ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

# Reweighted $l_1$ Dual Averaging Approach for Sparse Stochastic Learning

Vilen Jumutc and Johan A.K. Suykens

KU Leuven, Department of Electrical Engineering, ESAT-STADIUS
Kasteelpark Arenberg 10, B-3001, Leuven, Belgium

**Abstract**. Recent advances in stochastic optimization and regularized dual averaging approaches revealed a substantial interest for a simple and scalable stochastic method which is tailored to some more specific needs. Among the latest one can find sparse signal recovery and $l_0$-based sparsity inducing approaches. These methods in particular can force many components of the solution shrink to zero thus clarifying the importance of the features and simplifying the evaluation. In this paper we concentrate on enhancing sparsity of the recently proposed $l_1$ Regularized Dual Averaging (RDA) method with a simple reweighting iterative procedure which in a limit applies the $l_0$-norm penalty. We present some theoretical justifications of a bounded regret for a sequence of convex repeated games where every game stands for a separate reweighted $l_1$-RDA problem. Numerical results show an enhanced sparsity of the proposed approach and some improvements over the $l_1$-RDA method in generalization error.

## 1 Introduction

Numerous results and publications on stochastic learning revealed a substantial interest for extending learning paradigms in stochastic optimization with different notions of sparsity (parsimony) inducing norms [1, 2] and outlier-ablating loss functions [3]. In retrospect we can see an increasing importance of correct sparsity patterns and proliferation of soft-thresholding methods [1, 4] in achieving a good and approximately sparse solution. There are many important contributions of the parsimony concept to the machine learning field, e.g. enhanced interpretability of a solution or simplified and easy to compute linear models. Although methods, such as Lasso and Elastic Net, were investigated in the context of stochastic optimization in many papers [1, 4] we are not aware of any $l_0$-penalty inducing approaches which were applied in this particular setting.

Recently Candès et al. [5] proposed an approximation to the $l_0$-norm through an iteratively reweighted $l_1$ minimization. In this paper we intend to close the gap and introduce a novel view on enhancing sparsity of stochastic models through a sequence of convex repeated games [6]. In this general setting we assume a composite optimization objective of the form $\phi_t(w) \triangleq \mathbb{E}_\xi[F(w,\xi)] + \lambda\psi_t(w)$, where $\xi = (x, y)$ is a random pair (input-output observation) drawn from an unknown underlying distribution and both $f_t(w) \triangleq \mathbb{E}_\xi[F(w,\xi)]$ and $\psi_t(w)$ are related to some convex but possibly non-smooth functions. The solution of every optimization problem in our approach is treated as a hypothesis of a learner at iteration $t$ induced by a loss function $f_t(w) \triangleq \mathbb{E}_{\mathcal{A}_t}[l(w;\mathcal{A}_t)]$ on a particular question-answer subset $\mathcal{A}_t \in \mathcal{S}, |\mathcal{A}_t| = k$ of pairs $\{(x_{1t}, y_{1t}), \ldots, (x_{kt}, y_{kt})\}$ and

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

regularized by a reweighted function $\psi_t(w) \equiv \psi_t(w; \Theta_t)$, where $\Theta_t$ is our diagonal reweighting matrix at iteration $t$. In the sequel we extend the recent work of Xiao [4] on $l_1$-Regularized Deal Averaging (RDA) and present some theoretical regret bounds for learning sparser model representations through a reweighted iterative $l_1$-minimization.

This paper is structured as follows. Section 2 describes our re-weighted stochastic $l_1$-RDA method. Section 3 gives an upper bound on a regret of the sequence of convex repeated games. Section 4 presents our numerical results while Section 5 concludes the paper.

## 2 Method

### 2.1 Problem definition

In the stochastic $l_1$-RDA approach developed by Xiao [4] we are approximating the expected loss function $f_t(w)$ by using a finite set of independent observations $\xi_1, \dots, \xi_T$, and we are minimizing the following optimization objective:

$$\min_w \frac{1}{T} \sum_{t=1}^{T} f_t(w, \xi_t) + \Psi(w). \tag{1}$$

We are dealing with the $l_1$-norm and $\Psi(w) \triangleq \lambda \|w\|_1$, where $\lambda$ is a trade-off constant. In our particular approach we will replace it with the re-weighted version $\Psi_t(w) \triangleq \lambda \psi_t(w; \Theta_t) = \lambda \|\Theta_t w\|_1$ which in the limit applies an approximation to the $l_0$-norm penalty. At every iteration $t$ we will be solving a separate convex instantaneous optimization problem (game) conditioned on a diagonal reweighting matrix $\Theta_t$.

According to a *simple dual averaging* scheme [7] and intuition given by [4] we can approach our primal solution using the following sequence of iterates $w_{t+1}$ :

$$w_{t+1} = \arg\min_w \{ \frac{1}{t} \sum_{\tau=1}^{t} \langle g_\tau, w \rangle + \Psi_t(w) + \frac{\beta_t}{t} h(w) \}, \tag{2}$$

where $h(w)$ is an auxiliary $\sigma$-strongly convex smoothing term, $g_t \in \partial f_t(w_t)$ represents a subgradient, and $\{\beta_t\}_{t \geq 1}$ is a non-negative and non-decreasing input sequence, which determines the convergence properties of the algorithm. For our re-weighted $l_1$-regularized dual averaging approach we set $\beta_t = \gamma \sqrt{t}$ and we replace $h(w)$ with a parameterized version:

$$h(w) = \frac{1}{2} \|w\|_2^2 + \rho \|w\|_1, \tag{3}$$

where the initial parameters of the enhanced $l_1$-RDA method in [4] remain unchanged. Hence Eq.(2) becomes:

$$w_{t+1} = \arg\min_w \{ \frac{1}{t} \sum_{\tau=1}^{t} \langle g_\tau, w \rangle + \lambda \|\Theta_t w\|_1 + \frac{\gamma}{\sqrt{t}} (\frac{1}{2} \|w\|_2^2 + \rho \|w\|_1) \}. \tag{4}$$

Each iterate has a closed form solution. Let us define $\eta_t^{(i)} = \Theta_t^{(ii)}\lambda + \gamma\rho/\sqrt{t}$ and give an entry-wise solution by:

$$w_{t+1}^{(i)} = \begin{cases} 0, & \text{if } |\hat{g}_t^{(i)}| \leq \eta_t^{(i)} \\ -\frac{\sqrt{t}}{\gamma}(\hat{g}_t^{(i)} - \eta_t^{(i)}\text{sign}(\hat{g}_t^{(i)})), & \text{otherwise} \end{cases}, \tag{5}$$

where $\hat{g}_t^{(i)} = \frac{t-1}{t}\hat{g}_{t-1}^{(i)} + \frac{1}{t}g_t^{(i)}$ is the $i$-th component of the averaged $g_t \in \partial f_t(w_t)$.

### 2.2 Algorithm

In this subsection we will outline our main algorithmic scheme. It consists of a simple initialization step, computation and averaging of the subgradient $g_t$, evaluation of the iterate $w_{t+1}$ and finally re-computation of the reweighting matrix $\Theta_{t+1}$.

**Algorithm 1:** Stochastic Reweighted $l_1$-Regularized Dual Averaging

**Data**: $\mathcal{S}, \lambda > 0, \gamma > 0, \rho \geq 0, \epsilon > 0, T > 1, k \geq 1, \varepsilon > 0$

Set $w_1 = 0, \hat{g}_0 = 0, \Theta_1 = diag([1, \ldots, 1])$
**for** $t = 1 \rightarrow T$ **do**
  Select $\mathcal{A}_t \subseteq \mathcal{S}$, where $|\mathcal{A}_t| = k$
  Calculate $g_t \in \partial f_t(w_t; \mathcal{A}_t)$
  Compute the dual average $\hat{g}_t = \frac{t-1}{t}\hat{g}_{t-1} + \frac{1}{t}g_t$
  Compute the next iterate $w_{t+1}$ by Eq.(5)
  Re-calculate the next $\Theta$ by $\Theta_{t+1}^{(ii)} = 1/(|w_{t+1}^{(i)}| + \epsilon)$
  **if** $\|w_{t+1} - w_t\| \leq \varepsilon$ **then**
    **return** $w_{t+1}$
  **end**
**end**
**return** $w_{T+1}$

From Algorithm 1 we can clearly see that it can operate in a stochastic ($k = 1$) and semi-stochastic mode ($k > 1$). We do not restrict ourselves to a particular choice of the loss function $f_t(w)$. In comparison with the $l_1$-RDA approach we have one additional input parameter $\epsilon$, which should be tuned or selected properly as described in [5].

## 3 Analysis

In this section we will briefly[1] discuss some of our convergence results and upper bounds for Algorithm 1. We concentrate mainly on the regret *w.r.t.* function $\phi_\tau(w)$, such that for all $w \in \mathbb{R}^n$ we have:

$$R_t(w) = \sum_{\tau=1}^{t}(\phi_\tau(w_\tau) - \phi_\tau(w)). \tag{6}$$

[1] due to the space limitations

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

From [7] and [4] we know that if we consider $\Delta\psi_\tau = \psi_\tau(w_\tau) - \psi_\tau(w)$ the following *gap sequence* $\delta_t$ holds:

$$\delta_t = \max_w \{ \sum_{\tau=1}^{t} (\langle g_\tau, w_\tau - w\rangle + \Delta\psi_\tau)\} \geq \sum_{\tau=1}^{t} (f_\tau(w_\tau) - f_\tau(w) + \Delta\psi_\tau) = R_t(w) \quad (7)$$

which due to the convexity of $f_\tau$ bounds the regret function from above [8]. Hence by ensuring the necessary condition of Eq.(49) in [4] we can show the upper bound on $\delta_t$ which immediately implies the same bound on $R_t(w)$.

**Theorem 1** *Let the sequences* $\{w_t\}_{t\geq 1}$*,* $\{g_t\}_{t\geq 1}$ *and* $\{\Theta_t\}_{t\geq 1}$ *be generated by the Algorithm 1. Assume* $\|\Theta_{t+1} w\|_1 \geq \|\Theta_t w\|_1$ *for any* $w \in \mathbb{R}^n$*,* $\psi_{t+1}(w_{t+1}) \leq \psi_t(w_t)$*,* $\|g_t\|_* \leq G$ *and* $h(w_t) \leq D$*, where* $\|\cdot\|_*$ *stands for the dual norm. Then:*

$$R_t(w) \leq (\gamma D + \frac{G^2}{\gamma})\sqrt{t}. \quad (8)$$

*Proof:* We start with redefining a conjugate-type functions $V_t(s)$ and $U_t(s)$ in [4] and replacing $\Psi(w)$ in each of them with our reweighted $l_1$ function $\lambda\|\Theta_1 x\|_1$. In Eq.(7) we can separate and bound the maximization part:

$$\max_w \{ \sum_{\tau=1}^{t} (\langle g_\tau, w_0 - w\rangle - \psi_\tau(w))\} \leq \max_w \{ \sum_{\tau=1}^{t} \langle g_\tau, w_0 - w\rangle - t\psi_1(w)\}, \quad (9)$$

`iff` $\|\Theta_{t+1} x\|_1 \geq \|\Theta_t x\|_1$. The right hand side of Eq.(9) is exactly $U_t(s)$ in [4]. On the other hand our second assumption guarantees Eq.(49) in [4] because $V_t(-s_t) + \psi_t(w_t) \leq V_t(-s_t) + \psi_1(w_1) \leq V_{t-1}(-s_t)$. All together this guarantees the bound on $\delta_t$ sequence motivated by Eq.(2.15) in [7] and thoroughly discussed in Appendix B of [4]. This bound immediately implies Corollary 2 of [4]. □

Our intuition is related to the asymptotic convergence properties of an iterative reweighting procedure discussed in [9] where with each iterate of $\Theta_t$ our approximated norm becomes $\|\Theta_t w\|_1 \simeq \|w\|_p$ with $p \to 0$ thus making it closer to the $l_0$-norm. In return this implies $p_{t+1} \leq p_t$ and $\|w\|_{p_{t+1}} \geq \|w\|_{p_t}$. In the next theorem we will slightly relax the necessary conditions in order to derive a new bound *w.r.t.* the maximal discrepancy of $\Theta_t$ and $\psi_t(w_t)$ iterates.

**Theorem 2** *Let the sequences* $\{w_t\}_{t\geq 1}$*,* $\{g_t\}_{t\geq 1}$ *and* $\{\Theta_t\}_{t\geq 1}$ *be generated by the Algorithm 1. Assume* $\psi_t(w) = \lambda\|\Theta_t w\|_1$ *,* $\|\Theta_t w\|_1 - \|\Theta_{t+\tau} w\|_1 \leq \nu_1/\tau$ *and* $\psi_{t+\tau}(w_{t+\tau}) - \psi_t(w_t) \leq \nu_2/\tau$ *for any* $\tau \geq 1$*,* $\nu_1, \nu_2 \geq 0$*,* $\lambda > 0$ *and* $w \in \mathbb{R}^n$*,* $\|g_t\|_* \leq G$ *and* $h(w_t) \leq D$*, where* $\|\cdot\|_*$ *stands for the dual norm. Then:*

$$R_t(w) \leq \log(t)(\lambda\nu_1 + \nu_2) + (\gamma D + \frac{G^2}{\gamma})\sqrt{t}. \quad (10)$$

*Proof:* The outline of the proof is the same except for the adjusted Eq.(9):

$$\max_w \{ \sum_{\tau=1}^{t} (\langle g_\tau, w_0 - w\rangle - \psi_\tau(w))\} \leq \max_w \{ \sum_{\tau=1}^{t} (\langle g_\tau, w_0 - w\rangle - (\psi_1(w) - \lambda\nu_1/\tau))\}, \quad (11)$$

which in return implies the additional $\lambda\nu_1\log(t)$ term in Lemma 9 of [4]. □

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

# 4 Experiments

## 4.1 Setup

For all our experiments we are comparing linear models with the hinge loss. For tuning $\lambda, \gamma, \rho$ hyperparameters in Algorithm 1 we used a 2-step procedure. This procedure consists of the DFO-based[2] global optimization technique: Randomized Direct Search (RDS) [10] and the simplex method [11] for the second step. We perform 10-fold cross-validation at each step.

All experiments with large-scale UCI datasets (Table 1) were repeated 50 times with the random split to the training and test sets in proportion 9:1. In the presence of 3 or more classes we performed binary classification where we learned to classify the first class versus all others. For Algorithm 1 we fixed parameters: $T = 1000$, $k = 1$, $\epsilon = 10^{-2}$ and $\varepsilon = 10^{-5}$.

Table 1: UCI Datasets

| Dataset | # of attributes | # of classes | # of data points |
|---|---|---|---|
| Pen Digits | 16 | 10 | 10992 |
| Opt Digits | 64 | 10 | 5620 |
| Semeion | 256 | 10 | 1593 |
| Spambase | 57 | 2 | 4601 |
| Magic | 11 | 2 | 19020 |
| Shuttle | 9 | 2 | 58000 |
| Skin | 4 | 2 | 245057 |
| Covertype | 54 | 7 | 581012 |

Table 2: Performance and sparsity

| Dataset | Test error | | | Sparsity $\sum_i I(\lvert w_i\rvert > 0)/d$ | |
|---|---|---|---|---|---|
| | (re)$l_1$-RDA | $l_1$-RDA | Pegasos | (re)$l_1$-RDA | $l_1$-RDA |
| Pen Digits | 0.082 | 0.073 | **0.066** | 0.186 | 0.350 |
| Opt Digits | 0.050 | 0.061 | **0.037** | 0.165 | 0.246 |
| Semeion | 0.042 | **0.039** | 0.088 | 0.124 | 0.182 |
| Spambase | 0.116 | 0.160 | **0.099** | 0.321 | 0.412 |
| Shuttle | 0.067 | 0.077 | **0.062** | 0.307 | 0.493 |
| Magic | **0.225** | 0.251 | 0.276 | 0.290 | 0.452 |
| Skin | **0.066** | 0.083 | 0.115 | 0.680 | 0.713 |
| Covertype | **0.269** | 0.284 | 0.281 | 0.132 | 0.163 |

## 4.2 Results

In this subsection we present some numerical results and evaluations. As we can notice from Table 2 our proposed reweighted modification of the $l_1$-RDA approach achieves better sparsity patterns on every dataset while for some of them it can even attain better generalization performance. For the sake of completeness we provide values for the $l_2$-norm approach, namely Pegasos [12], which is an SGD-based stochastic optimization routine but without sparsity

[2]Derivative-Free Optimization

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

inducing capabilities. Comparing test errors of the Reweighted $l_1$-RDA with other approaches for some datasets we can observe a minor degradation of the performance for attaining a sparser solution.

# 5 Conclusions

In this paper we considered the problem of approximating the $l_0$-norm penalty in the context of linear models and sparse stochastic learning via reweighted $l_1$ minimization. We studied a simple dual averaging scheme for optimization which enabled us with an elegant and complementary analysis for the regret bounds. We were able to show that under certain conditions we have a bounded regret even if the convergence of our auxiliary reweighting iterate $\Theta_t$ is ill-conditioned. Numerical results demonstrate the advantages of the proposed approach both in terms of the generalization error and sparsity over the original $l_1$-RDA method.

**Acknowledgements:** The work is supported by Research Council KUL, ERC AdG A-DATADRIVE-B, GOA/10/09MaNet, CoE EF/05/006, FWO G.0588.09, G.0377.12, SBO POM, IUAP P6/04 DYSCO.

# References

[1] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l1 regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 929–936, New York, NY, USA, 2009. ACM.

[2] Nicolas Le Roux, Mark W. Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2672–2680, 2012.

[3] Vilen Jumutc, Xiaolin Huang, and Johan A. K. Suykens. Fixed-size pegasos for hinge and pinball loss svm. In *2013 International Joint Conference on Neural Networks (IJCNN 2013)*, pages 1122–1128, 2013.

[4] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, December 2010.

[5] Emmanuel Candès, Michael Wakin, and Stephen Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

[6] Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and fenchel duality. In Bernhard Schölkopf, John Platt, and Thomas Hoffman, editors, *NIPS*, pages 1265–1272. MIT Press, 2006.

[7] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[8] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[9] Kaizhu Huang, Irwin King, and Michael R. Lyu. Direct zero-norm optimization for feature selection. In *ICDM*, pages 845–850. IEEE Computer Society, 2008.

[10] Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, USA, 2009.

[11] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[12] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 807–814, New York, NY, USA, 2007.