Reweighted l_2-Regularized Dual Averaging Approach for Highly Sparse Stochastic Learning

Vilen Jumutc and Johan A.K. Suykens

KU Leuven, ESAT-STADIUS, Kasteelpark Arenberg 10, B-3001, Leuven, Belgium
{vilen.jumutc, johan.suykens}@esat.kuleuven.be

Abstract. Recent advances in dual averaging schemes for primal-dual subgradient methods and stochastic learning revealed an ongoing and growing interest in making stochastic and online approaches consistent and tailored towards sparsity inducing norms. In this paper we focus on the reweighting scheme in the l_2-Regularized Dual Averaging approach which favors properties of a strongly convex optimization objective while approximating in a limit the l_0-type of penalty. In our analysis we focus on a regret and convergence criteria of such an approximation. We derive our results in terms of a sequence of strongly convex optimization objectives obtained via the smoothing of a sub-differential and non-smooth loss function, *e.g.* hinge loss. We report an empirical evaluation of the convergence in terms of the cumulative training error and the stability of the selected set of features. Experimental evaluation shows some improvements over the l_1-RDA method in the generalization error as well.

Keywords: Stochastic learning, l_0 penalty, regularization, sparsity

1 Introduction

In this paper we investigate an interplay between l_2-Regularized Dual Averaging (RDA) approach [18] in the context of stochastic learning and parsimony concepts arising from the application of sparsity inducing norms, like the l_0-type of penalty. Learning with $\|x\|_0$ pseudonorm regularization is a NP-hard problem [10] and is feasible only via the reweighting schemes [3], [5], [16] while lacking a proper theoretical analysis of convergence in the online and stochastic learning cases. Some methods, like [7], consider an embedded approach where one has to solve a sequence of QP-problems, which might be very computationally- and memory-wise expensive while still missing some proper convergence criteria.

There are many important contributions of the parsimony concept to the machine learning field, e.g. understanding the obtained solution or simplified and easy to extract decision rules. Many methods, such as Lasso and Elastic Net, were studied in the context of stochastic and online learning in several papers [15], [18], [4] but we are not aware of any l_0-norm sparsity inducing approaches which were applied in the context of Regularized Dual Averaging and stochastic optimization.

In many existing iterative reweighting schemes [5], [9] the analysis is provided in terms of the Restricted Isometry (RIP) or the Null Space Properties (NSP) [8]. In this paper we are trying to provide a supplementary analysis and sufficient convergence criteria for learning much sparser linear Pegasos-like [14], [13] models from random observations. We use the l_2-Regularized Dual Averaging approach and a sequence of strongly convex reweighted optimization objectives to accomplish this goal. The solution of every optimization problem at iteration t in our approach is treated as a hypothesis of a learner which is induced by an expectation of a non-smooth loss function (*e.g.* hinge loss) $f(w) \triangleq \mathbb{E}_{\xi}[l(w, \xi)]$, where the expectation is taken *w.r.t.* the random sequence of observations $\xi = \{\xi_\tau\}_{1 \leq \tau \leq t}$. We regularize it by a re-weighted l_2-norm at each iteration t. This approach in case of satisfying the sufficient conditions will converge to a global optimal solution *w.r.t.* our objective and the loss function which is generating a sequence of stochastic sub-gradients endowing our dual space E^* [12].

This paper is structured as follows. Section 2 describes our reweighted l_2-RDA method. Section 2.3 gives an upper bound on a regret for the sequence of strongly convex optimization objectives under the setting of stochastic learning. Section 3 presents our numerical results and Section 4 concludes the paper.

2 Proposed Method

2.1 Problem definition

In the Regularized Dual Averaging approach for stochastic learning developed by Xiao [18] we approximate the expected loss function $f(w) \triangleq \mathbb{E}_{\xi}[l(w, \xi)]$ on a particular random question-answer sequence $\{\xi_\tau\}_{1 \leq \tau \leq t}$, where $\xi_\tau = (x_\tau, y_\tau)$ and $y_\tau \in \{-1, 1\}$. In this particular setting the loss function is regularized by a general convex penalty and hence we are minimizing the following optimization objective:

$$\begin{aligned} \min_{w} \quad & \phi(w) \\ \text{s.t.} \quad & \phi(w) \triangleq \frac{1}{t} \sum_{\tau=1}^{t} f(w, \xi_\tau) + \Psi(w), \end{aligned} \tag{1}$$

where $\Psi(w)$ can be either a strongly convex $\|\cdot\|_2$ norm or a non-smooth sparsity promoting $\|\cdot\|_1$ norm.

In our particular setting we are dealing with the squared l_2 norm and $\Psi(w) \triangleq \lambda \|w\|_2^2$. For promoting additional sparsity we add to the l_2-norm the reweighted $\|\Theta_t^{1/2} w\|_2^2$ term such that we have $\Psi_t(w) \triangleq \lambda \|w\|_2^2 + \|\Theta_t^{1/2} w\|_2^2$. At every iteration t we will be solving a separate λ-strongly convex instantaneous optimization objective conditioned on a diagonal reweighting matrix Θ_t.

To solve problem in Eq.(1) we split it into a sequence of separated optimization problems which should be cheap to compute and hence should have a closed form solution. These problems are interconnected through the sequence of dual

variables $\tilde{g}_\tau \in \partial f(w, \xi_\tau), \tau \in \overline{1,t}$ which are averaged *w.r.t.* to the current iterate t. Because we are working with the non-smooth hinge loss the reweighted l_2-regularization is imposed via a composite smoothing term which is being gradually increased with every iteration t.

According to a simple dual averaging scheme [12], [18] we can solve Eq.(1) with the following sequence of iterates w_{t+1}:

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

where $h(w)$ is an auxiliary strongly convex smoothing term and $\{\beta_t\}_{t \geq 1}$ is a non-negative and either constant or increasing input sequence, which in case of non-strongly convex $\Psi_t(w)$ function entirely determines the convergence properties of the algorithm. In our reweighted l_2-RDA approach we use a zero β_t-sequence[1] such that we omit the auxiliary smoothing term $h(w)$ which is not necessary since our $\Psi_t(w)$ function is already smooth and λ-strongly convex. Hence the solution for every iterate w_{t+1} in our approach is given by

$$w_{t+1} = \arg\min_w \{\langle \hat{g}_t, w \rangle + \|\Theta_t^{1/2} w\|_2^2 + \lambda \|w\|_2^2\}, \tag{3}$$

where for derivations we do average stochastic sub-gradients as $\hat{g}_t = \frac{1}{t} \sum_{\tau=1}^{t} \tilde{g}_\tau$. We will explain the details regarding recalculation of Θ_t in the next subsection.

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 $\tilde{g}_\tau$, evaluation of the iterate w_{t+1} and finally recalculation of the reweighting matrix Θ_{t+1}. In Algorithm 1 we do not have any explicit sparsification mechanism for the iterate w_{t+1} except for the auxiliary function "`Sparsify`" which utilizes an additional hyperparameter ε to truncate the final solution w_t or any other w below the desired number precision as follows:

$$w^{(i)} := \begin{cases} 0, & \text{if } |w^{(i)}| \leq \varepsilon, \\ w^{(i)}, & \text{otherwise,} \end{cases} \tag{4}$$

where $w^{(i)}$ is i-th component of the vector w. In general we do not restrict ourselves to a particular choice of the loss function $f(w_t, \xi_t)$ but as it was mentioned before we stick to the hinge loss for the completeness. In comparison with the simple l_2-RDA approach [18] we have one additional hyperparameter ϵ, which enters the closed form solution for w_{t+1} and should be tuned or adjusted *w.r.t.* the iterate t as described in [3] and highlighted in [2].

In Algorithm 1 we perform an optimization *w.r.t.* to the intrinsic bias term b, which doesn't enter our decision function

$$\hat{y} = \text{sign}(w^T x), \tag{5}$$

[1] we assume $\beta_0 = \lambda$ and $\beta_t = 0, t \geq 1$ for completeness

Algorithm 1: Stochastic Reweighted l_2-Regularized Dual Averaging

Data: $\mathcal{S}, \lambda > 0, k \geq 1, \epsilon > 0, \varepsilon > 0, \delta > 0$
Set $w_1 = 0, \hat{g}_0 = 0, \Theta_0 = diag([1, \ldots, 1])$
for $t = 1 \rightarrow T$ **do**
  Select $\mathcal{A}_t \subseteq \mathcal{S}$, where $|\mathcal{A}_t| = k$
  Calculate $\tilde{g}_t \in \partial f(w_t, \mathcal{A}_t)$
  Compute the dual average $\hat{g}_t = \frac{t-1}{t}\hat{g}_{t-1} + \frac{1}{t}\tilde{g}_t$
  Compute the next iterate $w_{t+1}^{(i)} = -\hat{g}_t^{(i)}/(\lambda + \Theta_t^{(ii)})$
  Recalculate the next Θ by $\Theta_{t+1}^{(ii)} = 1/((w_{t+1}^{(i)})^2 + \epsilon)$
  if $\|w_{t+1} - w_t\| \leq \delta$ **then**
    `Sparsify`(w_{t+1}, ε)
  end
end
return `Sparsify`(w_{T+1}, ε)

but is appended to the final solution w. The trick is to append every input x_t in the subset $\mathcal{A}_t$ with an additional feature column which will be set to 1. This will alleviate the decision function with an offset in the input space. Empirically we have verified that sometimes this design has a crucial influence on the performance of a linear classifier.

2.3 Theoretical guarantees

In this subsection we will provide the theoretical guarantees for the upper bound on the regret of the function $\phi_t(w) \triangleq f(w, \xi_t) + \Psi_t(w)$, such that for any $w \in \mathbb{R}^n$ we have:

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

In this case we are interested in the guaranteed boundedness of the sum generated by this function applied to the sequences $\{\xi_1, \ldots, \xi_t\}$ and $\{\Theta_1, \ldots, \Theta_t\}$. From [12] and [18] we know that a particular gap function defined as $\delta_t = \max_w\{\sum_{\tau=1}^{t}(\langle \tilde{g}_\tau, w_\tau - w\rangle + \Psi_t(w_t) - \Psi_t(w))\}$ is an upper bound for the regret

$$\delta_t \geq \sum_{\tau=1}^{t} (\phi_\tau(w_\tau) - \phi_\tau(w)) = \mathbf{R}_t(w) \tag{7}$$

due to the convexity of $f(w, \xi_t)$ [1]. In the next theorem we will provide the sufficient conditions for the boundedness of δ_t if the imposed regularization is given by the reweighted λ-strongly convex term $\|\Theta_t^{1/2} w\|_2^2 + \lambda \|w\|_2^2$. Due to the page limitations the proof of the following theorem is not included hereafter but provided online[2].

[2] `ftp://ftp.esat.kuleuven.be/pub/stadius/vjumutc/proofs/proofs_rl2rda.pdf`

Theorem 1. *Let the sequences* $\{w_t\}_{t\geq 1}$, $\{\hat{g}_t\}_{t\geq 1}$ *and* $\{\Theta_t\}_{t\geq 1}$ *be generated by Algorithm 1. Assume* $\|\Theta_{t+1}^{1/2}w\|_2 \geq \|\Theta_t^{1/2}w\|_2$ *for any* $w \in \mathbb{R}^n$, $\Psi_t(w_t) \leq \Psi_1(w_1)$, $\|g_t\|_* \leq G$, *where* $\|\cdot\|_*$ *stands for the dual norm and constant* $\lambda > 0$ *is given for all* $\Psi_t(w)$. *Then:*

$$\mathbf{R}_t(w) \leq \frac{G^2}{2\lambda}(1 + \log(t)). \tag{8}$$

Our intuition is related to the asymptotic convergence properties of an iterative reweighting procedure discussed in [7] where with each iterate of Θ_t our approximated norm becomes $\|\Theta_t w\|_2 \simeq \|w\|_p$ with $p \to 0$ thus in a limit applying the l_0-type of a penalty. This implies $p_{t+1} \leq p_t$ and $\|w\|_{p_{t+1}} \geq \|w\|_{p_t}$. In the next theorem we will relax the sufficient conditions on $\Psi_t(w_t)$ and Θ_t. This will introduce into the bound a new term which governs the accumulation of an error *w.r.t.* these conditions.

Theorem 2. *Let the sequences* $\{w_t\}_{t\geq 1}$, $\{g_t\}_{t\geq 1}$ *and* $\{\Theta_t\}_{t\geq 1}$ *be generated by Algorithm 1. Assume* $\|\Theta_t^{1/2}w\|_2 - \|\Theta_{t+\tau}^{1/2}w\|_2 \leq \nu_1/\tau$ *and* $\Psi_{t+\tau}(w_{t+\tau}) - \Psi_t(w_t) \leq \nu_2/\tau$ *for some* $\tau \geq 1$, $\nu_1, \nu_2 \geq 0$ *and* $w \in \mathbb{R}^n$, $\|g_t\|_* \leq G$, *where* $\|\cdot\|_*$ *stands for the dual norm and constant* $\lambda > 0$ *is given for all* $\Psi_t(w)$. *Then:*

$$\mathbf{R}_t(w) \leq \log(t)(\lambda\nu_1 + \nu_2) + \frac{G^2}{2\lambda}(1 + \log(t)). \tag{9}$$

The above bound boils down to the bound in Theorem 1 if we set ν_1, ν_2 to zero.

3 Simulated experiments

3.1 Experimental setup

For all methods in our experiments we use a 2-step procedure for tuning hyperparameters. This procedure consists of Coupled Simulated Annealing [17] initialized with 5 random sets of parameters for the first step and the simplex method [11] for the second step. After CSA converges to some local minima we select a tuple of hyperparameters which attains the lowest cross-validation error and start the simplex procedure to refine our selection. On every iteration step for CSA and simplex method we proceed with a 10-fold cross-validation. In l_1-RDA and our reweighted l_2-RDA we are promoting additional sparsity with a slightly modified cross-validation criteria. We introduce an affine combination of the validation error and obtained sparsity in proportion 90% : 10% where sparsity is calculated as $\sum_i I(|w^{(i)}| > 0)/d$.

All experiments with large-scale UCI datasets [6] were repeated 50 times (iterations) with the random split to training and test sets in proportion 90% : 10%. Every iteration all methods are evaluated with the same test set to provide a consistent and fair comparison in terms of the generalization error and obtained p-values of a pairwise two-sample t-test. In the presence of 3 or more classes we perform binary classification where we learn to classify the first class versus all

others. For CT slices[3] dataset we performed a binarization of an output y_i by the median value. For URI dataset we took only "Day0" subset as a probe. For evaluation of the Algorithm 1 for UCI datasets we set $T = 1000$, $k = 1$, $\delta = 10^{-5}$ and other hyperparameters λ, ϵ and ε were determined using the cross-validation tuning procedure described above. For extremely sparse datasets with $d \gg n$, like Dexter and URI we increased k by 10 times. Information on all public UCI datasets one can find in [6].

3.2 Numerical results

In this subsection we will provide an outlook on the performance of l_1-RDA, our reweighted l_2-RDA and Pegasos [14] methods. We provide the results of the Pegasos approach for the completeness and a fair comparison in terms of the affected generalization error *w.r.t.* the obtained sparsity. In Table 1 one can see generalization errors with standard deviations (in brackets) for different UCI datasets. In Table 1 one can find asterisk symbols next to the results of our

Table 1. Performance

Dataset	Generalization (test) errors					
	(re)l_2-RDA		l_1-RDA		Pegasos	
Pen Digits	0.0745**	(±0.02)	0.1043	(±0.04)	**0.0573**	(±0.02)
Opt Digits	0.0680**	(±0.03)	0.0554	(±0.03)	**0.0356**	(±0.01)
Semeion	0.0619*	(±0.03)	**0.0414**	(±0.02)	0.0549	(±0.02)
Spambase	0.1228*	(±0.02)	0.1205	(±0.02)	**0.0989**	(±0.02)
Shuttle	0.0744*	(±0.02)	0.0734	(±0.02)	**0.0488**	(±0.02)
CT slices	0.0643*	(±0.02)	0.0845	(±0.13)	**0.0478**	(±0.01)
Magic	**0.2242**	(±0.01)	0.2259	(±0.02)	0.2254	(±0.01)
CNAE-9	**0.0109****	(±0.01)	0.0172	(±0.02)	0.0448	(±0.02)
Covertype	**0.2670***	(±0.01)	0.2715	(±0.03)	0.2791	(±0.01)
Dexter	0.0922*	(±0.02)	0.0956	(±0.01)	**0.0765**	(±0.01)
URI	0.0458**	(±0.01)	0.0623	(±0.03)	**0.0388**	(±0.01)

method ((re)l_2-RDA). These symbols indicate p-values < 0.05 of a pairwise two-sample t-test on generalization errors. Here p-values are reflecting the statistical significance of having the null-hypothesis true: the equivalence of normal distributions from which the test errors are drawn. By having two asterisk symbols we assume strong presumption against null hypothesis *w.r.t.* both competing methods, and by having one asterisk symbol - to at least one of them. Analyzing Table 1 we can conclude that for the majority of UCI datasets we are doing equally good *w.r.t.* l_1-RDA method and the significance of the obtained difference is quite high. One can see that for some datasets our reweighted l_2-RDA approach is doing better than Pegasos as well. This phenomenon could be understood from the underlying sparsity pattern which is likely to be very sparse for some datasets, for instance CNAE-9.

[3] originally it is a regression problem

3.3 Sparsity and stability

In this subsection we will provide some of the findings which highlight the enhanced sparsity of the reweighted l_2-RDA approach as well as the consistency and stability for the selected set of features (dimensions). In Table 2 one can observe the evidence of an additional sparsity promoted by the reweighting procedure which in some cases significantly reduce the number of non-zeros in the obtained solution. We do not provide any results for the Pegasos-based approach because it consists of a generic l_2-norm penalty and a projection step which all together do not provide sparse solutions. In Table 2 we provide the statistical significance of the given result by an asterisk symbol. By analyzing the results on immediately imply that in cases where we are performing equally good or slightly worse the p-values are quite high. Next we perform several series of

Table 2. Sparsity $\sum_i I(|w^{(i)}| > 0)/d$

Dataset	(re)l_2-RDA		l_1-RDA	
Pen Digits	0.12*	(±0.06)	**0.09**	(±0.11)
Opt Digits	**0.16***	(±0.09)	0.24	(±0.07)
Semeion	**0.13***	(±0.08)	0.19	(±0.05)
Spambase	0.35	(±0.07)	**0.34**	(±0.08)
Shuttle	**0.32**	(±0.17)	0.32	(±0.10)
CT slices	0.26*	(±0.08)	**0.21**	(±0.05)
Magic	**0.22***	(±0.05)	0.34	(±0.15)
CNAE-9	**0.02***	(±0.01)	0.03	(±0.03)
Covertype	**0.06***	(±0.03)	0.09	(±0.06)
Dexter	**0.08***	(±0.07)	0.17	(±0.06)
URI	**0.0012***	(±0.0011)	0.0027	(±0.0007)

experiments with UCI datasets to reveal the consistency and stability of our algorithm *w.r.t.* the selected sparsity patterns. For every dataset first we tune the hyperparameters with all available data. We run our reweighted l_2-RDA approach and l_1-RDA [18] method 100 times in order to collect frequencies of every feature (dimension) being non-zero in the obtained solution. In Figure 1 we present the corresponding histograms. As we can see our approach results in much more sparser solutions which are quite robust *w.r.t.* a sequence of random observations. l_1-RDA approach lacks these very important properties being relatively unstable under the stochastic setting.

In the next experiment we adopted a simulated setup from [4] and created a toy dataset of sample size 10000, where every input vector a is drawn from a normal distribution $\mathcal{N}(0, I_{d\times d})$ and the output label is calculated as follows $y = \text{sign}(a^T w_* + \epsilon)$, where $w_*^{(i)} = 1$ for $1 \leq i \leq \lfloor d/2 \rfloor$ and 0 otherwise and the noise is given by $\epsilon \sim \mathcal{N}(0, 1)$. We run each algorithm for 100 times and report the mean F1-score reflecting the performance of sparsity recovery. F1-score is defined as $2\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$, where

$$\text{precision} = \frac{\sum_{i=1}^d I(\hat{w}^{(i)}\neq 0, w_*^{(i)}=1)}{\sum_{i=1}^d I(\hat{w}^{(i)}\neq 0)}, \qquad \text{recall} = \frac{\sum_{i=1}^d I(\hat{w}^{(i)}\neq 0, w_*^{(i)}=1)}{\sum_{i=1}^d I(w_*^{(i)}=1)}.$$

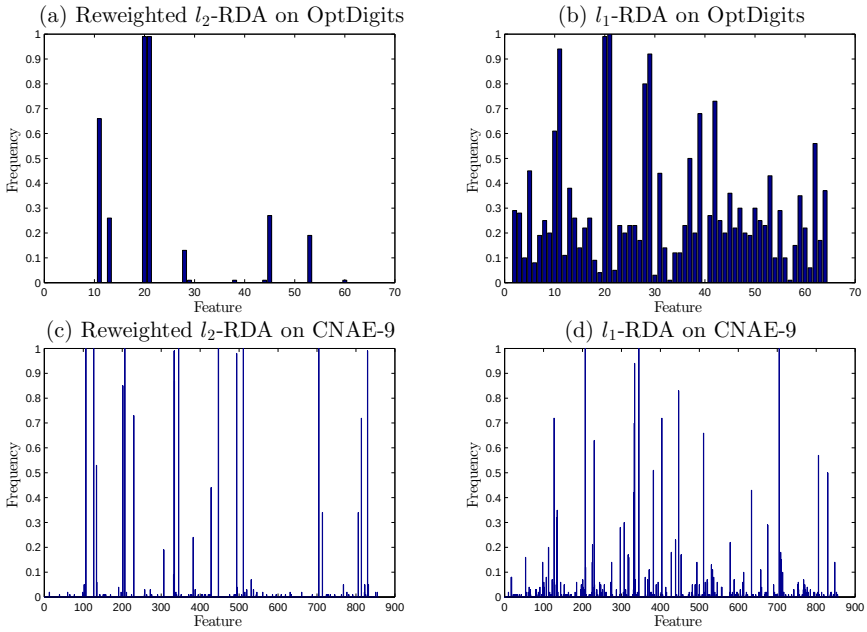

Fig. 1. Frequency of being non-zero for the features of Opt Digits and CNAE-9 datasets. In the left subfigures (a,c) we present the results for the reweighted l_2-RDA approach, while the right subfigures (b,d) correspond to l_1-RDA method.

Figure 2 shows that the reweighted l_2-RDA approach selects irrelevant features much less frequently as in comparison to l_1-RDA approach. As it was empirically verified before for UCI datasets we perform better both in terms of the stability of the selected set of features and the robustness to the stochasticity and randomness.

The higher the F1-score is, the better the recovery of the sparsity pattern. In Figure 3 we present an evaluation of our approach and l_1-RDA method *w.r.t.* to ability to identify the right sparsity pattern as the number of features increases. We clearly do outperform l_1-RDA method in terms of F1-score for $d \leq 300$. In conclusion we want to point out some of the inconsistencies that we've discovered comparing our F1-scores with [4]. Although the authors in [4] use a batch-version of the accelerated l_1-RDA method and a quadratic loss function they obtain very low F1-score (0.67) for the feature vector of size 100. In our experiments all F1-scores were above 0.7. For the dimension of size 100 our method obtains F1-score ≈ 0.95 while authors in [4] have only 0.87.

4 Conclusion

In this paper we presented a novel and promising approach, namely Reweighted l_2-Regularized Dual Averaging. This approach helps to approximate very efficient l_0-type of a penalty using a proven and reliable simple dual averaging

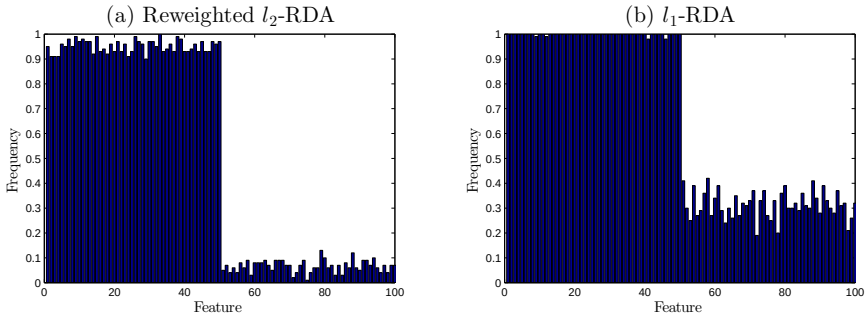

Fig. 2. Frequency of being non-zero for the features of our toy dataset ($d = 100$). Only the first half of features do correspond to the encoded sparsity pattern. In the left subfigure (a) we present the results for the reweighted l_2-RDA approach, while the right subfigure (b) corresponds to l_1-RDA method.

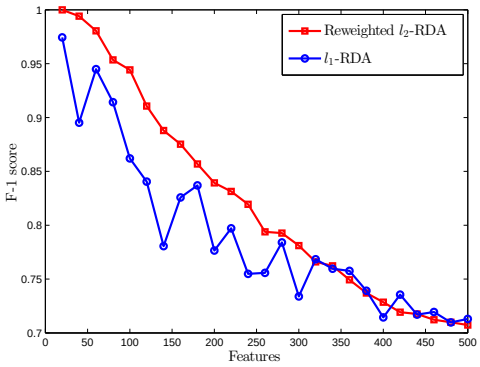

Fig. 3. F1-score as the function of the number of features. We ranged the number of features from 20 to 500 with the step size of 20.

scheme. Our method is suitable both for online and stochastic learning, while our numerical and theoretical results mainly consider only stochastic setting. We provided theoretical guarantees of the boundedness of the regret under different conditions and demonstrated the empirical convergence of the cumulative training error (loss). Experimental results validate the usefulness and promising capabilities of the proposed approach in obtaining much sparser and consistent solutions while keeping the convergence of Pegasos-like approaches at hand.

For the future we consider to improve our algorithm in terms of the accelerated convergence discussed in [4], [12], [18] and develop some further extensions towards online and stochastic learning applied to the huge-scale[4] data.

Acknowledgements: EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-

[4] both in terms of dimensions and number of samples

2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information. Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants Flemish Government: FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants IWT: projects: SBO POM (100031); PhD/Postdoc grants iMinds Medical Information Technologies SBO 2014 Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017)

References

1. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York, NY, USA (2004)
2. Candès, E., Wakin, M., Boyd, S.: Enhancing sparsity by reweighted l1 minimization. Journal of Fourier Analysis and Applications 14(5), 877–905 (2008)
3. Chartrand, R., Yin, W.: Iteratively reweighted algorithms for compressive sensing. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008. pp. 3869–3872 (March 2008)
4. Chen, X., Lin, Q., Peña, J.: Optimal regularized dual averaging methods for stochastic optimization. In: Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) NIPS. pp. 404–412 (2012)
5. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S.: Iteratively reweighted least squares minimization for sparse recovery. Comm. Pure Appl. Math. 63(1), 1–38 (Jan 2010)
6. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
7. Huang, K., King, I., Lyu, M.R.: Direct zero-norm optimization for feature selection. In: ICDM. pp. 845–850 (2008)
8. Lai, M.J., Liu, Y.: The null space property for sparse recovery from multiple measurement vectors. Applied and Computational Harmonic Analysis 30(3), 402–406 (2011)
9. Lai, M.J., Xu, Y., Yin, W.: Improved iteratively reweighted least squares for unconstrained smoothed l_q minimization. SIAM J. Numerical Analysis 51(2), 927–957 (2013)
10. Lázaro, J.L., De Brabanter, K., Dorronsoro, J.R., Suykens, J.A.K.: Sparse LS-SVMs with l_0-norm minimization. In: ESANN. pp. 189–194 (2011)
11. Nelder, J.A., Mead, R.: A simplex method for function minimization. Computer Journal 7, 308–313 (1965)
12. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Mathematical Programming 120(1), 221–259 (2009)
13. Shalev-Shwartz, S., Singer, Y.: Logarithmic regret algorithms for strongly convex repeated games. Tech. rep., The Hebrew University (2007)
14. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In: Proceedings of the 24th international conference on Machine learning. pp. 807–814. ICML '07, New York, NY, USA (2007)
15. Shalev-Shwartz, S., Tewari, A.: Stochastic methods for l1 regularized loss minimization. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 929–936. ICML '09, ACM, New York, NY, USA (2009)
16. Wipf, D.P., Nagarajan, S.S.: Iterative reweighted l_1 and l_2 methods for finding sparse solutions. J. Sel. Topics Signal Processing 4(2), 317–329 (2010)
17. Xavier-De-Souza, S., Suykens, J.A.K., Vandewalle, J., Bollé, D.: Coupled simulated annealing. IEEE Trans. Sys. Man Cyber. Part B 40(2), 320–335 (Apr 2010)
18. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. 11, 2543–2596 (Dec 2010)