Random Block-Coordinate Douglas-Rachford Splitting for Binary Logistic Regression

Giovanni CHIERCHIA[1], Afef CHERNI[1,2], Émilie CHOUZENOUX[1,3], and Jean-Christophe PESQUET[3]

1 - Université Paris Est, LIGM UMR 8049, CNRS, ENPC, ESIEE Paris, UPEM, Noisy-le-Grand, France
2 - IGBMC, CNRS UMR 7104, Inserm U 964, Illkirch-Graffenstaden, France
3 - Centre pour la Vision Numérique, CentraleSupélec, INRIA Saclay, Châtenay-Malabry, France

In collaboration with

G. Chierchia

A. Cherni

J.-C. Pesquet

Linear binary classifier

Goal: Learn a function $d : \mathbb{R}^N \mapsto \{-1, +1\}$ from L training examples.

$$\mathcal{S} = \big\{(x_\ell, y_\ell) \in \mathbb{R}^N \times \{-1, +1\} \mid \ell \in \{1, \dots, L\}\big\},$$

- **Model form in linear classification:**

$$d_w(x) = \text{sign}\big(x^\top w\big)$$

  where $w \in \mathbb{R}^N$ is a vector of parameters for the classifier, to be estimated from the training set.

- **Geometric intuition:** The coefficients of w specify a hyperplane (linear separator) that separates points into -1 versus $+1$ class.

Risk minimization problem

MINIMIZATION PROBLEM

$$\underset{w \in \mathbb{R}^N}{\text{minimize}} \;\; f(w) + \sum_{\ell=1}^{L} h\left(y_\ell \, x_\ell^\top w\right),$$

- $h \in \Gamma_0(\mathbb{R})$: loss function.

Examples: quadratic, hinge, smoothed hinge, Huber, logistic [Bartlett et al., 2006] [Parikh and Boyd, 2014][Rosasco et al., 2004].

- $f \in \Gamma_0(\mathbb{R}^N)$: sparse regularization term.

Examples: ℓ_1, group-Lasso, non-convex potential [Bradley and Mangasarian, 1998][Laporte et al., 2014][Meier et al., 2008].

* $\Gamma_0(\mathcal{H})$ is the set of convex, lower semi-continuous proper functions of the Hilbert space $\mathcal{H}$.

Risk minimization problem

MINIMIZATION PROBLEM

$$\underset{w\in\mathbb{R}^N}{\text{minimize}} \;\; f(w) + \sum_{\ell=1}^{L} h\left(y_\ell\, x_\ell^\top w\right)$$

Challenges:

✗ Very large size L of the training set

$\rightsquigarrow$ Random block-alternating strategy

✗ Possibly non-smooth regularization function f

$\rightsquigarrow$ Proximal minimization algorithm

✗ Slow convergence rate when treating h through its gradient

$\rightsquigarrow$ Primal-dual scheme

Proposed approach

Proximity operator

Let $f \in \Gamma_0(\mathbb{R}^N)$.

CHARACTERIZATION OF PROXIMITY OPERATOR

$$(\forall v \in \mathbb{R}^N) \quad \widehat{w} = \mathrm{prox}_f(v) \Leftrightarrow v - \widehat{w} \in \partial f(\widehat{w}).$$

The proximity operator $\mathrm{prox}_f(v)$ of f at $v \in \mathbb{R}^N$ is the unique vector $\widehat{w} \in \mathbb{R}^N$ such that

$$f(\widehat{w}) + \frac{1}{2}\|\widehat{w} - v\|^2 = \inf_{w \in \mathbb{R}^N} f(w) + \frac{1}{2}\|w - v\|.$$

Proposed random block-coordinate strategy

MINIMIZATION PROBLEM

$$\underset{w\in\mathbb{R}^N}{\text{minimize}} \;\; f(w) + \sum_{\ell=1}^{L} h\left(y_\ell\, x_\ell^\top w\right)$$

General idea: At each iteration $i \in \mathbb{N}$, select randomly a subset $(x_\ell, y_\ell)_{\ell\in\mathbb{L}_i}$ of $\mathcal{S}$ with $\mathbb{L}_i \subset \{1,\ldots,L\}$, using the Douglas-Rachford proximal splitting scheme from [Combettes and Pesquet, 2016].

Proposed random block-coordinate strategy

Minimization problem

$$\underset{w \in \mathbb{R}^N}{\text{minimize}} \;\; f(w) + \sum_{\ell=1}^{L} h\left(y_\ell \, x_\ell^\top w\right)$$

General idea: At each iteration $i \in \mathbb{N}$, select randomly a subset $(x_\ell, y_\ell)_{\ell \in \mathbb{L}_i}$ of $\mathcal{S}$ with $\mathbb{L}_i \subset \{1, \ldots, L\}$, using the Douglas-Rachford proximal splitting scheme from [Combettes and Pesquet, 2016].

Related works:

* Coordinate ascent method [Shalev-Shwartz and Tewari, 2011]
* Stochastic forward-backward strategy [Combettes and Pesquet, 2015][Rosasco et al., 2016]
* Regularized dual ascent approach [Xiao, 2010]
* Stochastic primal-dual proximal algorithms [Chierchia et al., 2015][Pesquet and Repetti, 2016]

Minimization algorithm

$Q = \left(\mathrm{Id} + \sum_{\ell=1}^{L} x_\ell x_\ell^\top\right)^{-1}$

$t^{[0]} \in \mathbb{R}^N, (v_1^{[0]}, \dots, v_L^{[0]}) \in \mathbb{R}^L, u^{[0]} = \sum_{\ell=1}^{L} y_\ell x_\ell \, v_\ell^{[0]}$

$\gamma \in \left]0, +\infty\right[, \mu \in \left]0, 2\right[$

For $i = 0, 1, \dots$

- Select $\mathbb{L}_i \subset \{1, \dots, L\}$
- $w^{[i]} = Q\big(t^{[i]} + u^{[i]}\big)$
- $t^{[i+1]} = t^{[i]} + \mu\Big(\mathrm{prox}_{\gamma f}(2w^{[i]} - t^{[i]}) - w^{[i]}\Big)$
- $(\forall \ell \in \mathbb{L}_i) \quad v_\ell^{[i+1]} = v_\ell^{[i]} + \mu\Big(\mathrm{prox}_{\gamma h}(2y_\ell x_\ell^\top w^{[i]} - v_\ell^{[i]}) - y_\ell x_\ell^\top w^{[i]}\Big)$
- $(\forall \ell \notin \mathbb{L}_i) \quad v_\ell^{[i+1]} = v_\ell^{[i]}$
- $u^{[i+1]} = u^{[i]} + \sum_{\ell \in \mathbb{L}_i} \big(v_\ell^{[i+1]} - v_\ell^{[i]}\big) y_\ell x_\ell.$

Convergence result

Assume that the following conditions hold:

- The set of solutions $\mathcal{F}$ of the problem is nonempty;
- $t^{[0]}$ is a $\mathbb{R}^N$-valued random variable, and $(v_1^{[0]}, \ldots, v_L^{[0]})$ is an $\mathbb{R}^L$-valued random variable;
- The $(\mathbb{L}_i)_{i\in\mathbb{N}}$ are drawn in an independent and identical manner.

Then, $(w^{[i]})_{i\in\mathbb{N}}$ converges almost surely to an element of $\mathcal{F}$. Moreover, consider $\mathcal{F}^*$ the set of solutions to the associated dual problem. Then the sequence $(\gamma^{-1}[y_1 x_1^\top w - v_1, \ldots, y_L x_L^\top w - v_L])_{i\in\mathbb{N}}$ converges almost surely to an element of $\mathcal{F}^*$.

✓ The convergence result still holds when the involved proximity operators are computed up to summable errors.

Application to binary logistic regression

Binary logistic regression

Goal: Maximize the posterior probability of the weights given the training data i.e, optimize the product of the weight prior probability and the conditional data likelihood :

$$\underset{w\in\mathbb{R}^N}{\text{maximize}} \quad \varphi(w)\prod_{\ell=1}^{L}\pi(y_\ell\,|\,x_\ell,w)\theta_\ell(x_\ell|w).$$

⬇

Binary logistic loss

$$(\forall v\in\mathbb{R}) \qquad h(v)=\log\Big(1+\exp(-v)\Big).$$

Proximity operator of the binary logistic loss

Let $\gamma \in]0, +\infty[$. The proximity operator of the logistic loss is

$$(\forall v \in \mathbb{R}) \quad \mathrm{prox}_{\gamma h}(v) = v + \mathrm{W}_{\exp(-v)}\Big(\gamma \exp(-v)\Big),$$

- Hereabove, $\mathrm{W}_{\cdot}$ is the **generalized W-Lambert function** from [Mező et al., 2014], which solves transcendental equations in the form:

$$(\forall \bar{v} \in \mathbb{R})(\forall v \in \mathbb{R})(\forall r \in]\exp(-2), +\infty[)$$
$$\bar{v} \exp(\bar{v}) + r\bar{v} = v \quad \Leftrightarrow \quad \bar{v} = \mathrm{W}_r(v).$$

- This function can be efficiently evaluated through a Newton-based method devised by Mező *et al.*

Proximity operator of the binary logistic loss

Let $\gamma \in \,]0,+\infty[$. The proximity operator of the logistic loss is

$$(\forall v \in \mathbb{R}) \quad \operatorname{prox}_{\gamma h}(v) = v + \mathrm{W}_{\exp(-v)}\Big(\gamma \exp\left(-v\right)\Big),$$

- Exponentiation leads to arithmetic overflow when v tends to minus infinity.

  $\rightsquigarrow$ Study of **asymptotic behaviour**:

  Let $\gamma \in \,]0,+\infty[$. Then,

  $$\operatorname{prox}_{\gamma h}(v) \sim_{v \to -\infty} v + \gamma\big(1 - \exp(\gamma + v)\big).$$

- Similar results available for the Fenchel conjugate function of h.

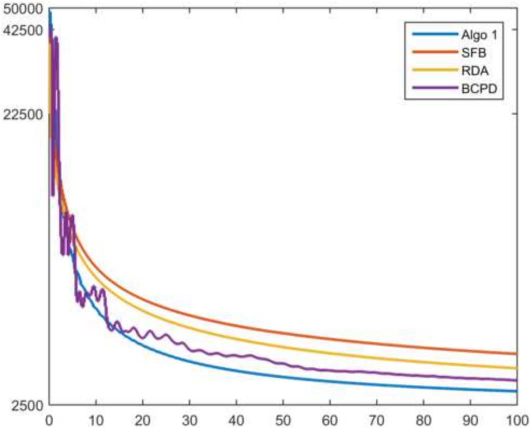
Experimental results

Experimental results

- Two standard data sets: MNIST ($N = 717$, $L = 60000$) and W8A ($N = 300$, $L = 49749$);
- h= binary logistic loss, f= ℓ_1 norm (with weight $\lambda = 1$);
- Mini-batches of size 1000 randomly selected using a uniform distribution;
- Initial vector $w^{[0]}$ randomly drawn from a normal distribution with zero mean and unit variance;

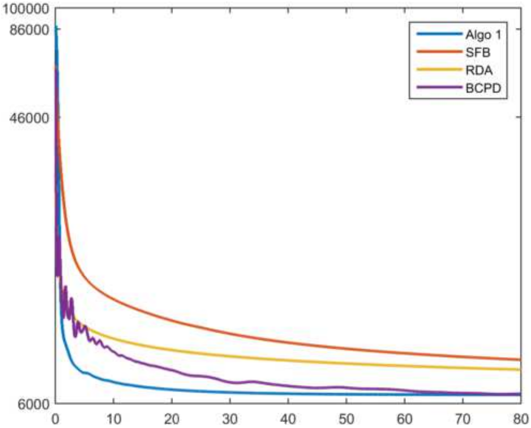
Algorithm	**Parameters**
Stochastic Forward Backward (SFB) [Rosasco et al., 2016]	$\gamma = 10^{-4}$
Regularized Dual Ascent (RDA) [Xiao, 2010]	$\gamma = 10^{-4}$
Block Coordinate primal-dual algorithm (BCPD) [Chierchia et al., 2015]	$\sigma = \tau^{-1} \left\| \sum_{\ell=1}^{L} x_\ell x_\ell^\top \right\|^{-1}$
Proposed algorithm	$\gamma = 10^{-4}$, $\mu = 1.8$

Experimental results



Evolution of the cost function along iterations for dataset MNIST

Experimental results



Evolution of the cost function along iterations for dataset W8A

Conclusion

- ✓ Proposition of a random block-coordinate Douglas-Rachford algorithm for sparse linear regression at a large scale;
- ✓ Convergence guaranteed under mild assumptions on the algorithmic parameters;
- ✓ Derivation of a closed-form expression for the proximity operator of the logistic loss;
- ✓ Training performance compares favorably to state-of-the-art stochastic methods;
- ✓ Coming soon : An improved version of the algorithm.

Bibliography

P. Combettes and J.-C Pesquet
Stochastic Quasi-Fejér Block-Coordinate Fixed Point Iterations with Random Sweeping
SIAM Journal on Optimization, 25(2), pp. 1221-1248, 2015.

I. Mezo and A. Baricz
On the generalization of the lambert W function
to appear in *Transactions of the AMS*, 2017.

J.-C. Pesquet and A. Repetti
A Class of Randomized Primal-Dual Algorithms for Distributed Optimization
Journal of Nonlinear and Convex Analysis, 16(12), pp. 2453–2490, 2015.

M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero and S. McLaughlin
A Survey of Stochastic Simulation and Optimization Methods in Signal Processing
IEEE Journal of Selected Topics in Signal Processing, 10(2), pp. 224-241, Mar. 2016.

L. Rosasco, S. Villa, and B. C. Vu
Stochastic forward-backward splitting for monotone inclusions
Journal of Optimization Theory and Applications, 169(2), pp. 388–406, May 2016.

L. Xiao
Dual averaging methods for regularized stochastic learning and online optimization
Journal of Machine Learning Research, 11, pp. 2543–2596, Oct. 2010.