ to x is not enough to study the long term behavior of the iterates $(x_n^{\gamma,a})_n$. The second step of the proof is to prove a stability result for the Feller Markov chain $(x_n^{\gamma,a})_n$. Denote by P_γ its transition kernel. The deterministic counterpart of this step of the proof is the so-called *Fejér monotonicity* of the sequence (x_n) of the algorithm (2). Even if some work has been done [5, 18], there is no immediate way to adapt the Fejér monotonicity to our random setting, mainly because of the constant step γ. As an alternative, we assume Hypotheses 5-6, and prove the existence of positive numbers α, C and γ_0, such that for every $\gamma \in (0, \gamma_0]$,

$$\mathbb{E}_n \|x_{n+1}^{\gamma,a} - x_\star\|^2 \leq \|x_n^{\gamma,a} - x_\star\|^2 \qquad (6)$$
$$- \alpha\gamma(F^\gamma + G^\gamma)(x_n^{\gamma,a}) + \gamma C.$$

In this inequality, $\mathbb{E}_n$ denotes the conditional expectation with respect to the sigma-algebra $\sigma(x_0^\gamma, x_1^\gamma, \ldots, x_n^\gamma)$ and

$$F^\gamma(x) = \int f_\gamma(x,s)\,\mu(ds), \quad G^\gamma(x) = \int g_\gamma(x,s)\,\mu(ds).$$

Since $\gamma \mapsto F^\gamma(x) + G^\gamma(x)$ is decreasing [6, 15], the function $F^\gamma + G^\gamma$ can be replaced by $F^{\gamma_0} + G^{\gamma_0}$. Besides, the coercivity of $F + G$ (Assumption 7) implies the coercivity of $F^{\gamma_0} + G^{\gamma_0}$ ([6, 15]). Therefore, assuming 5–7 and setting $\Psi = F^{\gamma_0} + G^{\gamma_0}$, there exist positive numbers α, C and γ_0, such that for every $\gamma \in (0, \gamma_0]$,

$$\mathbb{E}_n \|x_{n+1}^{\gamma,a} - x_\star\|^2 \leq \|x_n^{\gamma,a} - x_\star\|^2 - \alpha\gamma\Psi(x_n^{\gamma,a}) + \gamma C. \quad (7)$$

Equation (7) can alternatively be seen as a tightness result. It implies that the set I_γ of invariant measures of the Markov kernel P_γ is not empty for every $\gamma \in (0, \gamma_0]$, and that the set

$$\text{Inv} = \cup_{\gamma \in (0, \gamma_0]} I_\gamma \qquad (8)$$

is *tight*([19, 20]).

It remains to characterize the cluster points of Inv as $\gamma \to 0$. To that end, the dynamical behavior result and the stability result are combined. Let Assumptions 1– 8 hold true. [2] Then,

---

1. In the case where the domains are common, *i.e* $s \mapsto D(s)$ is μ-a.s constant, the moment Assumptions 1 and 2 are sufficient to state the dynamical behavior result. See [12] for an applicative context where the domains $D(s)$ are distinct.

2. Assumptions 3, 4 and 8 are not needed if the domains $D(s)$ are common.

the set Inv is tight, and, as $\gamma \to 0$, every cluster point of Inv is an invariant measure for the semi-flow Φ. The Theorem 1 is a consequence of this fact.

5. APPLICATION TO STRUCTURED REGULARIZATION

In this section is provided an application of the stochastic Douglas Rachford (3) algorithm to solve a regularized optimization problem. Consider problem (1), where F is a cost function that is written as an expectation, and G is a regularization term. Towards solving (1), many approaches involve the computation of the proximity operator of the regularization term G. In the case where G is a structured regularization term, its proximity operator is often difficult to compute. When G is a graph-based regularization, it is possible to apply a stochastic proximal method to address the regularization [7]. We shall concentrate on the case where G is an overlapping group regularization. In this case, the computation of the proximity operator of G is known to be a bottleneck [21]. We shall apply the algorithm (3) to overcome this difficulty.

Consider $\mathsf{X} = \mathbb{R}^N$, $N \in \mathbb{N}^\star$, and $g \in \mathbb{N}^\star$. Consider g subsets of $\{1, \dots, N\}$, $S_1, \dots, S_g$, possibly overlapping. Set $G(x) = \sum_{j=1}^{g} \|x_{S_j}\|$, where x_{S_j} denotes the restriction of x to the set of index S_j and $\| \cdot \|$ denotes the Euclidean norm. Set $F(x) = \mathbb{E}_{(\xi, \eta)}(h(\eta \langle x, \xi \rangle))$ where h denotes the hinge loss $h(z) = \max(0, 1 - z)$ and (ξ, η) is a r.v defined on some probability space with values in $\mathsf{X} \times \{-1, +1\}$. In this case, the problem (1) is also called the SVM classification problem, regularized by the overlapping group lasso. It is assumed that the user is provided with i.i.d copies $((\xi_n, \eta_n))_n$ of the r.v (ξ, η) online.

To solve this problem, we implement a stochastic Douglas Rachford strategy. To that end, the regularization G is rewritten $G(x) = \mathbb{E}_J(g \| x_{S_J} \|)$ where J is an uniform r.v over $\{1, \dots, g\}$. At each iteration n of the stochastic Douglas Rachford algorithm, the user is provided with the realization (ξ_n, η_n) and sample a group J_n uniformly in $\{1, \dots, g\}$. Then, a Douglas Rachford step is done, involving the computation of the proximity operators of the functions $g_n : x \mapsto \|x_{S_{J_n}}\|$ and $f_n : x \mapsto h(\eta_n \langle x, \xi_n \rangle)$.

This strategy is compared with a partially stochastic Douglas Rachford algorithm, deterministic in the regularization G, where the fast subroutine Fog-Lasso [21] is used to compute the proximity operator of the regularization G. At each iteration n, the user is provided with (ξ_n, η_n). Then, a Douglas Rachford step is done, involving the computation of the proximity operators of the functions G and $f_n : x \mapsto h(\eta_n \langle x, \xi_n \rangle)$. Figure 1 demonstrates the advantage of treating the regularization term in a stochastic way.

In Figure 1 "Stochastic D-R" denotes the stochastic Douglas Rachford algorithm and "Partially stochastic D-R" denotes the partially stochastic Douglas Rachford where the subroutine FoG-Lasso [21] is used at each iteration to compute
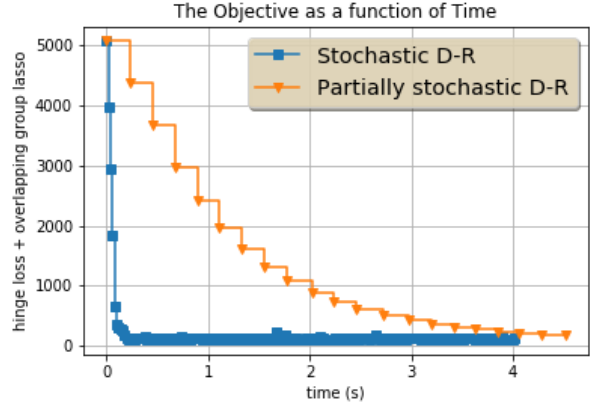

Fig. 1. The objective function $F + G$ as a function of time in seconds for each algorithm
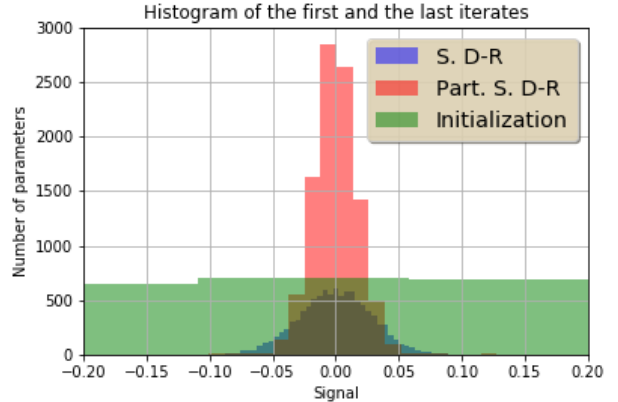

Fig. 2. Histogram of the Initialization and the last iterates of the Stochastic D-R (S. D-R) and the partially stochastic D-R (Part. S. D-R)

the true proximity operator of the regularization G. Figure 2 shows the appearance of the first and the last iterates. Even if a best performing procedure [21] is used to compute $\text{prox}_{\gamma G}$, we observe on Figure 1 that Stochastic D-R takes advantage of being a stochastic method. This advantage is known to be twofold ([22]). First, the iteration complexity of Stochastic D-R is moderate because $\text{prox}_{\gamma G}$ is never computed. Then, Stochastic D-R is faster than its partially deterministic counterpart which uses Fog-Lasso [21] as a subroutine, especially in the first iterations of the algorithms. Moreover, Stochastic D-R seems to perform globally better. This is because every proximity operators in Stochastic D-R can be efficiently computed ([23]). Contrary to the proximity operator of G [21], the proximity operator of g_n is easily computable. The proximity operator of f_n is easily computable as well. [3]

---

3. Even if $h(x) = \log(1 + \exp(-x))$ (logistic regression), the proximity operator of f_n is easily computable, see [2].

6. REFERENCES

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[2] G. Chierchia, A. Cherni, E. Chouzenoux, and J.-C. Pesquet, "Approche de douglas-rachford aléatoire par blocs appliquée à la régression logistique parcimonieuse," in *GRETSI*, 2017.

[3] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Numer. Anal.*, vol. 16, no. 6, pp. 964–979, 1979.

[4] J. Eckstein and D. P. Bertsekas, "On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, 1992.

[5] P. Bianchi and W. Hachem, "Dynamical behavior of a stochastic Forward-Backward algorithm using random monotone operators," *ArXiv e-prints, 1508.02845*, Aug. 2015, To be published in J. Optim. Theory Appl.

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

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

[8] Z. Shi and R. Liu, "Online and stochastic douglas-rachford splitting method for large scale machine learning," *arXiv preprint arXiv :1308.4757*, 2013.

[9] A. Chambolle, M. J Ehrhardt, P. Richtárik, and C.-B. Schönlieb, "Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging application," *arXiv preprint arXiv :1706.04957*, 2017.

[10] L. Rosasco, S. Villa, and B.-C. Vu, "Stochastic inertial primal-dual algorithms," *arXiv preprint arXiv :1507.00852*, 2015.

[11] P.-.L Combettes and J.-C. Pesquet, "Stochastic forward-backward and primal-dual approximation algorithms with application to online image restoration," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1813–1817.

[12] R. Mourya, P. Bianchi, A. Salim, and C. Richard, "An adaptive distributed asynchronous algorithm with application to target localization," in *CAMSAP*, 2017.

[13] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM review*, vol. 38, no. 3, pp. 367–426, 1996.

[14] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*, Springer-Verlag, Berlin, 1998.

[15] A. Salim, P. Bianchi, and W. Hachem, "A stochastic douglas-rachford algorithm with constant step size," Tech. Rep., see https ://adil-salim.github.io/Research, 2017.

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

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

[18] P.-L. Combettes and J.-C. Pesquet, "Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 1221–1248, 2015.

[19] J.-C. Fort and G. Pagès, "Asymptotic behavior of a Markovian stochastic algorithm with constant step," *SIAM J. Control Optim.*, vol. 37, no. 5, pp. 1456–1482 (electronic), 1999.

[20] P. Bianchi, W. Hachem, and A. Salim, "Asymptotics of constant step stochastic approximations involving differential inclusions," *arXiv preprint arXiv :1612.03831*, 2016.

[21] L. Yuan, J. Liu, and J. Ye, "Efficient methods for overlapping group lasso," in *Advances in NIPS*, 2011, pp. 352–360.

[22] L. Bottou, F. E Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *arXiv preprint arXiv :1606.04838*, 2016.

[23] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.