Supplementary (corrected) material to Reweighted l_2-Regularized Dual Averaging Approach for Highly Sparse Stochastic Learning

Vilen Jumutc and Johan A.K. Suykens

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

1 Reweighted l_2-RDA objective and algorithm

Reweighted l_2-RDA objective on every step t:

$$\phi_t(w) \triangleq \mathbb{E}_\xi[F(w,\xi)] + \psi_{l_2,t}(w) = \frac{1}{t}\sum_{\tau=1}^{t} f(w,\xi_\tau) + \|\Theta_t^{1/2} w\|_2^2 + \lambda\|w\|_2^2, \quad (1)$$

where $f(w,\xi_\tau)$ can represent any appropriate convex loss-function and $\psi_t(w)$ is at least λ-strongly convex function.

2 Reweighted l_2-RDA bounds for the regret function $\mathbf{R}_t(w)$

Theorem 1 *Let the sequences $\{w_t\}_{t\geq 1}$, $\{g_t\}_{t\geq 1}$ and $\{\Theta_t\}_{t\geq 1}$ be generated by Algorithm 1. 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:*

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

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**
  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)

Proof: We start our proof 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:

$$U_t(s) = \max_{w \in \mathcal{F}_\mathcal{D}} \{\langle s, w - w_0 \rangle - \sum_{\tau=1}^{t} \psi_{l_2,\tau}(w)\},$$
$$V_t(s) = \max_{w} \{\langle s, w - w_0 \rangle - \sum_{\tau=1}^{t} \psi_{l_2,\tau}(w) - \beta_t h(w)\}, \tag{3}$$

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

Proof 1

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

In 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 Algorithm 1 (line 6). And 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 by

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

with respect to $\psi_{l_2,t}(w)$. From [2] 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:

$$\begin{aligned} \delta_t &= \max_w \{\sum_{\tau=1}^{t} (\langle g_\tau, w_\tau - w\rangle \ + \Delta\psi_{l_2,\tau})\} \\ &\geq \sum_{\tau=1}^{t} (f(w_\tau, \xi_\tau) - f(w, \xi_\tau) + \Delta\psi_{l_2,\tau}) = \mathbf{R}_t(w) \end{aligned} \tag{4}$$

bounds the regret function from above due to the convexity of f [3]. 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.(4) then for any $w \in \mathbb{R}^n$ we get:

$$\begin{aligned} \delta_t &= \sum_{\tau=1}^{t} (\langle g_\tau, w_\tau - w_0\rangle \ + \psi_{l_2,\tau}(w_\tau)) + \max_w \{\langle s_t, w_0 - w\rangle - \sum_{\tau=1}^{t} \psi_{l_2,\tau}(w)\} \\ &\leq \sum_{\tau=1}^{t} (\langle g_\tau, w_\tau - w_0\rangle \ + \psi_{l_2,\tau}(w_\tau)) + V_t(-s_t) + \beta_t D, \end{aligned} \tag{5}$$

where $s_t = \sum_{\tau=1}^{t} g_\tau$. The last maximization term on the first line in Eq.(5) 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 [4, 2] for any $\tau \geq 1$ we have:

$$\begin{aligned} V_\tau(-s_\tau) + \psi_{l_2,\tau}(w_\tau) &\leq V_{\tau-1}(-s_\tau) \\ &= V_{\tau-1}(-s_{\tau-1} - g_\tau) \\ &\leq V_{\tau-1}(-s_{\tau-1}) + \langle -g_\tau, \nabla V_{\tau-1}(-s_{\tau-1})\rangle \\ &\quad + \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))} \\ &= V_{\tau-1}(-s_{\tau-1}) + \langle -g_\tau, \hat{w}_\tau - w_0\rangle \\ &\quad + \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))}, \end{aligned} \tag{6}$$

where $\|\cdot\|_*$ is a dual norm, the second inequality is due to the result of [4]: $V_\beta(s+g) \leq V_\beta(s) + \langle g, \nabla V_\beta(s)\rangle + \|g\|_*/(2\sigma\beta),\ \forall s, g \in E^*$, where $\sigma\beta$ term refers to the smoothing part, which is quite different in case of $V_t(s)$ in Eq.(3). Based on the result in [4] and Lemma 1 in [2], 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.(3) we have the Lipschitz continuous gradient of $V_t(s)$ with constant $\frac{1}{\beta_t + \lambda t}$, such that:

$$\|\nabla V_t(s_1) - \nabla V_t(s_2)\| \leq \frac{1}{\beta_t + \lambda t}\|s_1 - s_2\|_*, \quad \forall s_1, s_2 \in E^* \tag{7}$$

On the other hand we know a closed form for such a gradient in case of Eq.(6) and in view of conjugate support-type function $V_t(s)$ in Eq.(3):

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

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:

$$\begin{aligned}\sum_{\tau=1}^{t} (\langle g_\tau, \hat{w}_\tau - w_0\rangle + \psi_{l_2,\tau}(w_\tau)) + V_t(-s_t) \\ \leq V_0(-s_0) + \sum_{\tau=1}^{t} \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))}.\end{aligned} \tag{8}$$

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.(8) and taking into account Eq.(5) we get:

$$\mathbf{R}_t(w) \leq \delta_t \leq \beta_t D + \sum_{\tau=1}^{t} \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))}. \tag{9}$$

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

$$\mathbf{R}_t(w) \leq \frac{\|g_1\|_*^2}{2\lambda} + \sum_{\tau=1}^{t-1} \frac{\|g_{\tau+1}\|_*^2}{2\lambda\tau} \leq \frac{G^2}{2\lambda}(1 + \int_1^t \tau^{-1} d\tau) = \frac{G^2}{2\lambda}(1 + \log t). \tag{10}$$

□

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* $\psi_{l_2,t}(w_t) - \psi_{l_2,t}(w_{t+1}) \leq \nu/t$ *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*:*

$$\mathbf{R}_t(w) \leq \frac{G^2}{2\lambda} + \frac{G^2 + 2\lambda\nu}{2\lambda}\log t. \tag{11}$$

Proof: 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 1.

Lemma 2 *For each* $t \geq 1$ *in Algorithm 1 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 2

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

and to get the final result we use our assumption:

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

Finally to derive our bound in Theorem 2 we note that Eq.(6) in view of Lemma 2 is given as follows:

$$\begin{aligned}
V_\tau(-s_\tau) + \psi_{l_2,\tau}(w_\tau) &\leq \nu/\tau + V_{\tau-1}(-s_{\tau-1}) \\
&+\langle -g_\tau, \hat{w}_\tau - w_0\rangle + \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))}.
\end{aligned} \tag{12}$$

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

$$\begin{aligned}
\sum_{\tau=1}^{t}(\langle g_\tau, \hat{w}_\tau - w_0\rangle + \psi_{l_2,\tau}(w_\tau)) + V_t(-s_t) &\leq \nu\log t \\
&+V_0(-s_0) + \sum_{\tau=1}^{t}\frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau-1))}.
\end{aligned} \tag{13}$$

The inequality above is almost the same as Eq.(8) 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.(5), $V_0(-s_0) = V_0(0) = 0$ and the first-order optimality conditions of the function f we get:

$$\mathbf{R}_t(w) \leq \delta_t \leq \nu \log t + \beta_t D + \sum_{\tau=1}^{t} \frac{\|g_\tau\|_*^2}{2(\beta_{\tau-1} + \lambda(\tau - 1))}, \tag{14}$$

where one part of the right-hand side of inequality is exactly the same as in Eq.(9), so that we can immediately substitute it with the right-hand side of Eq.(10) and get:

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

□

References

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

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

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

[4] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.