SDCA without Duality, Regularization, and Individual Convexity

Shai Shalev-Shwartz SHAIS@CS.HUJI.AC.IL
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

Abstract

Stochastic Dual Coordinate Ascent is a popular method for solving regularized loss minimization for the case of convex losses. We describe variants of SDCA that do not require explicit regularization and do not rely on duality. We prove linear convergence rates even if individual loss functions are non-convex, as long as the expected loss is strongly convex.

1. Introduction

We consider the following loss minimization problem:

$$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) \ .$$

An important sub-class of problems is when each f_i can be written as $f_i(w) = \phi_i(w) + \frac{\lambda}{2}\|w\|^2$, where ϕ_i is L_i-smooth and convex. A popular method for solving this sub-class of problems is Stochastic Dual Coordinate Ascent (SDCA), and (Shalev-Shwartz & Zhang, 2013) established the convergence rate of $\tilde{O}((L_{\max}/\lambda + n)\log(1/\epsilon))$, where $L_{\max} = \max_i L_i$.

As its name indicates, SDCA is derived by considering a dual problem. In this paper, we consider the possibility of applying SDCA for problems in which individual f_i do not necessarily have the form $\phi_i(w) + \frac{\lambda}{2}\|w\|^2$, and can even be non-convex (e.g., deep learning optimization problems, or problems arising in fast calculation of the top singular vectors (Jin et al., 2015)). In many such cases, the dual problem is meaningless. Instead of directly using the dual problem, we describe and analyze a variant of SDCA in which only gradients of f_i are being used. Following (Johnson & Zhang, 2013), we show that SDCA is a member of the Stochastic Gradient Descent (SGD) family of algorithms, that is, its update is based on an unbiased estimate of the gradient, but unlike the vanilla SGD, for SDCA the variance of the estimation of the gradient tends to zero as we converge to a minimum.

Our analysis assumes that F is λ-strongly convex and each f_i is L_i-smooth. When each f_i is also convex we establish the convergence rate of $\tilde{O}(\bar{L}/\lambda + n)$, where $\bar{L}$ is the average of L_i and the $\tilde{O}$ notation hides logarithmic terms, including the factor $\log(1/\epsilon)$. This matches the best known bound for SVRG given in (Xiao & Zhang, 2014). Lower bounds have been derived in (Arjevani et al., 2015; Agarwal & Bottou, 2014). Applying an acceleration technique ((Shalev-Shwartz & Zhang, 2015; Lin et al., 2015)) we obtain the convergence rate $\tilde{O}(n^{1/2}\sqrt{\bar{L}/\lambda} + n)$. If f_i are non-convex we first prove that SDCA enjoys the rate $\tilde{O}(\bar{L}^2/\lambda^2 + n)$. Finally, we show how the acceleration technique yields the bound $\tilde{O}\left(n^{3/4}\sqrt{\bar{L}/\lambda} + n\right)$. That is, we have the same dependency on the square root of the condition number, $\sqrt{\bar{L}/\lambda}$, but this term is multiplied by $n^{3/4}$ rather than by $n^{1/2}$. Understanding if this factor can be eliminated is left to future work.

Related work: In recent years, many randomized methods for optimizing average of functions have been proposed. For example, SAG (Le Roux et al., 2012), SVRG (Johnson & Zhang, 2013), Finito (Defazio et al., 2014b), SAGA (Defazio et al., 2014a), S2GD (Konečný & Richtárik, 2013), and UniVr (Allen-Zhu & Yuan, 2015). All of these methods have similar convergence rates for strongly convex and smooth problems. Here we show that SDCA achieves the best known convergence rate for the case in which individual loss functions are convex, and a slightly worse rate for the case in which individual loss functions are non-convex. A systematic study of the convergence rate of the different methods under non-convex losses is left to future work.

This version of the paper improves upon a previous unpublished version of the paper (Shalev-Shwartz, 2015) in three aspects. First, the convergence rate here depends on $\bar{L}$ as opposed to $L_{\max}$ in (Shalev-Shwartz, 2015). Second, the version in (Shalev-Shwartz, 2015) only deals with the regularized case, while here we show that the same rate can be obtained for unregularized objectives. Last, for the non-convex case, here we derive

Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

the bound $\tilde{O}\left(n^{3/4}\sqrt{\bar{L}/\lambda}+n\right)$ while in (Shalev-Shwartz, 2015) only the bound of $\tilde{O}(L_{\max}^2/\lambda^2+n)$ has been given.

(Csiba & Richtárik, 2015) extended the work of (Shalev-Shwartz, 2015) to support arbitrary mini-batching schemes, and (He & Takáč, 2015) extended the work of (Shalev-Shwartz, 2015) to support adaptive sampling probabilities. A primal form of SDCA has been also given in (Defazio, 2014). Using SVRG for non-convex individual functions has been recently studied in (Shamir, 2015; Jin et al., 2015), in the context of fast computation of the top singular vectors of a matrix.

2. SDCA without Duality

We start the section by describing a variant of SDCA that do not rely on duality. To simplify the presentation, we start in Section 2.1 with regularized loss minimization problems. In Section 2.2 we tackle the non-regularized case and in Section 2.3 we tackle the non-convex case.

We recall the following basic definitions: A (differentiable) function f is λ-strongly convex if for every u, w we have $f(w) - f(u) \geq \nabla f(u)^\top (w-u) + \frac{\lambda}{2}\|w-u\|^2$. We say that f is convex if it is 0-strongly convex. We say that f is L-smooth if $\|\nabla f(w) - \nabla f(u)\| \leq L\|w-u\|$. It is well known that smoothness and convexity also implies that $f(w) - f(u) \leq \nabla f(u)^\top (w-u) + \frac{L}{2}\|w-u\|^2$.

2.1. Regularized problems

In regularized problems, each f_i can be written as $f_i(w) = \phi_i(w) + \frac{\lambda}{2}\|w\|^2$. Similarly to the original SDCA algorithm, we maintain vectors $\alpha_1, \ldots, \alpha_n$, where each $\alpha_i \in \mathbb{R}^d$. We call these vectors pseudo-dual vectors. The algorithm is described below.

Algorithm 1: Dual-Free SDCA for Regularized Objectives

Goal: Minimize $F(w) = \frac{1}{n}\sum_{i=1}^n \phi_i(w) + \frac{\lambda}{2}\|w\|^2$
Input: Objective F, number of iterations T,
step size η,
Smoothness parameters $L_1, \ldots, L_n$
Initialize: $w^{(0)} = \frac{1}{\lambda n}\sum_{i=1}^n \alpha_i^{(0)}$
for some $\alpha^{(0)} = (\alpha_1^{(0)}, \ldots, \alpha_n^{(0)})$
$\forall i \in [n]$, $q_i = (L_i + \bar{L})/(2n\bar{L})$
where $\bar{L} = \frac{1}{n}\sum_{i=1}^n L_i$
For $t = 1, \ldots, T$
Pick $i \sim q$, denote $\eta_i = \frac{\eta}{q_i n}$
Update:
$\alpha_i^{(t)} = \alpha_i^{(t-1)} - \eta_i \lambda n\left(\nabla\phi_i(w^{(t-1)}) + \alpha_i^{(t-1)}\right)$
$w^{(t)} = w^{(t-1)} - \eta_i\left(\nabla\phi_i(w^{(t-1)}) + \alpha_i^{(t-1)}\right)$

Observe that SDCA keeps the primal-dual relation

$$w^{(t-1)} = \frac{1}{\lambda n}\sum_{i=1}^n \alpha_i^{(t-1)}$$

Observe also that the update of α can be rewritten as

$$\alpha_i^{(t)} = (1-\beta_i)\alpha_i^{(t-1)} + \beta_i\left(-\nabla\phi_i(w^{(t-1)})\right) ,$$

where $\beta_i = \eta_i \lambda n$. Namely, the new value of α_i is a convex combination of its old value and the negative gradient. Finally, observe that, conditioned on the value of $w^{(t-1)}$ and $\alpha^{(t-1)}$, we have that

$$\begin{aligned}
&\mathbb{E}_{i\sim q}[w^{(t)}] = w^{(t-1)} - \eta\sum_i \frac{q_i}{q_i n}(\nabla\phi_i(w^{(t-1)}) + \alpha_i^{(t-1)}) \\
&= w^{(t-1)} - \eta\left(\nabla\frac{1}{n}\sum_{i=1}^n \phi_i(w^{(t-1)}) + \lambda w^{(t-1)}\right) \\
&= w^{(t-1)} - \eta\nabla P(w^{(t-1)}) .
\end{aligned}$$

That is, SDCA is in fact an instance of Stochastic Gradient Descent (SGD). As we will see shortly, the advantage of SDCA over a vanilla SGD algorithm is because the *variance* of the update goes to zero as we converge to an optimum.

Our convergence analysis relies on bounding the following potential function, defined for every $t \geq 0$,

$$C_t = \frac{\lambda}{2}\|w^{(t)} - w^*\|^2 + \frac{\eta}{n^2}\sum_{i=1}^n\left[\frac{1}{q_i}\|\alpha_i^{(t)} - \alpha_i^*\|^2\right] , \quad (1)$$

where

$$w^* = \operatorname*{argmin}_w F(w), \text{ and } \forall i,\ \alpha_i^* = -\nabla\phi_i(w^*) . \quad (2)$$

Intuitively, C_t measures the distance to the optimum both in primal and pseudo-dual variables. Observe that if F is L_F-smooth and convex then

$$F(w^{(t)}) - F(w^*) \leq \frac{L_F}{2}\|w^{(t)} - w^*\|^2 \leq \frac{L_F}{\lambda}C_t ,$$

and therefore a bound on C_t immediately implies a bound on the sub-optimality of $w^{(t)}$.

The following theorem establishes the convergence rate of SDCA for the case in which each ϕ_i is convex.

Theorem 1 *Assume that each ϕ_i is L_i-smooth and convex, and Algorithm 1 is run with $\eta \leq \min\left\{\frac{1}{4\bar{L}}, \frac{1}{4\lambda n}\right\}$. Then, for every $t \geq 1$,*

$$\mathbb{E}[C_t] \leq (1-\eta\lambda)^t C_0 ,$$

where C_t is as defined in (1). In particular, to achieve $\mathbb{E}[F(w^{(T)}) - F(w^*)] \leq \epsilon$ it suffices to set $\eta = \min\left\{\frac{1}{4\bar{L}}, \frac{1}{4\lambda n}\right\}$ and

$$T \geq \tilde{\Omega}\left(\frac{\bar{L}}{\lambda} + n\right) .$$

Variance Reduction: The lemma below tells us that the variance of the SDCA update decreases as we get closer to the optimum.

Lemma 1 *Under the same conditions of Theorem 1, the expected value of* $\|w^{(t)} - w^{(t-1)}\|^2$ *conditioned on* $w^{(t-1)}$ *satisfies:*

$$\mathbb{E}[\|w^{(t)}-w^{(t-1)}\|^2] \le 3\,\eta \left(\tfrac{1}{2}\|w^{(t-1)} - w^*\|^2 + C_{t-1}\right) .$$

2.2. SDCA without regularization

We now turn to the case in which the objective is not explicitly regularized. The algorithm below tackles this problem by a reduction to the regularized case. In particular, we artificially add regularization to the objective and compensate for it by adding one more loss function that cancels out the regularization term. While the added function is not convex (in fact, it is concave), we prove that the same convergence rate holds due to the special structure of the added loss function.

Algorithm 2: Dual-Free SDCA for Non-Regularized Objectives

Goal: Minimize $F(w) = \frac{1}{n}\sum_{i=1}^n f_i(w)$
Input: Objective F, number of iterations T,
step size η, Strong convexity parameter λ,
Smoothness parameters $L_1, \ldots, L_n$
Define:
For all $i \in [n]$, $\phi_i(w) = \frac{n+1}{n} f_i(w)$, $\tilde{L}_i = \frac{n+1}{n} L_i$
For $i = n+1$, $\phi_i(w) = \frac{-\lambda\, i}{2}\|w\|^2$, $\tilde{L}_i = \lambda\, i$
Solve:
Rewrite F as $F(w) = \frac{1}{n+1}\sum_{i=1}^{n+1}\phi_i(w) + \frac{\lambda}{2}\|w\|^2$
Call Algorithm 1 with F above and with $\{\tilde{L}_i\}$

Theorem 2 *Assume that F is λ-strongly convex, that each f_i is L_i-smooth and convex, and that Algorithm 2 is run with* $\eta \le \min\left\{\frac{1}{8(\bar{L}+\lambda)}, \frac{1}{4\,\lambda(n+1)}\right\}$. *Then, for every* $t \ge 1$,

$$\mathbb{E}[C_t] \le (1-\eta\lambda)^t\, C_0 ,$$

where C_t is as defined in (1). In particular, to achieve $\mathbb{E}[F(w^{(T)}) - F(w^*)] \le \epsilon$ *it suffices to set* $\eta = \min\left\{\frac{1}{8(\bar{L}+\lambda)}, \frac{1}{4\,\lambda(n+1)}\right\}$ *and*

$$T \ge \tilde{\Omega}\left(\frac{\bar{L}}{\lambda} + n\right) .$$

2.3. The non-convex case

We now consider the non-convex case. For simplicity, we focus on the regularized setting. In the non-regularized setting we can simply replace every f_i with $\phi_i(w) = f_i(w) - \frac{\lambda}{2}\|w\|^2$ and apply the regularized setting. Note that this does not change significantly the smoothness (because λ is typically much smaller than the average smoothness of the f_i).

We can apply Algorithm 1 for the non-convex case, and the only change is the choice of η, as reflected in the theorem below.

Theorem 3 *Consider running algorithm 1 on F which is λ-strongly convex, assume that each ϕ_i is L_i-smooth, and* $\eta \le \min\left\{\frac{\lambda}{4\bar{L}^2}, \frac{1}{4\,\lambda n}\right\}$. *Then, for every* $t \ge 1$,

$$\mathbb{E}[C_t] \le (1-\eta\lambda)^t\, C_0 ,$$

where C_t is as defined in (1). In particular, to achieve $\mathbb{E}[F(w^{(T)}) - F(w^*)] \le \epsilon$ *it suffices to set* $\eta = \min\left\{\frac{\lambda}{4\bar{L}^2}, \frac{1}{4\,\lambda n}\right\}$ *and*

$$T \ge \tilde{\Omega}\left(\frac{\bar{L}^2}{\lambda^2} + n\right) .$$

As can be seen, the dependence of T on the condition number, $\frac{\bar{L}}{\lambda}$, is quadratic for the non-convex case, as opposed to a linear dependency for the convex case. We next show how to improve the bound using acceleration.

2.4. Acceleration

Accelerated SDCA (Shalev-Shwartz & Zhang, 2015) is obtained by solving (using SDCA) a sequence of problems, where at each iteration, we add an artificial regularization of the form $\frac{\kappa}{2}\|w - y^{(t-1)}\|^2$, where $y^{(t-1)}$ is a function of $w^{(t-1)}$ and $w^{(t-2)}$. The algorithm has been generalized in (Lin et al., 2015) to allow the inner solver to be any algorithm. For completeness, we provide the pseudo-code of the "Catalyst" algorithm of (Lin et al., 2015) and its analysis.

Algorithm 3: Acceleration

Goal: Minimize a λ-strongly convex function $F(w)$
Parameters: κ, T
Initialize:
Initial solution $w^{(0)}$
ϵ_0 s.t. $\epsilon_0 \ge F(w^{(0)}) - F(w^*)$
$y^{(0)} = w^{(0)}$, $q = \frac{\lambda}{\lambda+\kappa}$
For: $t = 1, \ldots, T$
Define $G_t(w) = F(w) + \frac{\kappa}{2}\|w - y^{(t-1)}\|^2$
Set $\epsilon_t = (1 - 0.9\sqrt{q})\,\epsilon_{t-1}$
Find $w^{(t)}$ s.t. $G_t(w^{(t)}) - \min_w G_t(w) \le \epsilon_t$
Set $y^{(t)} = w^{(t)} + \frac{\sqrt{q}-q}{\sqrt{q}+q}(w^{(t)} - w^{(t-1)})$
Output: $w^{(T)}$

Lemma 2 *Fix $\epsilon > 0$ and suppose we run the Acceleration algorithm (Algorithm 3) for*

$$T = \Omega\left(\sqrt{\frac{\lambda+\kappa}{\lambda}}\,\log\left(\frac{\lambda+\kappa}{\lambda\,\epsilon}\right)\right)$$

iterations. Then, $F(w^{(T)}) - F(w^*) \le \epsilon$.

Proof The lemma follows directly from Theorem 3.1 of (Lin et al., 2015) by observing that Algorithm 3 is a specification of Algorithm 1 in (Lin et al., 2015) with $\alpha_0 = \sqrt{q}$ (which implies that $\alpha_t = \alpha_0$ for every t), with $\epsilon_t = \epsilon_0(1-\rho)^t$, and with $\rho = 0.9\sqrt{q}$. ■

Theorem 4 *Let $F = \frac{1}{n}\sum_{i=1}^n \phi_i(w) + \frac{\lambda}{2}\|w\|^2$, assume that each ϕ_i is L_i smooth and that F is λ-strongly convex. Assume also that $(\bar{L}/\lambda)^2 \ge 3n$ (otherwise we can simply apply $\tilde{O}(n)$ iterations of Algorithm 1). Then, running Algorithm 3 with parameters $\kappa = \bar{L}/\sqrt{n}$, $T = \tilde{\Omega}\left(1 + n^{-1/4}\sqrt{\bar{L}/\lambda}\right)$, and while at each iteration of Algorithm 3 using $\tilde{\Omega}\left(n\right)$ iterations of Algorithm 1 to minimize G_t, guarantees that $F(w^{(T)}) - F(w^*) \le \epsilon$ (with high probability). The total required number of iterations of Algorithm 1 is therefore bounded by $\tilde{O}\left(n + n^{3/4}\sqrt{\bar{L}/\lambda}\right)$.*

Observe that for the case of convex individual functions, accelerating Algorithm 1 yields the upper bound $\tilde{O}\left(n + n^{1/2}\sqrt{\bar{L}/\lambda}\right)$. Therefore, the convex and non-convex cases have the same dependency on the condition number, but the non-convex case has a worse dependence on n.

3. Proofs

3.1. Proof of Theorem 1

Observe that $0 = \nabla F(w^*) = \frac{1}{n}\sum_i \nabla\phi_i(w^*) + \lambda w^*$, which implies that $w^* = \frac{1}{\lambda n}\sum_i \alpha_i^*$, where $\alpha_i^* = -\nabla\phi_i(w^*)$.

Define $u_i = -\nabla\phi_i(w^{(t-1)})$ and $v_i = -u_i + \alpha_i^{(t-1)}$. We also denote two potentials:

$$A_t = \sum_{j=1}^n \frac{1}{q_j}\|\alpha_j^{(t)} - \alpha_j^*\|^2 \quad , \quad B_t = \|w^{(t)} - w^*\|^2 \;.$$

We will first analyze the evolution of A_t and B_t. If on round t we update using element i then $\alpha_i^{(t)} = (1-\beta_i)\alpha_i^{(t-1)} + \beta_i u_i$. It follows that,

$$A_{t-1} - A_t = -\frac{1}{q_i}\|\alpha_i^{(t)} - \alpha_i^*\|^2 + \frac{1}{q_i}\|\alpha_i^{(t-1)} - \alpha_i^*\|^2 \tag{3}$$

$$\begin{aligned}
&= -\frac{1}{q_i}\|(1-\beta_i)(\alpha_i^{(t-1)} - \alpha_i^*) + \beta_i(u_i - \alpha_i^*)\|^2 \\
&\quad + \frac{1}{q_i}\|\alpha_i^{(t-1)} - \alpha_i^*\|^2 \\
&= \frac{1}{q_i}(-(1-\beta_i)\|\alpha_i^{(t-1)} - \alpha_i^*\|^2 - \beta_i\|u_i - \alpha_i^*\|^2 \\
&\quad + \beta_i(1-\beta_i)\|\alpha_i^{(t-1)} - u_i\|^2 + \|\alpha_i^{(t-1)} - \alpha_i^*\|^2\ ) \\
&= \frac{\beta_i}{q_i}\left(\|\alpha_i^{(t-1)} - \alpha_i^*\|^2 - \|u_i - \alpha_i^*\|^2 + (1-\beta_i)\|v_i\|^2\right) \\
&= \frac{\eta\,\lambda}{q_i^2}\left(\|\alpha_i^{(t-1)} - \alpha_i^*\|^2 - \|u_i - \alpha_i^*\|^2 + (1-\beta_i)\|v_i\|^2\right) \;.
\end{aligned} \tag{4}$$

Taking expectation w.r.t. $i \sim q$ we obtain

$$\mathbb{E}[A_{t-1} - A_t] = \tag{5}$$

$$\begin{aligned}
&\sum_{i=1}^n \frac{\eta\lambda}{q_i}\left(\|\alpha_i^{(t-1)} - \alpha_i^*\|^2 - \|u_i - \alpha_i^*\|^2 + (1-\beta_i)\|v_i\|^2\right) \\
&= \eta\lambda\left(A_{t-1} + \sum_{i=1}^n \frac{1}{q_i}\left(-\|u_i - \alpha_i^*\|^2 + (1-\beta_i)\|v_i\|^2\right)\right) .
\end{aligned}$$

As to the second potential, we have

$$\begin{aligned}
B_{t-1} - B_t &= -\|w^{(t)} - w^*\|^2 + \|w^{(t-1)} - w^*\|^2 \\
&= 2\,(w^{(t-1)} - w^*)^\top(\eta\,v_i) - \eta_i^2\|v_i\|^2 \;.
\end{aligned} \tag{6}$$

Taking expectation w.r.t. $i \sim q$ and noting that $\mathbb{E}_{i\sim q}(\eta_i v_i) = \eta\nabla F(w^{(t-1)})$ we obtain

$$\begin{aligned}
\mathbb{E}[B_{t-1} - B_t] =& 2\eta\,(w^{(t-1)} - w^*)^\top\nabla F(w^{(t-1)}) \\
& - \frac{\eta^2}{n^2}\sum_i \frac{1}{q_i}\|v_i\|^2 \;.
\end{aligned} \tag{7}$$

We now take a potential of the form $C_t = c_a A_t + c_b B_t$. Combining (5) and (7) we obtain

$$\begin{aligned}
\mathbb{E}[C_{t-1} - C_t] &= c_a\eta\lambda A_{t-1} - c_a\eta\lambda\sum_i \frac{1}{q_i}\|u_i - \alpha_i^*\|^2 \\
&\quad + 2c_b\eta(w^{(t-1)} - w^*)^\top\nabla F(w^{(t-1)}) \\
&\quad + \sum_i \frac{1}{q_i}\|v_i\|^2\left(c_a\eta\lambda(1-\beta_i) - \frac{c_b\eta^2}{n^2}\right)
\end{aligned} \tag{8}$$

We will choose the parameters η, c_a, c_b such that

$$\eta \le \min\left\{\frac{q_i}{2\lambda}\,,\,\frac{1}{4\bar{L}}\right\} \quad\text{and}\quad \frac{c_b}{c_a} = \frac{\lambda n^2}{2\eta} \tag{9}$$

This implies that $\beta_i = \eta_i \lambda n = \frac{\eta\lambda}{q_i} \le 1/2$, and therefore the term in (8) is non-negative. Next, due to strong convexity of F we have that

$$(w^{(t-1)} - w^*)^\top \nabla F(w^{(t-1)}) \ge F(w^{(t-1)}) - F(w^*) + \frac{\lambda}{2}\|w^{(t-1)} - w^*\|^2 .$$

Therefore,

$$\mathbb{E}[C_{t-1} - C_t] = c_a \eta\lambda A_{t-1} - c_a \eta\lambda \sum_i \frac{1}{q_i}\|u_i - \alpha_i^*\|^2 + 2c_b\eta(F(w^{(t-1)}) - F(w^*)) + c_b\eta\lambda B_{t-1}$$
$$= \eta\,\lambda\,C_{t-1} + \eta\left(2c_b(F(w^{(t-1)}) - F(w^*)) - c_a\lambda \sum_i \frac{1}{q_i}\|u_i - \alpha_i^*\|^2\right). \tag{10}$$

Note that $u_i - \alpha_i^* = \nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)$. In Lemma 3 we show that when ϕ_i is L_i smooth and convex then

$$\|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 \le 2\,L_i\,(\phi_i(w^{(t-1)}) - \phi_i(w^*) - \nabla\phi_i(w^*)^\top(w^{(t-1)} - w^*)) \tag{11}$$

Therefore, denoting $\tau = \left(2\,\max_i \frac{L_i}{q_i}\right)$ we obtain that

$$\sum_i \frac{1}{q_i}\|u_i - \alpha_i^*\|^2 = \sum_i \frac{1}{q_i}\|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 \tag{12}$$
$$\le \tau \sum_i (\phi_i(w^{(t-1)}) - \phi_i(w^*) - \nabla\phi_i(w^*)^\top(w^{(t-1)} - w^*))$$
$$= \tau\, n\left(F(w^{(t-1)}) - F(w^*) - \frac{\lambda}{2}\|w^{(t-1)} - w^*\|^2\right)$$
$$\le \tau\, n\left(F(w^{(t-1)}) - F(w^*)\right) . \tag{13}$$

The definition of q_i implies that for every i,

$$\frac{L_i}{q_i} = 2n\bar{L}\,\frac{L_i}{L_i + \bar{L}} \le 2n\bar{L} . \tag{14}$$

Combining this with (12) and (10) we obtain

$$\mathbb{E}[C_{t-1} - C_t] \ge \eta\,\lambda\,C_{t-1} + \eta\left(2c_b - 4n^2\bar{L}\lambda c_a\right)(F(w^{(t-1)}) - F(w^*))$$

Plugging the value of $c_b = \frac{c_a\lambda n^2}{2\eta}$ yields that the coefficient in the last term is

$$2\frac{c_a\lambda n^2}{2\eta} - 4n^2\bar{L}\lambda c_a = c_a\lambda n^2\left(\frac{1}{\eta} - 4\bar{L}\right) \ge 0 ,$$

where we used the choice of $\eta \le \frac{1}{4\bar{L}}$. In summary, we have shown that $\mathbb{E}[C_{t-1} - C_t] \ge \eta\,\lambda\,C_{t-1}$, which implies that

$$\mathbb{E}[C_t] \;\le\; (1 - \eta\,\lambda)\,C_{t-1} .$$

Taking expectation over C_{t-1} and continue recursively, we obtain that $\mathbb{E}[C_t] \;\le\; (1-\eta\,\lambda)^t\,C_0 \;\le\; e^{-\eta\,\lambda\,t}\,C_0$.

Finally, since $q_i \ge 1/(2n)$ for every i, we can choose

$$\eta = \min\left\{\frac{1}{4\bar{L}}\,,\;\frac{1}{4\,\lambda n}\right\}$$

and therefore

$$\frac{1}{\eta\lambda} \le 4\left(n + \frac{\bar{L}}{\lambda}\right) .$$

The proof is concluded by choosing $c_b = \lambda/2$ and $c_a = \eta/n^2$.

3.2. Proof of Lemma 1

We have:

$$\mathbb{E}[\|w^{(t)} - w^{(t-1)}\|^2] = \sum_i q_i\eta_i^2\|\nabla\phi_i(w^{(t-1)}) + \alpha_i^{(t-1)}\|^2$$
$$\le \frac{3\eta^2}{n^2}\sum_i \frac{1}{q_i}(\|\nabla\phi_i(w^{(t-1)}) + \alpha_i^*\|^2 + \|\alpha_i^{(t-1)} - \alpha_i^*\|^2) \quad \text{(triangle inequality)}$$
$$= \frac{3\eta^2}{n^2}\sum_i (\tfrac{1}{q_i}\|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2 + \tfrac{1}{q_i}\|\alpha_i^{(t-1)} - \alpha_i^*\|^2)$$
$$\le \frac{3\eta^2}{n^2}\sum_i \left(2n\bar{L}\,\|w^{(t-1)} - w^*\|^2 + \tfrac{1}{q_i}\|\alpha_i^{(t-1)} - \alpha_i^*\|^2\right) \quad \text{(smoothness and (14))}$$
$$\le 3\,\eta\,\left(\tfrac{1}{2}\|w^{(t-1)} - w^*\|^2 + C_{t-1}\right) \quad \text{(because } \eta \le \tfrac{1}{4\bar{L}}\text{)} .$$

3.3. Proof of Theorem 2

The beginning of the proof is identical to the proof of Theorem 1. The change starts in (12), where we cannot apply (11) to ϕ_{n+1} because it is not convex. To overcome this, we first apply (11) to $\phi_1, \ldots, \phi_n$, and obtain that

$$\sum_{i=1}^n \frac{1}{q_i}\|u_i - \alpha_i^*\|^2 = \sum_{i=1}^n \frac{1}{q_i}\|\nabla\phi_i(w^{(t-1)}) - \nabla\phi_i(w^*)\|^2$$
$$\le \left(2\,\max_i \frac{\tilde{L}_i}{q_i}\right) \cdot \sum_{i=1}^n (\phi_i(w^{(t-1)}) - \phi_i(w^*) - \nabla\phi_i(w^*)^\top(w^{(t-1)} - w^*))$$
$$= 2\,(n+1)\left(\max_i \frac{\tilde{L}_i}{q_i}\right)(F(w^{(t-1)}) - F(w^*)) ,$$

where the last equality follows from the fact that $\sum_{i=1}^n \phi_i(w) = (n+1)F(w)$, which also implies that $\sum_i \nabla\phi_i(w^*) = 0$. In addition, since $\phi_{n+1}(w) = -\frac{\lambda(n+1)}{2}\|w\|^2$, we have

$$\begin{aligned}&\frac{1}{q_{n+1}}\|\nabla\phi_{n+1}(w) - \nabla\phi_{n+1}(w^*)\|^2\\ &= \frac{\lambda^2(n+1)^2}{q_{n+1}}\|w-w^*\|^2\\ &= 2\,(n+1)\,\frac{\tilde{L}_{n+1}}{q_{n+1}}\cdot\frac{\lambda}{2}\,\|w-w^*\|^2\\ &\le 2\,(n+1)\,\frac{\tilde{L}_{n+1}}{q_{n+1}}(F(w)-F(w^*))\ ,\end{aligned}$$

where the last inequality is because of the λ-strong convexity of F. Combining the two inequalities, we obtain an analogue of (12),

$$\begin{aligned}&\sum_{i=1}^{n+1}\frac{1}{q_i}\|u_i-\alpha_i^*\|^2\\ &\le\ 4\,(n+1)\left(\max_{i\in[n+1]}\frac{\tilde{L}_i}{q_i}\right)\,(F(w^{(t-1)})-F(w^*))\ .\end{aligned}$$

The rest of the proof is almost identical, except that we have n replaced by $n+1$ and $\bar{L}$ replaced by $\tilde{L} := \frac{1}{n+1}\sum_{i=1}^n \tilde{L}_i$. We now need to choose

$$\eta = \min\left\{\frac{1}{8\tilde{L}}\ ,\ \frac{1}{4\,\lambda(n+1)}\right\}\ .$$

Observe that,

$$(n+1)\tilde{L} = \frac{n+1}{n}\left(\sum_{i=1}^n L_i\right)+\lambda(n+1) = (n+1)(\bar{L}+\lambda)\ ,$$

so we can rewrite

$$\eta = \min\left\{\frac{1}{8(\bar{L}+\lambda)}\ ,\ \frac{1}{4\,\lambda(n+1)}\right\}\ .$$

This yields

$$\frac{1}{\eta\lambda} \le 4\left(n+3+\frac{2\bar{L}}{\lambda}\right)\ .$$

3.4. Proof of Theorem 3

The beginning of the proof is identical to the proof of Theorem 1 up to (8).

We will choose the parameters η, c_a, c_b such that

$$\eta \le \min\left\{\frac{q_i}{2\lambda}\ ,\ \frac{1}{4\bar{L}}\right\}\quad\text{and}\quad \frac{c_b}{c_a} = \frac{\lambda n^2}{2\eta} \tag{15}$$

This implies that $\beta_i = \eta_i\lambda n = \frac{\eta\lambda}{q_i} \le 1/2$, and therefore the term in (8) is non-negative. Next, due to strong convexity of F we have that

$$\begin{aligned}&(w^{(t-1)}-w^*)^\top\nabla F(w^{(t-1)})\\ &\ge\ F(w^{(t-1)})-F(w^*)+\frac{\lambda}{2}\|w^{(t-1)}-w^*\|^2\\ &\ge\ \lambda\|w^{(t-1)}-w^*\|^2\ .\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}[C_{t-1}-C_t]\\ &= c_a\eta\lambda A_{t-1} - c_a\eta\lambda\sum_i\frac{1}{q_i}\|u_i-\alpha_i^*\|^2 + 2c_b\eta\lambda B_{t-1}\\ &= \eta\,\lambda\,C_{t-1}+\eta\,\lambda\left(c_bB_{t-1}-c_a\sum_i\frac{1}{q_i}\|u_i-\alpha_i^*\|^2\right)\ .\end{aligned} \tag{16}$$

Next, we use the smoothness of the ϕ_i to get

$$\begin{aligned}&\sum_i\frac{1}{q_i}\|u_i-\alpha_i^*\|^2 = \sum_i\frac{1}{q_i}\|\nabla\phi_i(w^{(t-1)})-\nabla\phi_i(w^*)\|^2\\ &\le\sum_i\frac{L_i^2}{q_i}\|w^{(t-1)}-w^*\|^2 = B_{t-1}\sum_i\frac{L_i^2}{q_i}\ .\end{aligned}$$

The definition of q_i implies that for every i,

$$\frac{L_i}{q_i} = 2n\bar{L}\,\frac{L_i}{L_i+\bar{L}} \le 2n\bar{L}\ ,$$

so by combining with (16) we obtain

$$\mathbb{E}[C_{t-1}-C_t] \ge \eta\,\lambda\,C_{t-1}+\eta\lambda\left(c_b-2n^2\bar{L}^2c_a\right)B_{t-1}$$

The last term will be non-negative if $\frac{c_b}{c_a}\ge 2n^2\bar{L}^2$. Since we chose $\frac{c_b}{c_a}=\frac{\lambda n^2}{2\eta}$ we obtain the requirement

$$\frac{\lambda n^2}{2\eta}\ge 2n^2\bar{L}^2\ \Rightarrow\ \eta\le\frac{\lambda}{4\bar{L}^2}\ .$$

In summary, we have shown that $\mathbb{E}[C_{t-1}-C_t]\ge\eta\,\lambda\,C_{t-1}$. The rest of the proof is identical, but the requirement on η is

$$\eta\le\min\left\{\frac{\lambda}{4\bar{L}^2}\ ,\ \frac{1}{4\,\lambda n}\right\}\ ,$$

and therefore

$$\frac{1}{\eta\lambda}\le 4\left(n+\frac{\bar{L}^2}{\lambda^2}\right)\ .$$

4. Proof of Theorem 4

Proof Each iteration of Algorithm 3 requires to minimize G_t to accuracy $\epsilon_t \le O(1)\,(1-\rho)^t$, where $\rho = 0.9\,\sqrt{q}$. If

$t \le T$ where T is as defined in Lemma 2, then we have that,

$$-t\log(1-\rho) \le -T\log(1-\rho) = \frac{-\log(1-\rho)}{\rho}\log\left(\frac{800}{q\,\epsilon}\right)$$

Using Lemma 4, $\frac{-\log(1-\rho)}{\rho} \le 2$ for every $\rho \in (0, 1/2)$. In our case, ρ is indeed in $(0, 1/2)$ because of the definition of κ and our assumption that $(\bar{L}/\lambda)^2 \ge 3n$. Hence,

$$\log(\tfrac{1}{\epsilon_t}) \;=\; O(\log((\lambda+\kappa)/(\lambda\epsilon)))\ .$$

Combining this with Theorem 3, and using the definition of G_t, we obtain that the number of iterations required[1] by each application of Algorithm 3 is

$$\tilde{O}\left(\frac{(\bar{L}+\kappa)^2}{(\lambda+\kappa)^2} + n\right) = \tilde{O}(n)\ ,$$

where in the equality we used the definition of κ. Finally, multiplying this by the value of T as given in Lemma 2 we obtain (ignoring log-terms):

$$\sqrt{1+\frac{\kappa}{\lambda}}\,n \;\le\; (1+\sqrt{\frac{\kappa}{\lambda}})\,n = n + n^{3/4}\sqrt{\frac{\bar{L}}{\lambda}}\ .$$

■

4.1. Technical Lemmas

Lemma 3 *Assume that ϕ is L-smooth and convex. Then, for every w and u,*

$$\|\nabla\phi(w) - \nabla\phi(u)\|^2 \le 2L\left[\phi(w) - \phi(u) - \nabla\phi(u)^\top(w-u)\right]\ .$$

Proof For every i, define

$$g(w) = \phi(w) - \phi(u) - \nabla\phi(u)^\top(w-u)\ .$$

Clearly, since ϕ is L-smooth so is g. In addition, by convexity of ϕ we have $g(w) \ge 0$ for all w. It follows that g is non-negative and smooth, and therefore, it is self-bounded (see Section 12.1.3 in (Shalev-Shwartz & Ben-David, 2014)):

$$\|\nabla g(w)\|^2 \le 2Lg(w)\ .$$

Using the definition of g, we obtain

$$\begin{aligned} &\|\nabla\phi(w) - \nabla\phi(u)\|^2 \\ &= \|\nabla g(w)\|^2 \le 2Lg(w) \\ &= 2L\left[\phi(w) - \phi(u) - \nabla\phi(u)^\top(w-u)\right]\ . \end{aligned}$$

■

[1] While Theorem 3 bounds the expected sub-optimality, by techniques similar to (Shalev-Shwartz & Zhang, 2015) it can be converted to a bound that holds with high probability.

Lemma 4 *For $a \in (0, 1/2)$ we have $-\log(1-a)/a \le 1.4$.*

Proof Denote $g(a) = -\log(1-a)/a$. It is easy to verify that the derivative of g in $(0, 1/2)$ is positive and that $g(0.5) \le 1.4$. The proof follows. ■

5. Summary

We have described and analyzed a dual free version of SDCA that supports non-regularized objectives and non-convex individual loss functions. Our analysis shows a linear rate of convergence for all of these cases. Two immediate open questions are whether the worse dependence on the condition number for the non-accelerated result for the non-convex case is necessary, and whether the factor $n^{3/4}$ in Theorem 4 can be reduced to $n^{1/2}$.

Acknowledgements

In a previous draft of this paper, the bound for the non-convex case was $n^{5/4} + n^{3/4}\sqrt{\bar{L}/\lambda}$. We thank Ohad Shamir for showing us how to derive the improved bound of $n + n^{3/4}\sqrt{\bar{L}/\lambda}$. The work is supported by ICRI-CI and by the European Research Council (TheoryDL project).

References

Agarwal, Alekh and Bottou, Leon. A lower bound for the optimization of finite sums. In *ICML*, 2014.

Allen-Zhu, Zeyuan and Yuan, Yang. Univr: A universal variance reduction framework for proximal stochastic gradient method. *arXiv preprint arXiv:1506.01972*, 2015.

Arjevani, Yossi, Shalev-Shwartz, Shai, and Shamir, Ohad. On lower and upper bounds for smooth and strongly convex optimization problems. *arXiv preprint arXiv:1503.06833*, 2015.

Csiba, Dominik and Richtárik, Peter. Primal method for erm with flexible mini-batching schemes and non-convex losses. *arXiv preprint arXiv:1506.02227*, 2015.

Defazio, Aaron. *New Optimisation Methods for Machine Learning*. PhD thesis, Australian National Univer- sity, 2014.

Defazio, Aaron, Bach, Francis, and Lacoste-Julien, Simon. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014a.

Defazio, Aaron J, Caetano, Tibério S, and Domke, Justin. Finito: A faster, permutable incremental gradient method for big data problems. *arXiv preprint arXiv:1407.2710*, 2014b.

He, Xi and Takáč, Martin. Dual free sdca for empirical risk minimization with adaptive probabilities. *arXiv preprint arXiv:1510.06684*, 2015.

Jin, Chi, Kakade, Sham M, Musco, Cameron, Netrapalli, Praneeth, and Sidford, Aaron. Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation. *arXiv preprint arXiv:1510.08896*, 2015.

Johnson, Rie and Zhang, Tong. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

Konečnỳ, Jakub and Richtárik, Peter. Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*, 2013.

Le Roux, Nicolas, Schmidt, Mark, and Bach, Francis. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

Lin, Hongzhou, Mairal, Julien, and Harchaoui, Zaid. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pp. 3366–3374, 2015.

Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming SERIES A and B (to appear)*, 2015.

Shalev-Shwartz, Shai. Sdca without duality. *arXiv preprint arXiv:1502.06177*, 2015.

Shalev-Shwartz, Shai and Ben-David, Shai. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

Shalev-Shwartz, Shai and Zhang, Tong. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599, Feb 2013.

Shamir, Ohad. A stochastic pca and svd algorithm with an exponential convergence rate. In *ICML*, 2015.

Xiao, Lin and Zhang, Tong. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.