Reweighted Stochastic Learning

Vilen Jumutc[a,*], Johan A. K. Suykens[a]

[a] *KU Leuven, ESAT-STADIUS, Kasteelpark Arenberg 10, B-3001, Leuven, Belgium*

Abstract

Recent advances in stochastic learning, such as dual averaging schemes for proximal subgradient-based methods and simple but theoretically well-grounded solvers for linear Support Vector Machines (SVMs), revealed an ongoing interest in making these approaches consistent, robust and tailored towards sparsity inducing norms. In this paper we study reweighted schemes for stochastic learning (specifically in the context of classification problems) based on linear SVMs and dual averaging methods with primal-dual iterate updates. All these methods favour properties of a convex and composite optimization objective. The latter consists of a convex regularization term and loss function with Lipschitz continuous subgradients, *e.g.* l_1-norm ball together with hinge loss. Some approaches approximate in a limit the l_0-type of a 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 convex and strongly convex optimization objectives. These objectives are obtained via the smoothing of a generic sub-differential and possibly non-smooth composite function by the global proximal operator. We report an extended evaluation and comparison of the reweighted schemes against different state-of-the-art techniques and solvers for linear SVMs. Our experimental study indicates the usefulness of the proposed methods for obtaining sparser and better solutions. We show that reweighted schemes can outperform state-of-the-art traditional approaches in terms of generalization error as well.

*Corresponding author

Email addresses: vilen.jumutc@esat.kuleuven.be (Vilen Jumutc), johan.suykens@esat.kuleuven.be (Johan A. K. Suykens)

Keywords: linear SVMs, stochastic learning, l_0 penalty, sparsity, regularization.

1. Introduction

In many domains dealing with online and stochastic learning, the input instances are of very high dimension, yet within any particular instance several features are non-zero. Therefore specific stochastic and online approaches crafted with sparsity inducing regularization are of particular interest for many machine learning researchers and practitioners. This paper investigates an interplay between Regularized Dual Averaging (RDA) approaches [1] (along with other techniques for solving linear SVMs in the context of stochastic learning [2]) and parsimony concepts arising from the application of sparsity inducing norms, like the l_0-type of a penalty.

One can see an increasing importance of correctly identified sparsity patterns and proliferation of proximal and soft-thresholding subgradient-based methods [1], [3], [4]. There are many important contributions of the parsimony concept to the machine learning field. One may allude to the understanding of the obtained solution and simplified or easy to extract decision rules [5, 6, 7]. On the other hand the informativeness of the obtained features might be useful for a better generalization on unseen data [5]. Approaches based on l_1-regularized loss minimization were studied in the context of stochastic and online learning by several research groups [1], [3], [8], [9] but we are not aware of any l_0-norm inducing methods which were applied in the context of Regularized Dual Averaging and stochastic optimization.

In this paper we are trying to provide a supplementary analysis and sufficient regret bounds for learning sparser linear Regularized Dual Averaging (RDA) [1] models from random observations. We extend and modify our previous research [10], [11] and present complementary proofs with fewer assumptions and discussion for the reported theoretical findings. We use sequences of (strongly) convex reweighted optimization objectives to accomplish this goal.

This paper is structured as follows. Section 2 describes previous work on l_0-norm induced learning and some existing solutions to stochastic optimization with regularized loss. Section 3.1 presents a problem statement for the reweighted algorithms. Sections 3.2 and 3.5 introduce our reweighted l_1-RDA and l_2-RDA methods respectively while Section 3.8 presents completely novel approach based on probabilistic reweighted Pegasos-like linear SVM solver. Sections 3.4 and 3.7 provide a theoretic background for our reweighted RDA approaches. Section 4 presents our numerical results and Section 5 concludes the paper.

2. Related work

Learning with $\|w\|_0$ pseudonorm regularization is a NP-hard problem [12] and can be approached via the reweighting schemes [13], [14], [15], [16] while lacking a proper theoretical analysis of convergence in the online and stochastic learning cases. Some methods, like [17], 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.

In many existing iterative reweighting schemes [14], [18] the analysis is provided in terms of the Restricted Isometry Property (RIP) or the Null Space Property (NSP) [19], [14]. These approaches solely rely on the properties which are difficult to access beforehand in a data-driven fashion. This might be crucial if one decides to evaluate methods for their potential applicability. For instance in case of the Restricted Isometry Property, which is characterizing matrix Φ, one is interested to find a constant $\delta \in (0, 1)$ such that for each vector w we would have:

$$(1 - \delta)\|w\|_2 \leq \|\Phi w\|_2 \leq (1 + \delta)\|w\|_2.$$

The RIP was introduced by Candes and Tao [20] in their study of compressed sensing and l_1-minimization. But it cannot be directly applied in the context of online and stochastic optimization because we cannot observe matrix Φ immediately. This fact directly impedes the successful implication of convergence

guarantees based on the RIP or other related properties.

Other groups stemmed their research from the follow-the-regularized-leader (FTRL) family of algorithms [1], [9], [21] and complementary analysis for sparsity-induced learning. In primal-dual subgradient methods arising from this family of algorithms one aims at making a prediction $w_t \in \mathbb{R}^d$ on round t using the average subgradient of the loss function. The update encompasses a trade-off between a gradient-dependent linear term, the regularizer $\psi(w_t)$ and a strongly-convex term h_t for well-conditioned predictions. Our research is based on FTLR algorithms with primal-dual iterate updates, such as RDA [1], and corresponding theoretical guarantees are very much along the lines of the latter.

3. Reweighted Methods

3.1. Problem statement

In the stochastic Regularized Dual Averaging approach developed by Xiao [1] one approximates the loss function $f(w)$ by using a finite set of independent observations $\mathcal{S} = \{\xi_t\}_{1 \leq t \leq T}$. Under this setting one minimizes the following optimization objective:

$$\min_w \frac{1}{T} \sum_{t=1}^{T} f(w, \xi_t) + \psi(w), \tag{1}$$

where $\psi(w)$ represents a regularization term. Every observation is given as a pair of input-output variables $\xi = (x, y)$. In the above setting one deals with a simple classification model $\hat{y}_t = \text{sign}(\langle w, x_t \rangle)$ and calculates the corresponding loss $f(w, \xi_t)$ accordingly[1]. It is common to acknowledge Eq.(1) as an online learning problem if $T \to \infty$.

The problem in Eq.(1) can be approached using a sequence of strongly convex optimization objectives. The solution of every optimization problem at iteration t is treated as a hypothesis of a learner which is induced by an expectation of possibly non-smooth loss function, *i.e.* $\mathbb{E}_\xi[f(w, \xi)]$. One can regularize it by

[1] throughout this paper we will fix $f(w)$ to the hinge loss $f(w, \xi_t) = \max\{0, 1 - y_t \langle w, x_t \rangle\}$

a reweighted norm at each iteration t. This approach in case of satisfying the sufficient conditions will induce a bounded regret *w.r.t.* the loss function which is generating a sequence of stochastic subgradients endowing our dual space E^* [22].

For promoting sparsity we define an iterate-dependent regularization $\psi_t(w) \triangleq \lambda\|\Theta_t w\|$ which in the limit ($t \to \infty$) applies an approximation to the l_0-norm penalty. At every iteration t we will be solving a separate convex instantaneous optimization problem conditioned on a combination of the diagonal reweighting matrices Θ_t. Specific variations of $\psi_t(w)$ for different norms (*e.g.* l_1- and l_2-norm) will be presented in the next subsections. By using a *simple dual averaging* scheme [22] we can solve our problem effectively by the following sequence of iterates w_{t+1} :

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

where $h(w)$ is an auxiliary 1-strongly convex smoothing term (proximal operator defined as $h(w) = \frac{1}{2}\|w - w_0\|$, where w_0 is set to origin), $g_t \in \partial f_t(w_t)$ represents a subgradient and $\{\beta_t\}_{t\geq 1}$ is a non-negative and non-decreasing sequence, which determines the boundedness of the regret function of our algorithms[2].

In detail Eq.(2) is derived using a different optimization objective where we have replaced static regularization term $\psi(w)$ in Eq.(1) with the iterate-dependent term $\psi_t(w)$. In the latter case our optimization objective becomes

$$\min_w \frac{1}{T}\sum_{t=1}^{T}\phi_t(w), \tag{3}$$

where composite function $\phi_t(w)$ is defined as $\phi_t(w) \triangleq f(w, \xi_t) + \psi_t(w)$. Using the aforementioned *dual averaging* scheme from [22] it is easy to show that the sequence w_t in Eq.(2) will approximate an optimal solution to Eq.(3) if we linearly approximate an accumulated loss function $f(w, \xi_t)$ from $\phi_t(w)$ and add a smoothing term $h(w)$. For exact details the interested reader can refer to Eq.(2.14) or in depth to Theorem 1 in [22].

[2]see Sections 3.4 and 3.7

3.2. Reweighted l_1-Regularized Dual Averaging

For promoting additional sparsity to the l_1-Regularized Dual Averaging method [1] we define $\psi_t(w) = \psi_{l_1,t}(w) \triangleq \lambda\|\Theta_t w\|_1$. Hence Eq.(2) becomes:

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

For our reweighted l_1-RDA approach we set $\beta_t = \gamma\sqrt{t}$ and we replace $h(w)$ in Eq.(2) with the parameterized version:

$$h_{l_1}(w) = \frac{1}{2}\|w\|_2^2 + \rho\|w\|_1. \quad (5)$$

Each iterate has a closed form solution. Let us define $\eta_t^{(i)} = \frac{\lambda}{t}\sum_{\tau=1}^{t}\Theta_\tau^{(ii)} + \gamma\rho/\sqrt{t}$ and give an entry-wise solution by:

$$w_{t+1}^{(i)} = \begin{cases} 0, & \text{if } |\hat{g}_t^{(i)}| \le \eta_t^{(i)} \\ -\frac{\sqrt{t}}{\gamma}(\hat{g}_t^{(i)} - \eta_t^{(i)}\text{sign}(\hat{g}_t^{(i)})), & \text{otherwise} \end{cases}, \quad (6)$$

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 $g_t \in \partial f_t(w_t)$. *i.e.* $\hat{g}_t = \frac{1}{t}\sum_{\tau=1}^{t} g_\tau$.

3.3. Reweighted l_1-RDA Algorithm

In this subsection we will outline and explain our main algorithmic scheme for the Reweighted l_1-RDA method. It consists of a simple initialization step, drawing a sample $\mathcal{A}_t \subseteq \mathcal{S}$ from the dataset $\mathcal{S}$, computation and averaging of the subgradient g_t, evaluation of the iterate w_{t+1} and finally re-computation of the reweighting matrix Θ_{t+1}. By analyzing Algorithm 1 we can clearly see that it can operate in a stochastic ($k = 1$, where $k = |\mathcal{A}_t|$) and semi-stochastic mode ($k > 1$) such that one could draw only one or multiple samples from the dataset $\mathcal{S}$. 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 ϵ, which should be tuned or selected properly as described in [16]. This additional hyperparameter ϵ controls the stability of the reweighting scheme and usually is set to a small number to avoid ill-conditioning of the matrix Θ_t.

Algorithm 1: Stochastic Reweighted l_1-Regularized Dual Averaging [11]

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**

  Draw a sample $\mathcal{A}_t \subseteq \mathcal{S}$ of size 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.(6)

  Re-calculate the next Θ 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}

The time complexity of Algorithm 1 is driven by the computational budget T and subsample size k we use therein as our input parameters. In the worst case scenario the time complexity is of order $\mathcal{O}(dkT)$ where d is the input dimension.

3.4. Analysis for the Reweighted l_1-RDA method

In this section we will briefly discuss some of our convergence results and upper bounds for Algorithm 1. We concentrate mainly on the regret *w.r.t.* function $\phi_t(w) \triangleq f(w, \xi_t) + \psi_{l_1,t}(w)$ in Eq.(3) such that for all $w \in \mathbb{R}^n$ and iterates t we have:

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

In Eq.(7) $\mathbf{R}_t(w)$ denotes an accumulated gap between function evaluations at the solution w_τ obtained in a closed form at iterate τ as for instance in Eq.(6) and any $w \in \mathbb{R}^n$. In detail we pay a fixed regret at each iterate τ if we take w_τ instead of an optimal solution w^* for Eq.(3). From [22] and [1] we know that if

we consider $\Delta\psi_{l_1,\tau} = \psi_{l_1,\tau}(w_\tau) - \psi_{l_1,\tau}(w)$ the following *gap sequence* δ_t holds:

$$\delta_t = \max_w \{\sum_{\tau=1}^{t}(\langle g_\tau, w_\tau - w\rangle \ + \Delta\psi_{l_1,\tau})\} \geq \sum_{\tau=1}^{t}(f(w_\tau) - f(w) + \Delta\psi_{l_1,\tau}) = \mathbf{R}_t(w), \tag{8}$$

which due to the convexity of f bounds the regret function from above [23]. Hence by ensuring the necessary condition of Eq.(49) in [1] we can show the upper bound on δ_t, which immediately implies the same bound on $\mathbf{R}_t(w)$.

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* $\psi_{l_1,t}(w_t) \leq \psi_{l_1,t}(w_{t+1})$, $\|g_t\|_* \leq G$, *where* $\|\cdot\|_*$ *stands for the dual norm, and for any* $w \in \mathbb{R}^n$ *holds* $h_{l_1}(w) \leq D$. *Then for any fixed decision* w*:*

$$\mathbf{R}_t(w) \leq (\gamma D + \frac{G^2}{\gamma})\sqrt{t}. \tag{9}$$

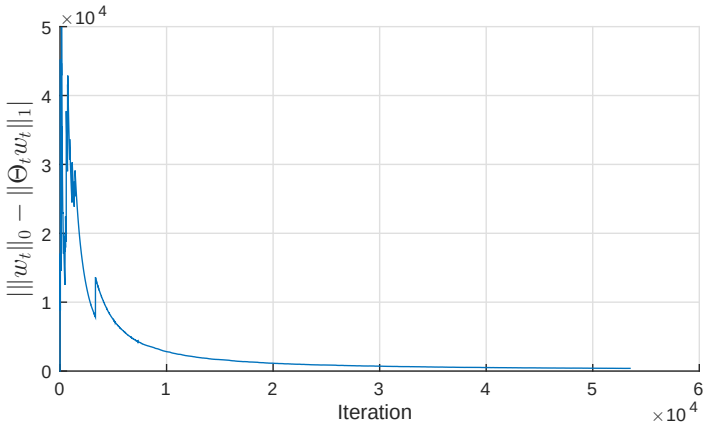

Figure 1: Difference between the reweighted l_1-norm and the true l_0-norm for CT slices dataset [24] at each iterate w_t.

Our intuition is related to the asymptotic convergence properties of an iterative reweighting procedure discussed in [17] where in the limit ($t \to \infty$) iterate Θ_t implies $\|\Theta_t w\|_1 \simeq \|w\|_{p_t}$ with $p_t \to 0$. Hereby we do not give any theoretical justification of the averaging effect implied by Eq.(2) on the approximation. Instead we present an empirical evidence of the convergence in Figure 1. We ran

the Reweighted l_1-RDA algorithm on the CT slices dataset [24] only once using all data points in an online learning setting with the parameters of Algorithm 1 set as follows: $\lambda, \rho, \gamma = 1, \epsilon = 0.1$. The number of iterations corresponds to the total number of available examples in $\mathcal{S}$.

In the next theorem we will slightly relax the necessary condition in order to derive a new bound *w.r.t.* maximal discrepancy of $\psi_{l_1,t}$ function evaluations at subsequent 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* $\psi_{l_1,t}(w_t) - \psi_{l_1,t}(w_{t+1}) \leq \nu/t$ *for some* $\nu \geq 0$, $\|g_t\|_* \leq G$, *where* $\|\cdot\|_*$ *stands for the dual norm, and for any* $w \in \mathbb{R}^n$ *holds* $h_{l_1}(w) \leq D$. *Then for any fixed decision* w*:*

$$\mathbf{R}_t(w) \leq \nu \log t + (\gamma D + \frac{G^2}{\gamma})\sqrt{t}. \tag{10}$$

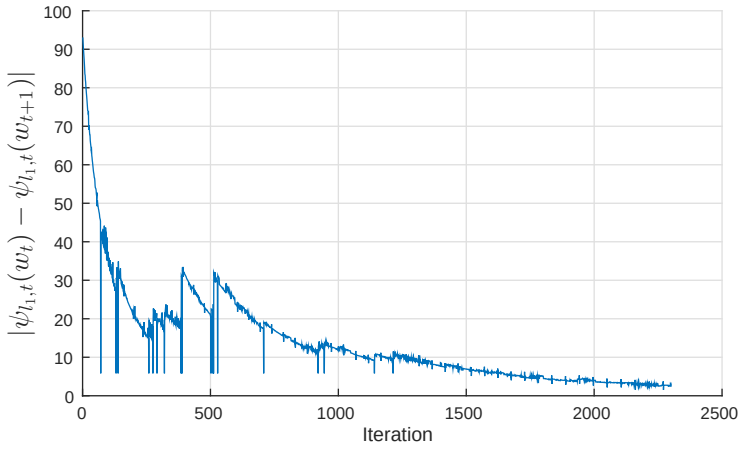

Figure 2: Near asymptotic convergence of the difference $|\psi_{l_1,t}(w_t) - \psi_{l_1,t}(w_{t+1})|$ at the first 2300 iterations for CT slices dataset.

From the above theorem it can be clearly seen that the dependence on the maximal discrepancy term is at most $\mathcal{O}(\log t)$ and is not linked directly to the non-strongly convex instantaneous optimization objective. Hence this discrepancy has less influence on the boundedness of the regret in comparison to the strong convexity assumption.

In Figure 2 we present an empirical evaluation of the near asymptotic convergence of $|\psi_{l_1,t}(w_t) - \psi_{l_1,t}(w_{t+1})|$ sequence, which verifies our assumption on the boundedness of this sequence in Theorem 2. The algorithmic setup is the same as in Figure 1. Proofs and lemmas related to Theorem 1 and Theorem 2 the interested reader can find in the Appendix A.

3.5. Reweighted l_2-Regularized Dual Averaging

Stemmed from the Section 3.2 this particular extension of the RDA approach [1] is dealing with the squared l_2 norm. For promoting additional sparsity we add the reweighted $\|\Theta_t^{1/2}w\|_2^2$ term such that we have $\psi_{l_2,t}(w) \triangleq \lambda\|w\|_2^2 + \|\Theta_t^{1/2}w\|_2^2$. At every iteration t we will be solving a separate λ-strongly convex instantaneous optimization objective conditioned on a combination of the diagonal reweighting matrices 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 $g_\tau \in \partial f(w, \xi_\tau), \tau \in \overline{1,t}$ and regularization terms which are averaged *w.r.t.* to the current iterate t.

Following the *dual averaging scheme* presented by Eq.(2) we can effectively solve our problem with a closed form solution. In our reweighted l_2-RDA approach we use a zero β_t-sequence[3] such that we omit the auxiliary smoothing term $h(w)$ in Eq.(2) which is not necessary since our $\psi_{l_2,t}(w)$ function is already smooth and strongly convex. Hence the solution for every iterate w_{t+1} in our approach is given by

$$w_{t+1} = \arg\min_w \{\frac{1}{t}\sum_{\tau=1}^{t}(\langle g_\tau, w\rangle + \|\Theta_\tau^{1/2}w\|_2^2) + \lambda\|w\|_2^2\}. \tag{11}$$

We will explain the details regarding recalculation of Θ_t and iterate w_{t+1} in the next subsection.

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

3.6. Reweighted l_2-RDA Algorithm

In this subsection we will outline and explain our main algorithmic scheme for the Reweighted l_2-RDA method. It consists of a simple initialization step, computation and averaging of the subgradient g_τ, evaluation of the iterate w_{t+1} and finally recalculation of the reweighting matrix Θ_{t+1}. In Algorithm 2 we do

Algorithm 2: Stochastic Reweighted l_2-Regularized Dual Averaging [10]

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 $t = 1 \to T$ **do**

  Draw a sample $\mathcal{A}_t \subseteq \mathcal{S}$ of size k

  Calculate $g_t \in \partial f(w_t, \mathcal{A}_t)$

  Compute the dual average $\hat{g}_t = \frac{t-1}{t}\hat{g}_{t-1} + \frac{1}{t}g_t$

  Compute the next iterate $w_{t+1}^{(i)} = -\hat{g}_t^{(i)}/(\lambda + \frac{1}{t}\sum_{\tau=1}^{t}\Theta_\tau^{(ii)})$

  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)

not have any explicit sparsification mechanism for the iterate w_{t+1} except for the auxiliary function "`Sparsify`" which utilizes an additional hyperparameter ε and uses it to truncate the final solution w_t below the desired number precision as follows:

$$w_t^{(i)} := \begin{cases} 0, & \text{if } |w_t^{(i)}| \leq \varepsilon, \\ w_t^{(i)}, & \text{otherwise,} \end{cases} \tag{12}$$

where $w_t^{(i)}$ is i-th component of the vector w_t. In comparison with the simple l_2-RDA approach [1] 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 [13] and highlighted in [16]. As in Section 3.3 this

hyperparameter is introduced to stabilize the reweighting scheme and make Θ_t well-conditioned if some entries of w_t are zeros.

The time complexity of Algorithm 2 is the same as for Algorithm 1 with a small extra cost for sparsifying the final solution.

In Algorithms 1 and 2 we perform an optimization *w.r.t.* to the intrinsic bias term b, which doesn't enter our decision function

$$\hat{y} = \text{sign}(\langle w, x \rangle), \tag{13}$$

but is appended to the final solution w. The trick for including a bias term is to augment every input x_t in the subset $\mathcal{A}_t$ with an additional component 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.

3.7. Analysis for the Reweighted l_2-RDA method

In this subsection we will provide the theoretical guarantees for the upper bound on the regret of the function $\phi_t(w) \triangleq f(w, \xi_t) + \psi_{l_2,t}(w)$ as defined in Eq.(7). 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 $\{\Theta_1, \ldots, \Theta_t\}$. 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 terms $\|\Theta_t^{1/2} w\|_2^2 + \lambda \|w\|_2^2$. Supplementary proofs and lemmas are provided in the Appendix B.

Theorem 3. *Let the sequences $\{w_t\}_{t\geq 1}$, $\{g_t\}_{t\geq 1}$ and $\{\Theta_t\}_{t\geq 1}$ be generated by Algorithm 2. Assume $\psi_{l_2,t}(w_t) \leq \psi_{l_2,t}(w_{t+1})$ and $\|g_t\|_* \leq G$, where $\|\cdot\|_*$ stands for the dual norm, and constant $\lambda > 0$ is given for all $\psi_{l_2,t}(w)$. Then for any fixed decision w:*

$$\mathbf{R}_t(w) \leq \frac{G^2}{2\lambda}(1 + \log t). \tag{14}$$

This theorem closely follows results presented in Section 3.2 of [1]. On the other hand our motivation and outline of the proof differs in many aspects. First we

have to maintain a sequence of different regularization terms $\{\psi_{l_2,\tau}(w)\}_{1\leq\tau\leq t}$. Second averaging of this sequence is crucial for proving the boundedness of the conjugate support-type function $V_\tau(s_\tau)$ in [25, 22] for any $\tau \geq 1$.

Theorem 4 provides the necessary condition for deriving a new bound *w.r.t.* maximal discrepancy of $\psi_{l_2,t}$ function evaluations at subsequent w_t iterates.

Theorem 4. *Let the sequences $\{w_t\}_{t\geq 1}$, $\{g_t\}_{t\geq 1}$ and $\{\Theta_t\}_{t\geq 1}$ be generated by Algorithm 2. Assume $\psi_{l_2,t}(w_t) - \psi_{l_2,t}(w_{t+1}) \leq \nu/t$ for some $\nu \geq 0$, $\|g_t\|_* \leq G$, where $\|\cdot\|_*$ stands for the dual norm, and constant $\lambda > 0$ is given for all $\psi_{l_2,t}(w)$. Then for any fixed decision w:*

$$\mathbf{R}_t(w) \leq \frac{G^2}{2\lambda} + \frac{G^2 + 2\lambda\nu}{2\lambda} \log t. \tag{15}$$

The above bound boils down to the bound in Theorem 2 if we set ν to zero. In contrast with Theorem 2 here relaxation implies order $\mathcal{O}(\log t)$ dependency on the total number of iterations and perfectly fits within original bound in Theorem 3 implied by the strong convexity of the instantaneous optimization objective and the zero β_t-sequence.

3.8. Reweighted Dropout Pegasos

In this section we present another novel development based on the Pegasos algorithm [2] for solving linear SVMs. Our finding is motivated by [9], [26] and is rooted in the observation that more dominant features might require more frequent regularization based on the previous values (defined by w_{t-1}).

In our approach we maintain a vector of discrete Bernoulli variables of the same size as w_t, where every success probability $p_t^{(i)}$ of the Bernoulli distribution for feature i at round t is defined as follows:

$$p_t^{(i)} = \frac{w_{t-1}^{(i)} \cdot w_{t-1}^{(i)}}{1 + w_{t-1}^{(i)} \cdot w_{t-1}^{(i)}}. \tag{16}$$

Hence if feature i is updated more frequently and growing by modulus it has higher chances under the Bernoulli distribution for being regularized by the standard l_2-norm penalty [2]. The value of a discrete Bernoulli variable depends upon the weighted previous iterate w_{t-1}. After obtaining a draw of the

particular Bernoulli distribution at round t we simply drop out from regularization features with zero-valued Bernoulli variables. This approach resembles "dropout" regularization applied in convolutional neural networks to prevent them from overfitting [26].

Algorithm 3: Reweighted Dropout Pegasos

Data: $\mathcal{S}, \lambda, T, k, \delta$

1 Select $w_0 = w_1$ randomly s.t. $\|w_1\| \leq 1/\sqrt{\lambda}$
2 **for** $t = 1 \to T$ **do**
3 Draw a sample $\mathcal{A}_t \subseteq \mathcal{S}$ of size k
4 Calculate $g_t \in \partial f(w_t, \mathcal{A}_t)$
5 Calculate $p_t^{(i)}$ by Eq.(16)
6 Draw a binary sample r_t from p_t
7 Set $\eta_t = \frac{1}{\lambda t}$
8 $w_{t+\frac{1}{2}} = w_t - \eta_t(\lambda w_t \circ r_t - \frac{1}{k} g_t)$
9 $w_{t+1} = \min\left\{1, \frac{1/\sqrt{\lambda}}{\|w_{t+\frac{1}{2}}\|}\right\} w_{t+\frac{1}{2}}$
10 **if** $\|w_{t+1} - w_t\| \leq \delta$ **then**
11 **return** w_{t+1}
12 **end**
13 **end**
14 **return** w_{T+1}

Reweighted Dropout Pegasos has a very simple outline and can be summarized in Algorithm 3, where $\circ$ stands for the element-wise multiplication and g_t is a subgradient of an arbitrary convex loss function f_t. By analyzing Algorithm 3 we can see that one major deviation from the Pegasos algorithm is formulated in terms of a binary sample r_t which is drawn from the Bernoulli distribution p_t. This sample is used to drop out the regularization counterpart $\lambda\eta_t w_t$ at the step 8 for some particular dominant features.

4. Simulated experiments

4.1. Experimental setup

For all methods in our experiments with UCI datasets [24] for tuning (*e.g.* to estimate the ubiquitous λ hyperparameter or tuples of hyperparameters employed in Algorithms 1 and 2) we use Coupled Simulated Annealing [27] initialized with 5 random sets of parameters. These random sets are made out of tuples of hyperparameters linked to one particular setup of an algorithm. At every iteration step for CSA we proceed with a 10-fold cross-validation. Within the cross-validation we are promoting additional sparsity with a slightly modified evaluation criterion. We introduce an affine combination of the validation error and the obtained sparsity in proportion of 95% : 5% for initially non-sparse datasets and 80% : 20% for sparse datasets where sparsity is calculated as $\sum_i I(|w^{(i)}| > 0)/d$. This novel cross-validation criterion could be summarized as follows:

$$\text{criterion}_{cv}(X_{valid}, w) = (1-\kappa)\text{error}(X_{valid}, w) + \kappa \sum_i I(|w^{(i)}| > 0)/d, \quad (17)$$

where d is the input dimension, κ is the amount of introduced sparsity (0.05 vs. 0.20) and $\text{error}(X_{valid}, w)$ is implemented as a misclassification rate.

All experiments with large-scale UCI datasets [24] were repeated 50 times (iterations) with the random split to training and test sets in proportion 90% : 10%. At 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 sparsity. 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[4] 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 all presented stochastic algorithms we set $T = 1000$, $k = 1$, $\delta = 10^{-5}$ and other hyperparameters were determined using the cross-validation tuning procedure described

[4] originally it is a regression problem

above. All methods were using a hinge loss as f_t given in Eq.(1) and implemented in `julia` technical computing language[5]. Corresponding software can be found online at `www.esat.kuleuven.be/stadius/ADB/software.php` and `github.com/jumutc/SALSA.jl`.

Table 1: Datasets

Dataset	# attributes	# classes	# data points
Pen Digits	16	10	10992
Opt Digits	64	10	5620
CNAE-9	1080	9	857
Semeion	256	10	1593
Spambase	57	2	4601
Magic	11	2	19020
Shuttle	9	2	58000
CT slices	386	2	53500
Covertype	54	7	581012
URI	3231961	2	16000
RCV1	47152	4	804414

A slightly different approach we took for the RCV1 corpus data [28]. We adopted the experimental setup described in [9]. We ran all competing algorithms only once using cross-validation to search for the optimal trade-off hyperparameter λ. All other hyperparameters were set to the default values in order to obtain at least 50% sparsity. We set the T hyperparameter for all algorithms to the total number of training points, such that we could work within the online learning setting. There are 4 high-level categories: Economics, Commerce, Medical, and Government (ECAT, CCAT, MCAT, GCAT), and multiple more specific categories. We focus on training binary classifiers for each of these

[5] `http://julialang.org/`

major categories. Originally RCV1 data is split into the test and training counterparts. We report our performance for both test and training data. One can find more information on the datasets in Table 1.

Table 2: Generalization performance

Dataset	Generalization (test) errors in % for UCI datasets											
	l_1-RDA$_{re}$ [11]		l_2-RDA$_{re}$ [10]		l_1-RDA$_{ada}$ [9]		l_1-RDA [1]		Pegasos [2]		Pegasos$_{drop}$	
Pen Digits	**6.3**	(±2.1)	7.9	(±2.5)	6.9	(±2.3)	9.2	(±13)	6.3	(±1.9)	**6.1**	(±2.2)
Opt Digits	**3.9**	(±1.7)	4.8	(±2.1)	4.0	(±1.9)	4.4	(±1.9)	**3.4**	(±1.2)	5.3	(±6.4)
Shuttle	**5.3**	(±2.4)	6.9	(±2.7)	5.8	(±2.1)	5.6	(±2.0)	5.3	(±1.7)	**4.7**	(±1.4)
Spambase	11.0	(±3.0)	11.5	(±2.0)	**10.8**	(±1.7)	12.6	(±13)	10.0	(±1.7)	**9.4**	(±1.6)
Magic	22.7	(±2.4)	**22.2**	(±1.3)	22.4	(±1.7)	22.6	(±2.0)	**22.2**	(±1.1)	25.3	(±2.8)
Covertype	28.3	(±1.8)	27.0	(±1.4)	**25.3**	(±1.1)	26.6	(±2.6)	**27.6**	(±1.0)	28.2	(±2.6)
CNAE-9	2.0	(±1.4)	3.6	(±3.7)	**1.9**	(±1.4)	2.3	(±1.8)	1.2	(±1.1)	**0.9**	(±0.9)
Semeion	**8.9**	(±2.6)	13.3	(±18)	10.0	(±3.0)	11.6	(±13)	5.6	(±1.9)	**5.3**	(±1.8)
CT slices	**5.6**	(±1.4)	8.9	(±4.0)	8.4	(±2.8)	8.0	(±1.9)	**5.0**	(±0.7)	5.2	(±1.0)
URI	4.4	(±1.7)	5.2	(±3.0)	**4.0**	(±1.0)	4.8	(±2.5)	**4.3**	(±1.8)	8.4	(±6.0)

4.2. Numerical results

In this subsection we will provide an outlook on the performance of l_1-RDA [1], adaptive l_1-RDA$_{ada}$ [9], our reweighted l_1-RDA$_{re}$ [11] and l_2-RDA$_{re}$ [10] methods as well as our novel Reweighted Dropout Pegasos algorithm (Pegasos$_{drop}$) together with original Pegasos [2] itself. In Table 2 one can see some generalization errors with standard deviations (in brackets) for all datasets.

In Table 2 we have highlighted dominant performances of sparsity inducing RDA-based algorithms and "non-sparse" Pegasos-based algorithms. Analyzing Table 2 we can conclude that for the majority of UCI datasets we are doing equally good *w.r.t.* the Adaptive l_1-RDA method and significantly better *w.r.t.* the original l_1-RDA approach. This fact can be understood from the similar (but

different in theory) underlying principles of the reweighted and adaptive RDA approaches. Indeed both approaches are relying on the importance and hence accumulated information by the represented features. But in the adaptive RDA method we are "reweighting" our closed form solution at round t by the norm over historical subgradients while in the reweighted RDA approaches we are explicitly maintaining diagonal matrix Θ_t which directly preserves the weights and gives as in the limit the approximation for the l_0-norm. In hindsight we can evaluate performance of the competing approaches by taking a closer look at the boxplot of test error distributions in Figure 3. Analyzing performance of

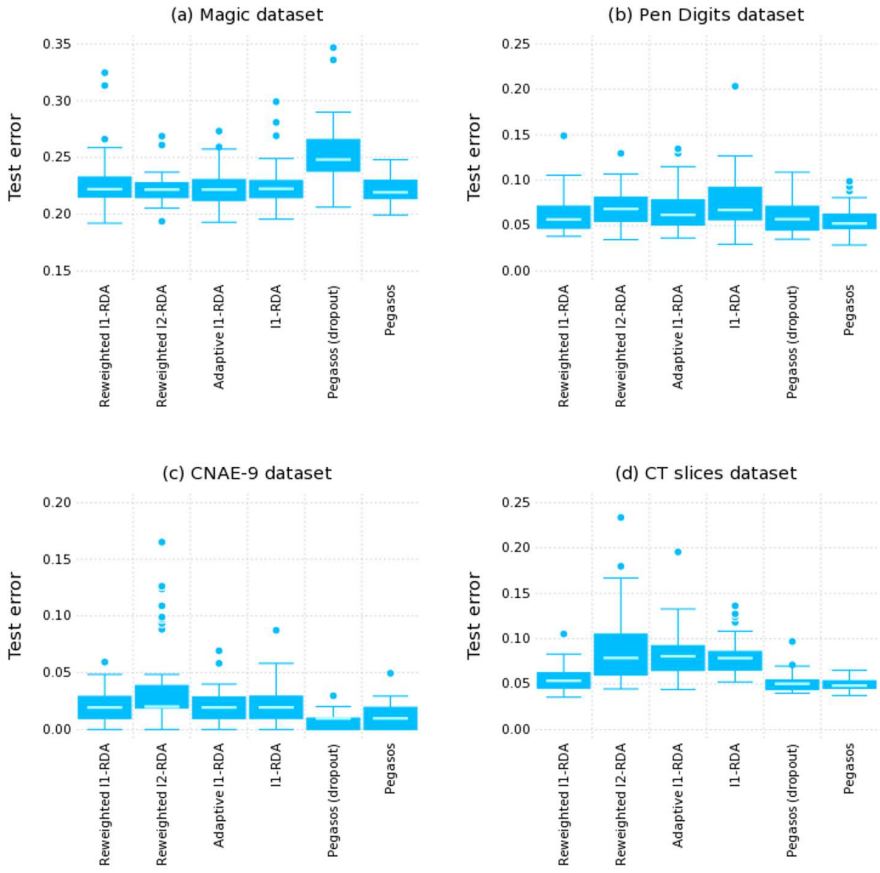

Figure 3: Comparison of the test error distribution over 50 runs for different UCI datasets.

the Pegasos-based approaches we can see that for some datasets our Reweighted Dropout approach outperforms Pegasos.

Table 3: Sparsity $\sum_i I(|w^{(i)}| > 0)/d$

Dataset	Sparsity in % for UCI datasets											
	l_1-RDA$_{re}$ [11]		l_2-RDA$_{re}$ [10]		l_1-RDA$_{ada}$ [9]		l_1-RDA [1]		Pegasos [2]		Pegasos$_{drop}$	
Pen Digits	98.6	(±4.6)	**26.1**	(±19)	48.3	(±18)	40.5	(±18)	100	(±0.0)	100	(±0.0)
Opt Digits	94.0	(±4.1)	33.4	(±18)	36.6	(±12)	**32.5**	(±10)	96.8	(±0.8)	94.8	(±13)
Shuttle	99.8	(±1.5)	**50.0**	(±24)	52.8	(±20)	51.3	(±15)	100	(±0.0)	100	(±0.0)
Spambase	98.2	(±3.8)	**56.9**	(±15)	57.7	(±17)	58.3	(±20)	100	(±0.0)	100	(±0.0)
Magic	93.2	(±12)	**30.8**	(±7.2)	32.8	(±11)	37.2	(±15)	100	(±0.0)	100	(±0.0)
Covertype	92.8	(±14)	**8.0**	(±5.4)	9.4	(±5.0)	12.4	(±7.1)	100	(±0.0)	98.1	(±14)
CNAE-9	**1.42**	(±0.8)	2.86	(±3.7)	1.74	(±1.3)	1.74	(±1.3)	17.9	(±2.1)	14.3	(±1.8)
Semeion	4.82	(±6.7)	6.20	(±20.2)	1.33	(±3.2)	**0.11**	(±0.8)	99.9	(±0.2)	99.8	(±0.2)
CT slices	84.7	(±18.7)	20.9	(±11.3)	**14.0**	(±3.4)	14.4	(±4.2)	98.9	(±0.5)	98.8	(±0.4)
URI	0.06	(±0.06)	0.1	(±0.06)	0.04	(±0.05)	**0.03**	(±.05)	1.4	(±0.07)	0.08	(±0.06)

4.3. Sparsity

In this subsection we will provide some of the findings which should highlight the enhanced sparsity of the reweighted RDA approaches. In Table 3 one can observe the evidence of an additional sparsity promoted by the reweighting procedures which in some cases significantly reduce the number of non-zeros in the obtained solution *w.r.t.* the adaptive and simple l_1-RDA approaches.

By analyzing Table 3 it is not difficult to imply that for almost all datasets we were able to find a good trade-off between sparsity and generalization error. For instance the Reweighted l_2-RDA method was able to find more sparsified solutions[6] with only a small increase in the generalization error for such datasets

[6] in comparison to other RDA-based approaches

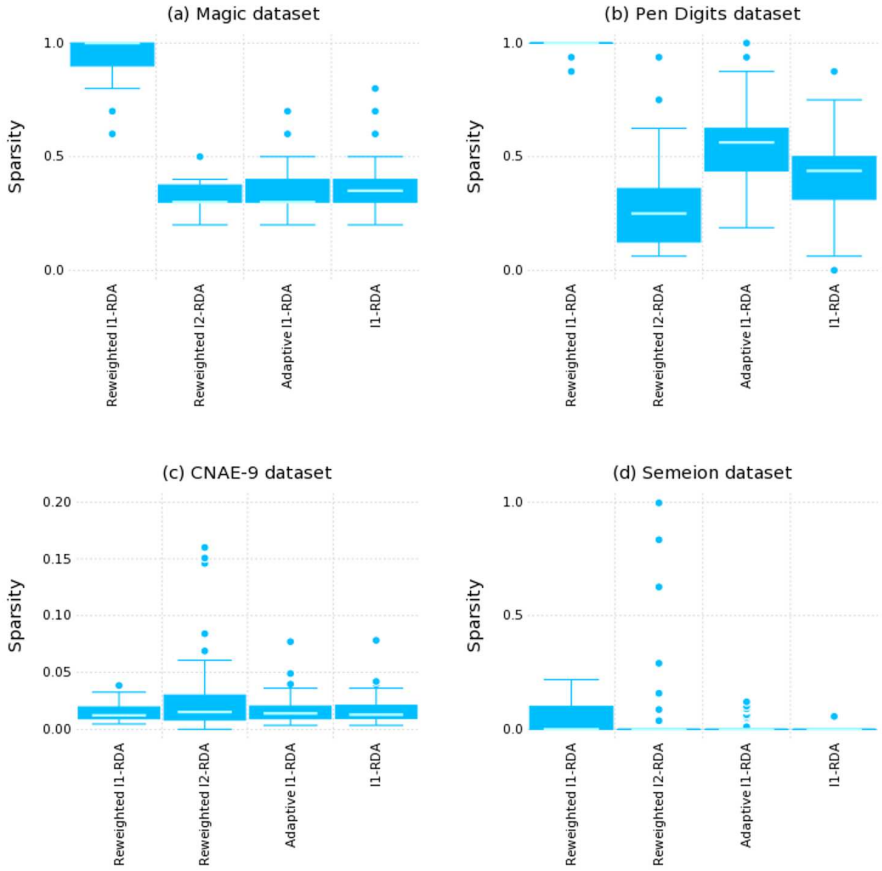

Figure 4: Comparison of the attained sparsity over 50 runs for different UCI datasets.

as Pen Digits, Spambase, Covertype or even better generalization, like for the Magic dataset. On the other hand the Reweighted l_1-RDA method was better in generalization for sparse datasets, like Semeion and CT slices, but less sparsifying than other RDA-based approaches. For one particular dataset (CNAE-9) the Reweighted l_1-RDA method was performing equally good in terms of generalization and sparsity. In hindsight we can evaluate the attained sparsity for different sparsity inducing methods and datasets over 50 trials in Figure 4.

By analyzing these distributions it is easy to verify that for some datasets, like (d) Semeion, most sparsity inducing methods are facing oversparsification

issues, which in turn imply a considerable decay in generalization. For other presented datasets we are able to obtain more consistent performance *w.r.t.* other RDA-based approaches.

Table 4: Performance on RCV1 dataset

Category	Training error (sparsity) (in %)			Test error in %		
	l_1-RDA$_{re}$	l_1-RDA$_{ada}$	l_1-RDA	l_1-RDA$_{re}$	l_1-RDA$_{ada}$	l_1-RDA
CCAT	**5.0** (32.7)	6.3 (47.5)	6.3 (**15.2**)	8.6	**7.9**	9.3
ECAT	**3.5** (24.9)	7.9 (28.5)	4.1 (**11.6**)	**6.5**	9.3	6.7
GCAT	**3.6** (**10.3**)	3.8 (40.0)	5.0 (11.7)	5.6	**5.0**	7.0
MCAT	6.6 (**9.5**)	3.8 (33.0)	**2.7** (10.3)	7.9	5.2	**4.7**

4.4. RCV1 dataset results

We present our results for RCV1 dataset [28] separately because of the different experimental setup and to concentrate on both training and test errors which are of the same interest in the online learning setup. We present both generalization- and sparsity-related performance in Table 4. Each row in the table represents the test and training errors (sparsity is given in brackets) of four different experiments in which we train our binary models *w.r.t.* one of the 4 major high-level categories, *i.e.*: Economics, Commerce, Medical, and Government (ECAT, CCAT, MCAT, GCAT respectively).

Numerical results are given only for l_1-RDA based approaches because we are interested in the comparison of sparsity-inducing methods within the same follow-the-regularized-leader (FTRL) framework, *i.e.* the l_1-norm regularization. This approach gives a clear reference (*w.r.t.* the original l_1-RDA method) how good we can perform using adaptive [9] and reweighted [10] reformulations.

By analyzing Table 4 it is easy to verify that at least two approaches are performing equally good at each category *w.r.t.* the test error and there is no distinct winner in this case. Our results are very different from [9] but our

experimental setup was slightly modified and we were using original TF-IDF features from [28]. In contrast the obtained training error and sparsity gives us more perspicuous outlook that the Reweighted l_1-RDA approach delivers better and faster learning rates in terms of generalization and sparsity while being relatively inferior only for one category, *i.e.* MCAT.

4.5. Stability

To test the stability of our approach (and specifically the Reweighted l_2-RDA method, which is less stable than the Reweighted l_1-RDA rival, due to the enforced and possibly unstable sparsification by `Sparsify`(w_t, ε) procedure in the end of Algorithm 2) we perform several series of experiments with UCI datasets to reveal the consistency and stability of our algorithm *w.r.t.* the obtained 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 [1] method 100 times in order to collect frequencies of every feature (dimension) being non-zero in the obtained solution. In Figure 5 we present the corresponding histograms. As we can see our approach results in much 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.

Additionally we adopt an experimental setup from [8] where we create 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 $y = \text{sign}(\langle w_*, a\rangle + \epsilon)$, 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

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

Figure 6 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 empir-

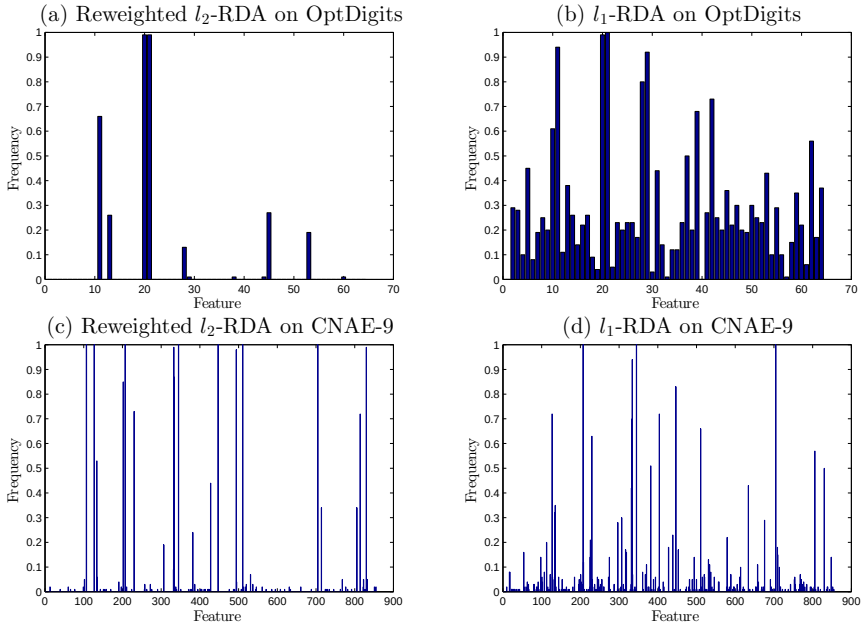

Figure 5: Frequency of being non-zero for the features of Opt Digits and CNAE-9 datasets. In the left subfigures (a,c) we present the results for the reweighted l_2-RDA approach, while the right subfigures (b,d) correspond to l_1-RDA method.

ically 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 7 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 discovered by comparing our F1-scores with [8]. Although the authors in [8] 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 [8] have only 0.87.

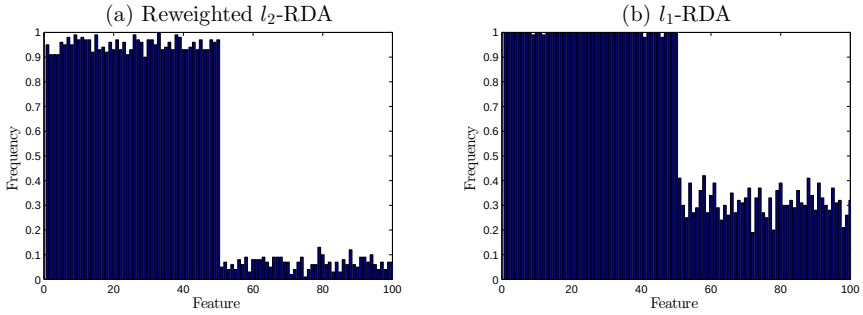

Figure 6: Frequency of being non-zero for the features of our toy dataset ($d = 100$). Only the first half of features do correspond to the encoded sparsity pattern. In the left subfigure (a) we present the results for the reweighted l_2-RDA approach, while the right subfigure (b) corresponds to l_1-RDA method.

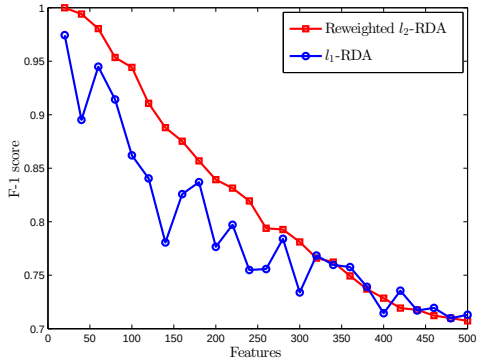

Figure 7: F1-score as the function of the number of features. We ranged the number of features from 20 to 500 with the step size of 20.

5. Conclusion

In this paper we studied reweighted stochastic learning in the context of dual averaging schemes and solvers for linear SVMs. We have presented two different directions for applying reweighting at each round t. The first approach helps to approximate very efficient l_0-type of a penalty using a reliable and proven dual averaging scheme [22]. We applied the reweighting procedure to different norms and elaborated two versions of the Regularized Dual Averaging method [1], namely Reweighted l_1- and l_2-RDA. The second approach stems from the

Pegasos algorithm [2] and applies regularization based on resampling from the Bernoulli distribution where any success probability for any feature i depends upon the weighted value of the previous iterate $w_{t-1}^{(i)}$.

Our methods are suitable both for online and stochastic learning, while our numerical and theoretical results consider only the stochastic setting. We provided theoretical guarantees of the boundedness for the regret under different conditions. Experimental results validate the usefulness and promising capabilities of the proposed approaches in obtaining sparser, consistent and more stable solutions while keeping the regret rates of the follow-the-regularized-leader methods at hand.

For the future we consider to improve our algorithms in terms of the accelerated convergence rate discussed in [8], [22] and develop some further extensions towards online and stochastic coordinate descent methods applied to the huge-scale[7] 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 (FP7/2007-2013) / 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)

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

References

[1] L. Xiao, Dual averaging methods for regularized stochastic learning and online optimization, J. Mach. Learn. Res. 11 (2010) 2543–2596.

[2] S. Shalev-Shwartz, Y. Singer, N. Srebro, Pegasos: Primal Estimated sub-GrAdient SOlver for SVM, in: Proceedings of the 24th international conference on Machine learning, ICML '07, New York, NY, USA, 2007, pp. 807–814.

[3] S. Shalev-Shwartz, A. Tewari, Stochastic methods for l_1 regularized loss minimization, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, ACM, New York, NY, USA, 2009, pp. 929–936.

[4] J. Duchi, Y. Singer, Efficient online and batch learning using forward backward splitting, J. Mach. Learn. Res. 10 (2009) 2899–2934.

[5] N. H. Barakat, A. P. Bradley, Rule extraction from support vector machines: A review., Neurocomputing 74 (1-3) (2010) 178–190.

[6] H. Núnez, C. Angulo, A. Català, Rule extraction from support vector machines, in: In Proceedings of European Symposium on Artificial Neural Networks, 2002, pp. 107–112.

[7] C. J. C. Burges, Simplified support vector decision rules, in: L. Saitta (Ed.), Proceedings of the 13th International Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA, 1996, pp. 71–77.

[8] X. Chen, Q. Lin, J. Peña, Optimal regularized dual averaging methods for stochastic optimization., in: P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), NIPS, 2012, pp. 404–412.

[9] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011) 2121–2159.

[10] V. Jumutc, J. A. K. Suykens, Reweighted l_2-regularized dual averaging approach for highly sparse stochastic learning, in: Advances in Neural Networks - 11th International Symposium on Neural Networks, ISNN 2014, Hong Kong and Macao, China, November 28 – December 1, 2014, pp. 232–242.

[11] V. Jumutc, J. A. K. Suykens, Reweighted l_1 dual averaging approach for sparse stochastic learning, in: 22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25, 2014.

[12] J. L. Lázaro, K. De Brabanter, J. R. Dorronsoro, J. A. K. Suykens, Sparse LS-SVMs with l_0-norm minimization, in: ESANN, 2011, pp. 189–194.

[13] R. Chartrand, W. Yin, Iteratively reweighted algorithms for compressive sensing, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008, 2008, pp. 3869–3872.

[14] I. Daubechies, R. DeVore, M. Fornasier, C. S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, Comm. Pure Appl. Math. 63 (1) (2010) 1–38. `arXiv:0807.0575`.

[15] D. P. Wipf, S. S. Nagarajan, Iterative reweighted l_1 and l_2 methods for finding sparse solutions, J. Sel. Topics Signal Processing 4 (2) (2010) 317–329.

[16] E. Candès, M. Wakin, S. Boyd, Enhancing sparsity by reweighted l_1 minimization, Journal of Fourier Analysis and Applications 14(5) (2008) 877–905.

[17] K. Huang, I. King, M. R. Lyu, Direct zero-norm optimization for feature selection., in: ICDM, 2008, pp. 845–850.

[18] M.-J. Lai, Y. Xu, W. Yin, Improved iteratively reweighted least squares for unconstrained smoothed l_q minimization, SIAM J. Numerical Analysis 51 (2) (2013) 927–957.

[19] M.-J. Lai, Y. Liu, The null space property for sparse recovery from multiple measurement vectors, Applied and Computational Harmonic Analysis 30 (3) (2011) 402–406.

[20] E. Candes, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, Information Theory, IEEE Transactions on 52 (12) (2006) 5406–5425.

[21] E. Hazan, A. Agarwal, S. Kale, Logarithmic regret algorithms for online convex optimization, Machine Learning 69 (2-3) (2007) 169–192.

[22] Y. Nesterov, Primal-dual subgradient methods for convex problems, Mathematical Programming 120 (1) (2009) 221–259.

[23] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, New York, NY, USA, 2004.

[24] A. Frank, A. Asuncion, UCI machine learning repository (2010).
URL `http://archive.ics.uci.edu/ml`

[25] Y. Nesterov, Smooth minimization of non-smooth functions, Mathematical Programming 103 (1) (2005) 127–152.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.

[27] S. Xavier-De-Souza, J. A. K. Suykens, J. Vandewalle, D. Bollé, Coupled simulated annealing, IEEE Trans. Sys. Man Cyber. Part B 40 (2) (2010) 320–335.

[28] D. D. Lewis, Y. Yang, T. G. Rose, F. Li, Rcv1: A new benchmark collection for text categorization research, J. Mach. Learn. Res. 5 (2004) 361–397.

Appendix A

Proof of Theorem 1

We start with redefining the conjugate-type functions $V_t(s)$ and $U_t(s)$ in [1] by replacing $\psi(w)$ in each of them with the sum of our reweighted l_1 functions $\psi_{l_1,t}(w) \triangleq \lambda\|\Theta_t w\|_1$, such that:

$$U_t(s) = \max_{w\in\mathcal{F}_\mathcal{D}} \{\langle s, w - w_0\rangle - \sum_{\tau=1}^{t} \psi_{l_1,\tau}(w)\}, \tag{18}$$

$$V_t(s) = \max_{w} \{\langle s, w - w_0\rangle - \sum_{\tau=1}^{t} \psi_{l_1,\tau}(w) - \beta_t h_{l_1}(w)\}, \tag{19}$$

where we define the domain of any $\psi_{l_1,\tau}(w)$ function to be identical and set as $\mathcal{F}_D = \{w \in \text{dom}\psi_{l_1} | h_{l_1}(w) \leq D\}$ (*i.e.* the domain of a simple non-reweighted l_1-norm as in [1]), $\{\beta_t\}_{t\geq 1}$ is a non-negative and non-decreasing sequence and $h_{l_1}(w)$ is a 1-strongly convex function defined in Eq.(5).

Next we refer to the important Lemma 9 in [1], as for any $s \in E^*$ and $t \geq 0$ we have: $U_t(s) \leq V_t(s) + \beta_t D$, because using the definition of $\mathcal{F}_D$ we can bound $U_t(s)$ as follows:

$$\begin{aligned}
U_t(s) &= \max_{w\in\mathcal{F}_\mathcal{D}} \{\langle s, w - w_0\rangle - \sum_{\tau=1}^{t} \psi_{l_1,\tau}(w)\} \\
&= \max_{w} \min_{\beta\geq 0} \{\langle s, w - w_0\rangle - \sum_{\tau=1}^{t} \psi_{l_1,\tau}(w) + \beta(D - h_{l_1}(w))\} \\
&\leq \min_{\beta\geq 0} \max_{w} \{\langle s, w - w_0\rangle - \sum_{\tau=1}^{t} \psi_{l_1,\tau}(w) + \beta(D - h_{l_1}(w))\} \\
&\leq \max_{w} \{\langle s, w - w_0\rangle - \sum_{\tau=1}^{t} \psi_{l_1,\tau}(w) + \beta_t(D - h_{l_1}(w))\} \\
&= V_t(s) + \beta_t D.
\end{aligned} \tag{20}$$

□

We will need this lemma to bound the regret as it will be shown below. Before bounding the regret we need to adjust one more lemma from [1]:

Lemma 1. *For each $t \geq 1$ we have $V_t(-s_t) + \psi_{l_1,t}(w_t) \leq V_{t-1}(-s_t)$.*

Proof 1.

$$\begin{aligned} V_{t-1}(-s_t) &= \max_w \{\langle s_t, w - w_0\rangle - \sum_{\tau=1}^{t-1} \psi_{l_1,\tau}(w) - \beta_{t-1} h_{l_1}(w)\} \\ &\geq \langle s_t, w_{t+1} - w_0\rangle - \sum_{\tau=1}^{t-1} \psi_{l_1,\tau}(w_{t+1}) - \beta_{t-1} h_{l_1}(w_{t+1}) \\ &\geq \{\langle s_t, w_{t+1} - w_0\rangle - \sum_{\tau=1}^{t} \psi_{l_1,\tau}(w_{t+1}) - \beta_t h_{l_1}(w_{t+1})\} + \psi_{l_1,t}(w_{t+1}) \\ &= V_t(-s_t) + \psi_{l_1,t}(w_{t+1}) \geq V_t(-s_t) + \psi_{l_1,t}(w_t). \end{aligned}$$

□

In this proof the first line follows from the definition of $V_t(s)$. The second line is immediately implied by the maximization objective. On the third line we used the main property of $\{\beta_t\}_{t\geq 1}$ sequence to be non-decreasing and non-negative. We derived an equality on the final line by noticing that the expression in curly brackets is exactly $V_t(-s_t)$ (*i.e.* the solution of the maximization problem defined in $V_{t-1}(-s_t)$ is exactly w_{t+1} by our algorithmic design in Eq.(2). And the final inequality follows from our assumption $\psi_{l_1,t}(w_t) \leq \psi_{l_1,t}(w_{t+1})$.

In the next part we will finally bound our regret function defined in Eq.(7). From [22] and [1] we know that if we consider $\Delta\psi_{l_1,\tau} = \psi_{l_1,\tau}(w_\tau) - \psi_{l_1,\tau}(w)$ the following *gap sequence* δ_t:

$$\delta_t = \max_w \{\sum_{\tau=1}^{t} (\langle g_\tau, w_\tau - w\rangle + \Delta\psi_{l_1,\tau})\} \geq \sum_{\tau=1}^{t} (f(w_\tau, \xi_\tau) - f(w, \xi_\tau) + \Delta\psi_{l_1,\tau}) = \mathbf{R}_t(w) \tag{21}$$

bounds the regret function from above due to the convexity of f [23]. Next we can derive an upper bound on δ_t using the properties of the conjugate-type functions $V_t(s)$ and $U_t(s)$.

If we add and subtract the $\sum_{\tau=1}^{t} \langle g_\tau, w_0\rangle$ term from the definition in Eq.(21)[8]

[8] for derivations below we have to substitute general index t with the running index $\tau \in \overline{1,t}$

then for any $w \in \mathbb{R}^n$ we get:

$$\begin{aligned}\delta_t &= \sum_{\tau=1}^{t}(\langle g_\tau, w_\tau - w_0\rangle \ + \psi_{l_1,\tau}(w_\tau)) + \max_w\{\langle s_t, w_0 - w\rangle - \sum_{\tau=1}^{t}\psi_{l_1,\tau}(w)\} \\ &\leq \sum_{\tau=1}^{t}(\langle g_\tau, w_\tau - w_0\rangle \ + \psi_{l_1,\tau}(w_\tau)) + V_t(-s_t) + \beta_t D, \end{aligned} \tag{22}$$

where $s_t = \sum_{\tau=1}^{t} g_\tau$. The last maximization term in Eq.(22) is exactly $U_t(-s_t)$ so it can be bounded by Eq.(20).

Applying well-known results on the boundedness of the conjugate-type functions from [25, 22] for any $\tau \geq 1$ and in view of Lemma 1 we have:

$$\begin{aligned} V_\tau(-s_\tau) + \psi_{l_1,\tau}(w_\tau) &\leq V_{\tau-1}(-s_\tau) \\ &= V_{\tau-1}(-s_{\tau-1} - g_\tau) \\ &\leq V_{\tau-1}(-s_{\tau-1}) + \langle -g_\tau, \nabla V_{\tau-1}(-s_{\tau-1})\rangle + \frac{\|g_\tau\|_*^2}{2\beta_{\tau-1}} \\ &= V_{\tau-1}(-s_{\tau-1}) + \langle -g_\tau, \hat{w}_\tau - w_0\rangle + \frac{\|g_\tau\|_*^2}{2\beta_{\tau-1}}, \end{aligned} \tag{23}$$

where $\|\cdot\|_*$ is a dual norm, second inequality is due to the result of [25]: $V_\beta(s+g) \leq V_\beta(s) + \langle g, \nabla V_\beta(s)\rangle + \|g\|_*/2\sigma\beta,\ \forall s, g \in E^*$, the fact that $\sigma = 1$ for $h_{l_1}(w)$ and

$$\nabla V_{\tau-1}(-s_{\tau-1}) = \hat{w}_\tau - w_0,\ \forall \tau \geq 1$$

where $\hat{w}_\tau \triangleq \arg\min_w\{\langle s_{\tau-1}, w\rangle + \sum_{\kappa=1}^{\tau-1}\psi_{l_1,\kappa}(w) + \beta_{\tau-1}h_{l_1}(w)\},\ \forall \tau \geq 1$.

Next we rearrange and sum up everything *w.r.t.* the index τ:

$$\sum_{\tau=1}^{t}(\langle g_\tau, \hat{w}_\tau - w_0\rangle + \psi_{l_1,\tau}(w_\tau)) + V_t(-s_t) \leq V_0(-s_0) + \sum_{\tau=1}^{t}\frac{\|g_\tau\|_*^2}{2\beta_{\tau-1}}. \tag{24}$$

We can further simplify this equation by assuming $V_0(-s_0) = V_0(0) = 0$ and from the first-order optimality conditions of the function f as follows:

$$0 \leq \langle g_\tau, \hat{w}_\tau - w_\tau\rangle \Longrightarrow \langle g_\tau, w_\tau\rangle \leq \langle g_\tau, \hat{w}_\tau\rangle.$$

Substituting all of above into Eq.(24) and taking into account Eq.(22) we get:

$$\mathbf{R}_t(w) \leq \delta_t \leq \beta_t D + \sum_{\tau=1}^{t}\frac{\|g_\tau\|_*^2}{2\beta_{\tau-1}}. \tag{25}$$

To obtain bound in Theorem 1 we assume $\|g_t\|_* \leq G$ and set $\beta_t = \gamma\sqrt{t}$, while keeping $\beta_0 = \beta_1 = \gamma$:

$$\mathbf{R}_t(w) \leq \gamma\sqrt{t}D + \frac{G^2}{2\gamma}(1 + \sum_{\tau=1}^{t-1}\frac{1}{\sqrt{\tau}}) \leq (\gamma D + \frac{G^2}{\gamma})\sqrt{t}, \tag{26}$$

where $\sum_{\tau=1}^{t-1}\tau^{-\frac{1}{2}} \leq 1 + \int_1^t \tau^{-\frac{1}{2}}d\tau = 2\sqrt{t} - 1$.

□

Proof of Theorem 2

The main part of the proof is exactly the same as for Theorem 1. We will refer to the parts of the proof of Theorem 1 as we will need them.

First we need to adjust Lemma 1 and introduce the dependency on the sufficient conditions of Theorem 2.

Lemma 2. *For each $t \geq 1$ we have $V_t(-s_t) + \psi_{l_1,t}(w_t) \leq V_{t-1}(-s_t) + \nu/t$.*

Proof 2.

$$\begin{aligned}
V_{t-1}(-s_t) &= \max_w\{\langle s_t, w - w_0\rangle - \sum_{\tau=1}^{t-1}\psi_{l_1,\tau}(w) - \beta_{t-1}h_{l_1}(w)\} \\
&\geq \langle s_t, w_{t+1} - w_0\rangle - \sum_{\tau=1}^{t-1}\psi_{l_1,\tau}(w_{t+1}) - \beta_{t-1}h_{l_1}(w_{t+1}) \\
&\geq \{\langle s_t, w_{t+1} - w_0\rangle - \sum_{\tau=1}^{t}\psi_{l_1,\tau}(w_{t+1}) - \beta_t h_{l_1}(w_{t+1})\} + \psi_{l_1,t}(w_{t+1}) \\
&= V_t(-s_t) + \psi_{l_1,t}(w_{t+1}).
\end{aligned}$$

And to get the final result we use our assumption in Theorem 2 that we have $\psi_{l_1,t}(w_t) - \psi_{l_1,t}(w_{t+1}) \leq \nu/t$ for any $t \geq 1$, $\nu \geq 0$, hence:

$$V_{t-1}(-s_t) \geq V_t(-s_t) + \psi_{l_1,t}(w_{t+1}) \Longrightarrow V_{t-1}(-s_t) + \nu/t \geq V_t(-s_t) + \psi_{l_1,t}(w_t).$$

□

Finally to derive our bound in Theorem 2 we note that Eq.(23) in view of Lemma 2 is given as follows:

$$V_\tau(-s_\tau) + \psi_{l_1,\tau}(w_\tau) \leq V_{\tau-1}(-s_{\tau-1}) + \nu/\tau + \langle -g_\tau, \hat{w}_\tau - w_0\rangle + \frac{\|g_\tau\|_*^2}{2\beta_{\tau-1}}. \tag{27}$$

Next we rearrange and sum up everything *w.r.t.* the index τ:

$$\sum_{\tau=1}^{t}(\langle g_\tau, \hat{w}_\tau - w_0\rangle + \psi_{l_1,\tau}(w_\tau)) + V_t(-s_t) \leq V_0(-s_0) + \nu\log t + \sum_{\tau=1}^{t}\frac{\|g_\tau\|_*^2}{2\beta_{\tau-1}}, \quad (28)$$

where we used $\sum_{\tau=1}^{t}\frac{1}{\tau} \leq \int_1^t \tau^{-1}d\tau = \log t$ to sum up ν/τ term. Next keeping in mind our initial bound on $U_t(s) \leq V_t(s) + \beta_t D$ and taking into account Eq.(22), $V_0(-s_0) = V_0(0) = 0$ and the first-order optimality conditions of the function f we get:

$$\mathbf{R}_t(w) \leq \delta_t \leq \nu(1 + \log t) + \beta_t D + \sum_{\tau=1}^{t}\frac{\|g_\tau\|_*^2}{2\beta_{\tau-1}}, \quad (29)$$

where one part of the right-hand side of inequality is exactly the same as in Eq.(25). So we can immediately substitute it with the right-hand side of Eq.(26) and get:

$$\mathbf{R}_t(w) \leq \nu\log + (\gamma D + \frac{G^2}{\gamma})\sqrt{t}. \quad (30)$$

□

Appendix B

Proof of Theorem 3

We stem our result from the proof of the Theorem 1 by redefining the conjugate-type functions $V_t(s)$ and $U_t(s)$ in [1] by replacing $\psi(w)$ in each of them with the sum of our reweighted l_2 functions $\psi_{l_2,t}(w) \triangleq \|\Theta_t^{1/2}w\|_2^2 + \lambda\|w\|_2^2$, such that:

$$U_t(s) = \max_{w\in\mathcal{F}_\mathcal{D}}\{\langle s, w - w_0\rangle - \sum_{\tau=1}^{t}\psi_{l_2,\tau}(w)\},$$
$$V_t(s) = \max_{w}\{\langle s, w - w_0\rangle - \sum_{\tau=1}^{t}\psi_{l_2,\tau}(w) - \beta_t h(w)\}, \quad (31)$$

where we define the domain of $\psi_{l_2,\tau}(w)$ as $\mathcal{F}_D = \{w \in \mathrm{dom}\psi_{l_2} | h(w) \leq D\}$ (*i.e.* the domain of a simple non-reweighted l_2-norm as in [1]), $\{\beta_t\}_{t\geq 1}$ is a non-negative and non-decreasing sequence and $h(w)$ is a 1-strongly convex function (*i.e.* smoothing term).

Next we refer to the important Lemma 9 in [1], as for any $s \in E^*$ and $t \geq 0$ we have: $U_t(s) \leq V_t(s) + \beta_t D$, because using the definition of $\mathcal{F}_D$ we can bound $U_t(s)$ as follows:

$$\begin{aligned}
U_t(s) &= \max_{w \in \mathcal{F}_{\mathcal{D}}} \{\langle s, w - w_0 \rangle - t\psi_{l_2,1}(w)\} \\
&= \max_{w} \min_{\beta \geq 0} \{\langle s, w - w_0 \rangle - t\psi_{l_2,1}(w) + \beta(D - h(w))\} \\
&\leq \min_{\beta \geq 0} \max_{w} \{\langle s, w - w_0 \rangle - t\psi_{l_2,1}(w) + \beta(D - h(w))\} \\
&\leq \max_{w} \{\langle s, w - w_0 \rangle - t\psi_{l_2,1}(w) + \beta_t(D - h(w))\} \\
&= V_t(s) + \beta_t D.
\end{aligned}$$

We will need this lemma to bound the regret as it will be shown below.

Lemma 3. *For each $t \geq 1$ in Algorithm 2 we have $V_t(-s_t) + \psi_{l_2,t}(w_t) \leq V_{t-1}(-s_t)$.*

Proof 3.

$$\begin{aligned}
V_{t-1}(-s_t) &= \max_{w} \{\langle s_t, w - w_0 \rangle - \sum_{\tau=1}^{t-1} \psi_{l_2,\tau}(w) - \beta_{t-1} h(w)\} \\
&\geq \langle s_t, w_{t+1} - w_0 \rangle - \sum_{\tau=1}^{t-1} \psi_{l_2,\tau}(w_{t+1}) - \beta_{t-1} h(w_{t+1}) \\
&\geq \{\langle s_t, w_{t+1} - w_0 \rangle - \sum_{\tau=1}^{t} \psi_{l_2,\tau}(w_{t+1}) - \beta_t h(w_{t+1})\} + \psi_{l_2,t}(w_{t+1}) \\
&= V_t(-s_t) + \psi_{l_2,t}(w_{t+1}) \geq V_t(-s_t) + \psi_{l_2,t}(w_t).
\end{aligned}$$

In the above proof we used the same evidence as for Lemma 1 but with respect to $\psi_{l_2,1}(w)$. The final inequality follows from our assumption $\psi_{l_2,t}(w_t) \leq \psi_{l_2,t}(w_{t+1})$.

In the next part we will finally bound our regret function defined in Eq.(7) but with respect to $\psi_{l_2,t}(w)$. From [22] and [1] we know that if we consider

$\Delta\psi_{l_2,\tau} = \psi_{l_2,\tau}(w_\tau) - \psi_{l_2,\tau}(w)$ the following *gap sequence* δ_t:

$$\begin{aligned}\delta_t &= \max_w\{\sum_{\tau=1}^{t}(\langle g_\tau, w_\tau - w\rangle \ + \Delta\psi_{l_2,\tau})\} \\ &\geq \sum_{\tau=1}^{t}(f(w_\tau,\xi_\tau) - f(w,\xi_\tau) + \Delta\psi_{l_2,\tau}) = \mathbf{R}_t(w) \end{aligned} \tag{32}$$

bounds the regret function from above due to the convexity of f [23]. Next we can derive an upper bound on δ_t using the properties of the conjugate-type functions $V_t(s)$ and $U_t(s)$.

If we add and subtract $\sum_{\tau=1}^{t}\langle g_\tau, w_0\rangle$ term from the definition in Eq.(32) then for any $w \in \mathbb{R}^n$ we get:

$$\begin{aligned}\delta_t &= \sum_{\tau=1}^{t}(\langle g_\tau, w_\tau - w_0\rangle \ + \psi_{l_2,\tau}(w_\tau)) + \max_w\{\langle s_t, w_0 - w\rangle - \sum_{\tau=1}^{t}\psi_{l_2,\tau}(w)\} \\ &\leq \sum_{\tau=1}^{t}(\langle g_\tau, w_\tau - w_0\rangle \ + \psi_{l_2,\tau}(w_\tau)) + V_t(-s_t) + \beta_t D, \end{aligned} \tag{33}$$

where $s_t = \sum_{\tau=1}^{t} g_\tau$. The last maximization term on the first line in Eq.(33) is exactly $U_t(-s_t)$, which was bounded earlier.

Using the well-known results on the boundedness of the conjugate support-type functions from [25, 22] for any $\tau \geq 1$ we have:

$$\begin{aligned} V_\tau(-s_\tau) + \psi_{l_2,\tau}(w_\tau) &\leq V_{\tau-1}(-s_\tau) \\ &= V_{\tau-1}(-s_{\tau-1} - g_\tau) \\ &\leq V_{\tau-1}(-s_{\tau-1}) + \langle -g_\tau, \nabla V_{\tau-1}(-s_{\tau-1})\rangle \\ &+ \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))} \\ &= V_{\tau-1}(-s_{\tau-1}) + \langle -g_\tau, \hat{w}_\tau - w_0\rangle \\ &+ \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))}, \end{aligned} \tag{34}$$

where $\|\cdot\|_*$ is a dual norm, the second inequality is due to the result of [25]: $V_\beta(s+g) \leq V_\beta(s) + \langle g, \nabla V_\beta(s)\rangle + \|g\|_*/(2\sigma\beta),\ \forall s, g \in E^*$, where $\sigma\beta$ term refers to the smoothing part, which is quite different in case of $V_t(s)$ in Eq.(31). Based on the result in [25] and Lemma 1 in [22], the fact that $\sigma = 1$ in $h(w)$ and our

particular choice of the smoothing strongly-convex term $\sum_{\tau=1}^{t}\psi_{l_2,\tau}(w)+\beta_t h(w)$ in Eq.(31) we have the Lipschitz continuous gradient of $V_t(s)$ with constant $\frac{1}{\beta_t+\lambda t}$, such that:

$$\|\nabla V_t(s_1)-\nabla V_t(s_2)\| \leq \frac{1}{\beta_t+\lambda t}\|s_1-s_2\|_*, \quad \forall s_1,s_2 \in E^* \tag{35}$$

On the other hand we know a closed form for such a gradient in case of Eq.(34) and in view of conjugate support-type function $V_t(s)$ in Eq.(31):

$$\nabla V_{\tau-1}(-s_{\tau-1}) = \hat{w}_\tau - w_0,\ \forall \tau \geq 1$$

where $\hat{w}_\tau \triangleq \arg\min_w\{\langle s_{\tau-1}, w\rangle + \sum_{\kappa=1}^{\tau-1}\psi_{l_2,\kappa}(w)+\beta_{\tau-1}h(w)\}$ for $\forall \tau \geq 1$.

Next we rearrange and sum up everything *w.r.t.* index τ:

$$\begin{aligned}\sum_{\tau=1}^{t}(\langle g_\tau, \hat{w}_\tau - w_0\rangle + \psi_{l_2,\tau}(w_\tau)) + V_t(-s_t) \\ \leq V_0(-s_0) + \sum_{\tau=1}^{t}\frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1}+\lambda(\tau-1))}.\end{aligned} \tag{36}$$

We can further simplify this equation by assuming $V_0(-s_0)=V_0(0)=0$ and by using the first-order optimality conditions of the function f as follows:

$$0 \leq \langle g_\tau, \hat{w}_\tau - w_\tau\rangle \Longrightarrow \langle g_\tau, w_\tau\rangle \leq \langle g_\tau, \hat{w}_\tau\rangle.$$

Substituting all of above into Eq.(36) and taking into account Eq.(33) we get:

$$\mathbf{R}_t(w) \leq \delta_t \leq \beta_t D + \sum_{\tau=1}^{t}\frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1}+\lambda(\tau-1))}. \tag{37}$$

To obtain the bound in Theorem 3 we assume $\|g_t\|_* \leq G$ and set $\beta_{t\geq 1}=0$, while keeping $\beta_0=\lambda$:

$$\mathbf{R}_t(w) \leq \frac{\|g_1\|_*^2}{2\lambda} + \sum_{\tau=1}^{t-1}\frac{\|g_{\tau+1}\|_*^2}{2\lambda\tau} \leq \frac{G^2}{2\lambda}(1+\int_1^t \tau^{-1}d\tau) = \frac{G^2}{2\lambda}(1+\log t). \tag{38}$$

□

Proof of Theorem 4

The main part of the proof is exactly the same as for Theorem 3. We will refer to the parts of the proof of Theorem 3 as we will need them.

First we need to adjust Lemma 3 and introduce the dependency on the sufficient conditions of Theorem 3.

Lemma 4. *For each $t \geq 1$ in Algorithm 2 assuming that $\psi_{l_2,t}(w_t) \leq \psi_{l_2,t}(w_{t+1}) + \nu/t$ for some $\nu \geq 0$ we have $V_t(-s_t) + \psi_{l_2,t}(w_t) \leq V_{t-1}(-s_t) + \nu/t$.*

Proof 4.

$$\begin{aligned} V_{t-1}(-s_t) &= \max_w \{\langle s_t, w - w_0\rangle - \sum_{\tau=1}^{t-1} \psi_{l_2,\tau}(w) - \beta_{t-1} h(w)\} \\ &\geq \langle s_t, w_{t+1} - w_0\rangle - \sum_{\tau=1}^{t-1} \psi_{l_2,\tau}(w_{t+1}) - \beta_{t-1} h(w_{t+1}) \\ &\geq \{\langle s_t, w_{t+1} - w_0\rangle - \sum_{\tau=1}^{t} \psi_{l_2,\tau}(w_{t+1}) - \beta_t h(w_{t+1})\} + \psi_{l_2,t}(w_{t+1}) \\ &= V_t(-s_t) + \psi_{l_2,t}(w_{t+1}), \end{aligned}$$

and to get the final result we use our assumption:

$$V_{t-1}(-s_t) \geq V_t(-s_t) + \psi_{l_2,t}(w_1) \Longrightarrow V_{t-1}(-s_t) + \nu/t \geq V_t(-s_t) + \psi_{l_2,t}(w_t).$$

In the above proof we used the same evidence as for Lemma 2 but with respect to $\psi_{l_2,1}(w)$.

Finally to derive our bound in Theorem 4 we note that Eq.(34) in view of Lemma 4 is given as follows:

$$\begin{aligned} V_\tau(-s_\tau) + \psi_{l_2,\tau}(w_\tau) &\leq \nu/\tau + V_{\tau-1}(-s_{\tau-1}) \\ &+\langle -g_\tau, \hat{w}_\tau - w_0\rangle + \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))}. \end{aligned} \tag{39}$$

Next we rearrange and sum up everything *w.r.t.* index τ:

$$\begin{aligned} \sum_{\tau=1}^{t} (\langle g_\tau, \hat{w}_\tau - w_0\rangle + \psi_{l_2,\tau}(w_\tau)) + V_t(-s_t) &\leq \nu \log t \\ &+V_0(-s_0) + \sum_{\tau=1}^{t} \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))}. \end{aligned} \tag{40}$$

The inequality above is almost the same as Eq.(36) except for an additional $\nu \log t$ term. Next keeping in mind our initial bound on $U_t(s) \leq V_t(s) + \beta_t D$ and

taking into account Eq.(33), $V_0(-s_0) = V_0(0) = 0$ and the first-order optimality conditions of the function f we get:

$$\mathbf{R}_t(w) \leq \delta_t \leq \nu \log t + \beta_t D + \sum_{\tau=1}^{t} \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau - 1))}, \tag{41}$$

where one part of the right-hand side of inequality is exactly the same as in Eq.(37), so that we can immediately substitute it with the right-hand side of Eq.(38) and get:

$$\mathbf{R}_t(w) \leq \frac{G^2}{2\lambda} + \frac{G^2 + 2\lambda\nu}{2\lambda} \log t. \tag{42}$$

□