Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization

Shai Shalev-Shwartz
School of Computer Science and Engineering
Hebrew University, Jerusalem, Israel

Tong Zhang
Department of Statistics
Rutgers University, NJ, USA

Abstract

Stochastic Gradient Descent (SGD) has become popular for solving large scale supervised machine learning optimization problems such as SVM, due to their strong theoretical guarantees. While the closely related Dual Coordinate Ascent (DCA) method has been implemented in various software packages, it has so far lacked good convergence analysis. We present a new analysis of Stochastic Dual Coordinate Ascent (SDCA) showing that this class of methods enjoy strong theoretical guarantees that are comparable or better than SGD. This analysis justifies the effectiveness of SDCA for practical applications. A long version of this paper is available in [15].

1 Introduction

We consider the following generic optimization problem associated with regularized loss minimization of linear predictors: Let $x_1, \ldots, x_n$ be vectors in $\mathbb{R}^d$, let $\phi_1, \ldots, \phi_n$ be a sequence of scalar convex functions, and let $\lambda > 0$ be a regularization parameter. Our goal is to solve

$$
P(w) = \left[ \frac{1}{n} \sum_{i=1}^n \phi_i(w^T x_i) + \frac{\lambda}{2} \|w\|^2 \right].
$$

For example, given labels $y_1, \ldots, y_n$ in $\{\pm 1\}$, the SVM problem (with linear kernels and no bias term) is obtained by setting $\phi_i(a) = \max\{0, 1 - y_i a\}$. Regularized logistic regression is obtained by setting $\phi_i(a) = \log(1 + \exp(-y_i a))$. Regression problems also fall into the above. For example, ridge regression is obtained by setting $\phi_i(a) = (a - y_i)^2$, regression with the absolute-value is obtained by setting $\phi_i(a) = |a - y_i|$, and support vector regression is obtained by setting $\phi_i(a) = \max\{0, |a - y_i| - \nu\}$, for some predefined insensitivity parameter $\nu > 0$.

Let $w^*$ be the optimum of (1). We say that a solution $w$ is $\epsilon_P$-sub-optimal if $P(w) - P(w^*) \leq \epsilon_P$.

We analyze the runtime of optimization procedures as a function of the time required to find an $\epsilon_P$-sub-optimal solution.

A simple approach for solving SVM is stochastic gradient descent (SGD) [16, 13, 1]. SGD finds an $\epsilon_P$-sub-optimal solution in time $O(1/(\lambda \epsilon_P))$. This runtime does not depend on $n$ and therefore is favorable when $n$ is very large. However, the SGD approach has several disadvantages. It does not have a clear stopping criterion; It tends to be too aggressive at the beginning of the optimization process, especially when $\lambda$ is very small; While SGD reaches a moderate accuracy quite fast, it’s convergence becomes rather slow when we are interested in more accurate solutions.

An alternative approach is dual coordinate ascent (DCA), which solves a dual problem of (1). Specifically, for each $i$ let $\phi_i^*: \mathbb{R} \rightarrow \mathbb{R}$ be the convex conjugate of $\phi_i$, namely, $\phi_i^*(u) = \sup_{a \in \mathbb{R}} (u a - \phi_i(a))$. Dual coordinate ascent updates $w$ and $y$ simultaneously, using the subgradient of $\phi_i^*$ w.r.t. $u$.

We present a new analysis of Stochastic Dual Coordinate Ascent (SDCA) showing that this class of methods enjoy strong theoretical guarantees that are comparable or better than SGD. This analysis justifies the effectiveness of SDCA for practical applications. A long version of this paper is available in [15].
max_z(zu - φ_i(z)). The dual problem is

\[
\max_{\alpha \in \mathbb{R}^m} D(\alpha) \quad \text{where} \quad D(\alpha) = \left[ \frac{1}{n} \sum_{i=1}^{n} -\phi_i^*(\alpha_i) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i x_i \right\|^2 \right].
\]

(2)

The dual objective in (2) has a different dual variable associated with each example in the training set. At each iteration of DCA, the dual objective is optimized with respect to a single dual variable, while the rest of the dual variables are kept in tact.

If we define

\[
w(\alpha) = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i x_i,
\]

then it is known that \(w(\alpha^*) = w^*\), where \(\alpha^*\) is an optimal solution of (2). It is also known that \(P(w^*) = D(\alpha^*)\) which immediately implies that for all \(w\) and \(\alpha\), we have \(P(w) \geq D(\alpha)\), and hence the duality gap defined as

\[P(w(\alpha)) - D(\alpha)\]

can be regarded as an upper bound of the primal sub-optimality \(P(w(\alpha)) - P(w^*)\).

We focus on a stochastic version of DCA, abbreviated by SDCA, in which at each round we choose which dual coordinate to optimize uniformly at random. The purpose of this paper is to develop theoretical understanding of the convergence of the duality gap for SDCA.

We analyze SDCA either for \(L\)-Lipschitz loss functions or for \((1/\gamma)\)-smooth loss functions, which are defined as follows.

**Definition 1.** A function \(\phi_i : \mathbb{R} \rightarrow \mathbb{R}\) is \(L\)-Lipschitz if for all \(a, b \in \mathbb{R}\), we have

\[|\phi_i(a) - \phi_i(b)| \leq L |a - b|.
\]

A function \(\phi_i : \mathbb{R} \rightarrow \mathbb{R}\) is \((1/\gamma)\)-smooth if it is differentiable and its derivative is \((1/\gamma)\)-Lipschitz. An equivalent condition is that for all \(a, b \in \mathbb{R}\), we have

\[\phi_i(a) \leq \phi_i(b) + \phi_i'(b)(a - b) + \frac{1}{2\gamma}(a - b)^2.
\]

It is well-known that if \(\phi_i(a)\) is \((1/\gamma)\)-smooth, then \(\phi_i^*(u)\) is \(\gamma\) strongly convex: for all \(u, v \in \mathbb{R}\) and \(s \in [0, 1]::

\[-\phi_i^*(su + (1 - s)v) \geq -s\phi_i^*(u) - (1 - s)\phi_i^*(v) + \frac{\gamma s(1 - s)}{2}(u - v)^2.
\]

Our main findings are: in order to achieve a duality gap of \(\epsilon\),

- For \(L\)-Lipschitz loss functions, we obtain the rate of \(\hat{O}(n + L^2/(\lambda\epsilon))\).
- For \((1/\gamma)\)-smooth loss functions, we obtain the rate of \(\hat{O}((n + 1/(\lambda\gamma))\log(1/\epsilon))\).
- For loss functions which are almost everywhere smooth (such as the hinge-loss), we can obtain rate better than the above rate for Lipschitz loss. A precise statement is given in the long version.

## 2 Related Work

DCA methods are related to decomposition methods [12, 5]. While several experiments have shown that decomposition methods are inferior to SGD for large scale SVM [13, 6], Hsieh et al. [3] recently argued that SDCA outperform the SGD approach in some regimes. For example, this occurs when we need relatively high solution accuracy so that either SGD or SDCA has to be run for more than a few passes over the data.

However, our theoretical understanding of SDCA is not satisfying. Several authors (e.g. [10, 3]) proved a linear convergence rate for solving SVM with DCA (not necessarily stochastic). The basic technique is to adapt the linear convergence of coordinate ascent that was established by Luo and
losses, assuming that $n \sim 1$. Another relevant approach is the Stochastic Average Gradient (SAG), that has recently been an-
is the same as the rate we derive for SDCA with a Lipschitz loss function. 

yields the SDCA algorithm for the hinge-loss. The rate of convergence [7] derived for their algorithm 
specifying one variant of their algorithm to binary classification with the hinge loss, 
recently, [7] derived a stochastic coordinate ascent for structural SVM based on the Frank-Wolfe 
behavior of cyclic dual coordinate ascent, is inferior to our analysis. 

In this paper we derive new bounds on the duality gap (hence, they also imply bounds on the primal 
sub-optimality) of SDCA. These bounds are superior to earlier results, and our analysis only holds 
for randomized (stochastic) dual coordinate ascent. As we will see from our experiments, random-
ized methods with similar results. This means that their analysis, which can be no better than the 
results to the dual SVM formulation. However, the resulting convergence rate is 

Quite different. Consequently their results are not directly comparable to results we obtain in this 

closest, then $\nu \to 0$. This dependency is problematic in the data laden domain, and we note that such a 
dependency does not occur in the analysis of SGD.

Second, the analysis only deals with the sub-optimality of the dual objective, while our real goal is 
to bound the sub-optimality of the primal objective. Given a dual solution $\alpha \in \mathbb{R}^n$ its corresponding 
primal solution is $w(\alpha)$ (see (3)). The problem is that even if $\alpha$ is $\epsilon_D$-sub-optimal in the dual, for 
some small $\epsilon_D$, the primal solution $w(\alpha)$ might be far from being optimal. For SVM, [4, Theorem 
2] showed that in order to obtain a primal $\epsilon_D$-sub-optimal solution, we need a dual $\epsilon_D$-sub-optimal 
solution with $\epsilon_D = O(\lambda^2 \gamma^2)$; therefore a convergence result for dual solution can only translate into 
a primal convergence result with worse convergence rate. Such a treatment is unsatisfactory, and 
this is what we will avoid in the current paper.

Some analyses of stochastic coordinate ascent provide solutions to the first problem mentioned 
avove. For example, Collins et al [2] analyzed an exponentiated gradient dual coordinate ascent 
algorithm for SVM and logistic regression. The algorithm analyzed there (exponentiated gradient) 
is different from the standard DCA algorithm which we consider here, and the proof techniques are 
quite different. Consequently their results are not directly comparable to results we obtain in this 
paper. Nevertheless we note that for SVM, their analysis shows a convergence rate of $O(n/\epsilon_D)$ 
in order to achieve $\epsilon_D$-sub-optimality on the dual while our analysis shows a convergence of 
$O(n \log \log n + 1/\lambda \epsilon)$ to achieve $\epsilon$ duality gap; for logistic regression, their analysis shows a convergence 
rate of $O((n + 1/\lambda) \log(1/\epsilon_D))$ in order to achieve $\epsilon_D$-sub-optimality on the dual while 
our analysis shows a convergence of $O((n + 1/\lambda) \log(1/\epsilon))$ to achieve $\epsilon$ duality gap.

In addition, [14], and later [11] have analyzed randomized versions of coordinate descent for uncon-
strained and constrained minimization of smooth convex functions. [3, Theorem 4] applied these 
results to the dual SVM formulation. However, the resulting convergence rate is $O(n/\epsilon_D)$ which is, 
as mentioned before, inferior to the results we obtain here. Furthermore, neither of these analyses 
can be applied to logistic regression due to their reliance on the smoothness of the dual objective 
function which is not satisfied for the dual formulation of logistic regression. We shall also point 
out again that all of these bounds are for the dual sub-optimality, while as mentioned before, we are 
interested in the primal sub-optimality.

In this paper we derive new bounds on the duality gap (hence, they also imply bounds on the primal 
sub-optimality) of SDCA. These bounds are superior to earlier results, and our analysis only holds 
for randomized (stochastic) dual coordinate ascent. As we will see from our experiments, random-
ization is important in practice. In fact, the practical convergence behavior of (non-stochastic) cyclic 
dual coordinate ascent (even with a random ordering of the data) can be slower than our theoretical 
bounds for SDCA, and thus cyclic DCA is inferior to SDCA. In this regard, we note that some of 
the earlier analysis such as [9] can be applied both to stochastic and to cyclic dual coordinate as-
cent methods with similar results. This means that their analysis, which can be no better than the 
behavior of cyclic dual coordinate ascent, is inferior to our analysis.

Recently, [7] derived a stochastic coordinate ascent for structural SVM based on the Frank-Wolfe 
algorithm. Specifying one variant of their algorithm to binary classification with the hinge loss, 
yields the SDCA algorithm for the hinge-loss. The rate of convergence [7] derived for their algorithm 
is the same as the rate we derive for SDCA with a Lipschitz loss function.

Another relevant approach is the Stochastic Average Gradient (SAG), that has recently been an-
alysed in [8]. There, a convergence rate of $O(n \log(1/\epsilon))$ rate is shown, for the case of smooth 
losses, assuming that $n \geq \frac{8}{\lambda^2}$. This matches our guarantee in the regime $n \geq \frac{8}{\lambda^2}$. 

3
The following table summarizes our results in comparison to previous analyses. Note that for SDCA with Lipschitz loss, we observe a faster practical convergence rate, which is explained with our refined analysis in the long version of this paper.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>type of convergence</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipschitz loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGD</td>
<td>primal</td>
<td>$O\left(\frac{1}{\lambda}\epsilon\right)$</td>
</tr>
<tr>
<td>online EG [2] (for SVM)</td>
<td>dual</td>
<td>$\tilde{O}\left(\frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>SDCA</td>
<td>primal-dual</td>
<td>$\tilde{O}(n + \frac{1}{\lambda})$ or faster</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>type of convergence</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smooth loss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGD</td>
<td>primal</td>
<td>$O\left(\frac{1}{\lambda}\epsilon\right)$</td>
</tr>
<tr>
<td>online EG [2] (for logistic regression)</td>
<td>dual</td>
<td>$\tilde{O}\left((n + \frac{1}{\lambda}) \log \frac{1}{\epsilon}\right)$</td>
</tr>
<tr>
<td>SDCA</td>
<td>primal-dual</td>
<td>$\tilde{O}\left((n + \frac{1}{\lambda}) \log \frac{1}{\epsilon}\right)$</td>
</tr>
</tbody>
</table>

Acknowledgments

The work of Shai Shalev-Shwartz is supported by the Israeli Science Foundation grant number 598-10 and by the John S. Cohen Senior Lectureship in Computer Science.

References

