A Stochastic Douglas Rachford Algorithm with Application to Non Separable Regularizations

Adil Salim, Pascal Bianchi and Walid Hachem

Telecom ParisTech and Universite Paris-Est-Marne-la-Vallee

Objectives

- F, R convex functions over $\mathsf{X} = \mathbb{R}^d$
- **Problem**:

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

- ξ and ζ are random variables
- **Data fitting** term $F(x) = \mathbb{E}_\xi(f(x, \xi))$
- **Regularization** term $R(x) = \mathbb{E}_\zeta(r(x, \zeta))$

Example: Overlapping Group Regularizations

Structured sparsity:

- $$F(x) = \sum_{i=1}^{N} f_i(x) \qquad (2)$$

  cost function associated with SVM or logistic regression
- $\mathcal{G}$ is a set of **possibly overlapping** subsets of $\{1, \ldots, d\}$
- $$R(x) = \sum_{g \in \mathcal{G}} r_g(x) \qquad (3)$$
- $\forall g \in \mathcal{G}$, $x_{|g}$ is the restriction of vector x to g ($e.g$ if $g = 1, 2, 4$ then $x = (x_1, x_2, x_4)$)
- $r_g(x) = \|x_{|g}\|_1$

Douglas Rachford Algorithm

Proximal methods for solving (1) are known for numerical stability. Proximity operator

$$\operatorname{prox}_{\gamma R}(x) = \arg\min_{y \in \mathsf{X}} \frac{1}{2\gamma}\|x - y\|^2 + R(y), \quad \gamma > 0$$

Standard method: Douglas-Rachford

$$\begin{aligned} y_{n+1} &= \operatorname{prox}_{\gamma F}(x_n) \\ z_{n+1} &= \operatorname{prox}_{\gamma R}(2y_{n+1} - x_n) \\ x_{n+1} &= x_n + z_{n+1} - y_{n+1} \end{aligned} \qquad (4)$$

Theorem ([1]) : $y_n \longrightarrow_{n \to +\infty} \arg\min F + R$

- Related to ADMM
- Converges with a constant step $\gamma > 0$
- Splitting method

More splitting?

Sometimes we need more splitting for both F and R.

- In adaptive signal processing/online learning, F is unknown but revealed through i.i.d realizations of ξ
- Even if F is known, $\operatorname{prox}_{\gamma F}$ is often intractable ($e.g$ (2))
- In many cases ($e.g$ (3)), $\operatorname{prox}_{\gamma R}$ is also intractable

Contact Information

- adil.salim,pascal.bianchi@telecom-paristech.fr
- walid.hachem@u-pem.fr

Stochastic Douglas Rachford Algorithm

Takes advantage of the numerical stability without the iteration complexity.

$$\begin{aligned} y_{n+1} &= \operatorname{prox}_{\gamma f(\cdot, \xi_{n+1})}(x_n^\gamma) \\ z_{n+1} &= \operatorname{prox}_{\gamma r(\cdot, \zeta_{n+1})}(2y_{n+1} - x_n^\gamma) \\ x_{n+1}^\gamma &= x_n^\gamma + z_{n+1} - y_{n+1} \end{aligned} \qquad (5)$$

where

- (ξ_n) (resp. (ζ_n)) are i.i.d copies of ξ (resp. ζ)
- Constant step $\gamma > 0$

The random functions $f(\cdot, \xi_{n+1})$ (resp. $r(\cdot, \xi_{n+1})$) can be much simpler than F (resp. R), see (2)-(3) ([2]).

Dynamical Behavior

Constant step $\gamma > 0$: No a.s convergence. Stochastic approximation technique [3]:

$$\mathsf{x}_\gamma(t) = x_n^\gamma + (t - n\gamma)\frac{x_{n+1}^\gamma - x_n^\gamma}{\gamma}, \qquad (6)$$

where $n > 0, \quad n\gamma \le t < (n+1)\gamma$.

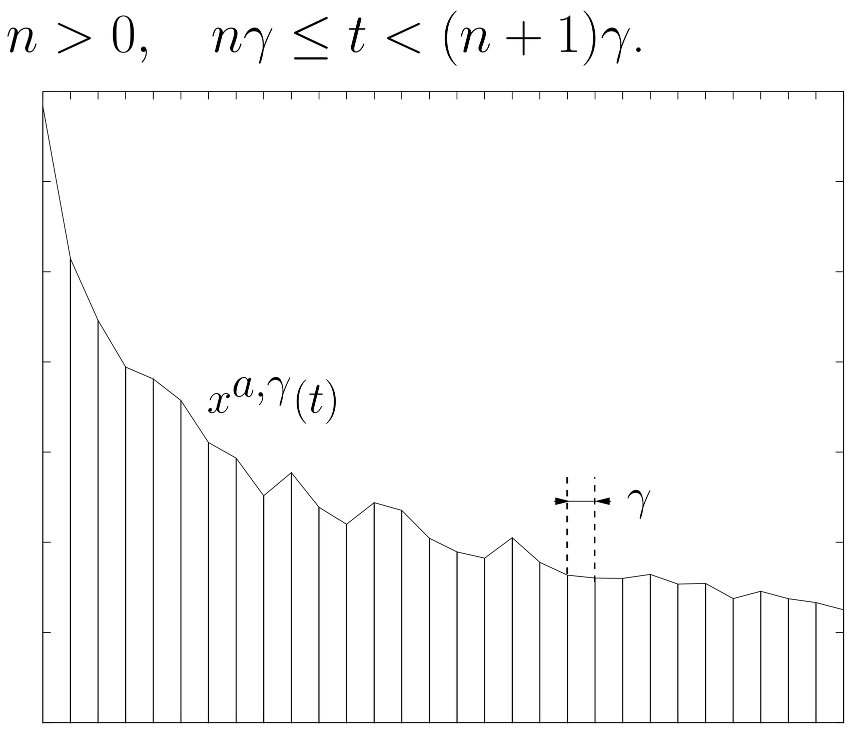

Figure 1: The linearly interpolated process of order γ: x_γ

Theorem

Under mild assumptions,

$$\mathsf{x}_\gamma \longrightarrow_{\gamma \to 0} \mathsf{x},$$

weakly where x *satisfies the Differential Inclusion ([4])*

$$\dot{\mathsf{x}}(t) \in \nabla F(\mathsf{x}(t)) + \partial R(\mathsf{x}(t)), \quad t \ge 0.$$

Long-run behavior

Known fact :

$$\mathsf{x}(t) \longrightarrow_{t \to +\infty} \arg\min F + R$$

We would like x_γ to "inherits" this property. OK under stability of the Markov chain (x_n^γ).

Theorem

Assume moreover

- $F(x) + R(x) \longrightarrow_{\|x\| \to +\infty} +\infty$
- $\nabla f(\cdot, \xi)$ *is Lipschitz continuous.*

Then, for every $\varepsilon > 0$,

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

Return to the Overlapping Group Lasso

Two Douglas Rachford strategies to solve the problem defined by (1)–(3).

First, Partially Stochastic Douglas Rachford

1. Sample $i_{n+1} \sim U(\{1, \ldots, N\})$
2. Compute $\operatorname{prox}_{\gamma f_{i_{n+1}}}$ using [2]
3. Compute $\operatorname{prox}_{\gamma R}$ using [5]

Second, Stochastic Douglas Rachford

1. Sample $i_{n+1} \sim U(\{1, \ldots, N\})$
2. Sample $g_{n+1} \sim U(\mathcal{G})$
3. Compute $\operatorname{prox}_{\gamma f_{i_{n+1}}}$ using [2]
4. Compute $\operatorname{prox}_{\gamma r_{g_{n+1}}}$ (easy, soft thresholding)

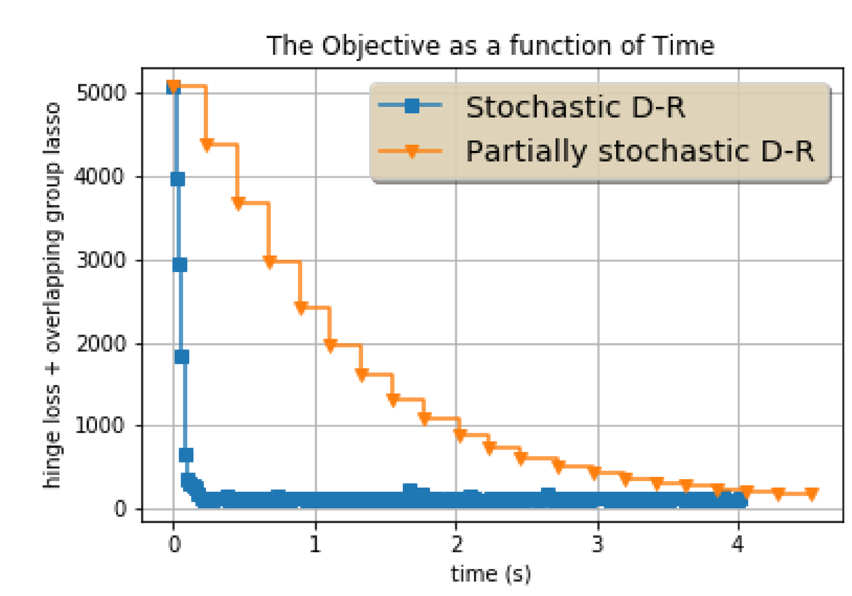

Figure 2: $F + R$ as a function of time for the Stochastic Douglas Rachford and the Partially Stochastic Douglas Rachford algorithms

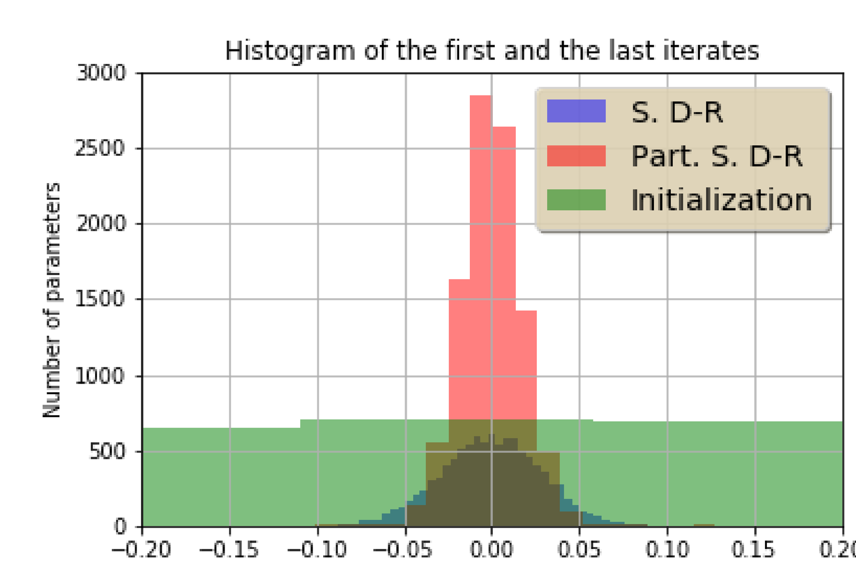

Figure 3: Histogram of the initialization and the last iterates of the two algorithms

References

[1] P.-L. Lions and B. Mercier.
Splitting algorithms for the sum of two nonlinear operators.
SIAM J. Numer. Anal., 16(6):964–979, 1979.

[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] 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.

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

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