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 $\min_{w \in \mathbb{R}^d} P(w)$ where

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

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 ϵ_P-sub-optimal if $P(w) - P(w^*) \le \epsilon_P$. We analyze the runtime of optimization procedures as a function of the time required to find an ϵ_P-sub-optimal solution.

A simple approach for solving SVM is stochastic gradient descent (SGD) [16, 13, 1]. SGD finds an ϵ_P-sub-optimal solution in time $\tilde{O}(1/(\lambda \epsilon_P))$. 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 λ 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 $\phi_i^* : \mathbb{R} \to \mathbb{R}$ be the convex conjugate of ϕ_i, namely, $\phi_i^*(u) =$

$\max_z(zu - \phi_i(z))$. The dual problem is

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

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

$$w(\alpha) = \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i x_i, \tag{3}$$

then it is known that $w(\alpha^*) = w^*$, where α^* is an optimal solution of (2). It is also known that $P(w^*) = D(\alpha^*)$ which immediately implies that for all w and α, 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 $\phi_i : \mathbb{R} \to \mathbb{R}$ 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 $\phi_i : \mathbb{R} \to \mathbb{R}$ 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 γ 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 ϵ,

- For L-Lipschitz loss functions, we obtain the rate of $\tilde{O}(n + L^2/(\lambda\epsilon))$.
- For $(1/\gamma)$-smooth loss functions, we obtain the rate of $\tilde{O}((n + 1/(\lambda\gamma)) \log(1/\epsilon))$.
- 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

Tseng [9]. The linear convergence means that it achieves a rate of $(1-\nu)^k$ after k passes over the data, where $\nu > 0$. This convergence result tells us that after an unspecified number of iterations, the algorithm converges faster to the optimal solution than SGD.

However, there are two problems with this analysis. First, the linear convergence parameter, ν, may be very close to zero and the initial unspecified number of iterations might be very large. In fact, while the result of [9] does not explicitly specify ν, an examine of their proof shows that ν is proportional to the smallest nonzero eigenvalue of $X^\top X$, where X is the $n \times d$ data matrix with its i-th row be the i-th data point x_i. For example if two data points $x_i \neq x_j$ becomes closer and closer, 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 α is ϵ_D-sub-optimal in the dual, for some small ϵ_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 ϵ_P-sub-optimal solution, we need a dual ϵ_D-sub-optimal solution with $\epsilon_D = O(\lambda\epsilon_P^2)$; 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 above. 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 ϵ_D-sub-optimality (on the dual) while our analysis shows a convergence of $O(n \log\log n + 1/\lambda\epsilon)$ to achieve ϵ duality gap; for logistic regression, their analysis shows a convergence rate of $O((n + 1/\lambda)\log(1/\epsilon_D))$ in order to achieve ϵ_D-sub-optimality on the dual while our analysis shows a convergence of $O((n + 1/\lambda)\log(1/\epsilon))$ to achieve ϵ duality gap.

In addition, [14], and later [11] have analyzed randomized versions of coordinate descent for unconstrained 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, randomization 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 ascent 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 analyzed in [8]. There, a convergence rate of $\tilde{O}(n\log(1/\epsilon))$ rate is shown, for the case of smooth losses, assuming that $n \geq \frac{8}{\lambda\gamma}$. This matches our guarantee in the regime $n \geq \frac{8}{\lambda\gamma}$.

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.

Lipschitz loss

Algorithm	type of convergence	rate
SGD	primal	$\tilde{O}(\frac{1}{\lambda\epsilon})$
online EG [2] (for SVM)	dual	$\tilde{O}(\frac{n}{\epsilon})$
SDCA	primal-dual	$\tilde{O}(n + \frac{1}{\lambda\epsilon})$ or faster

Smooth loss

Algorithm	type of convergence	rate
SGD	primal	$\tilde{O}(\frac{1}{\lambda\epsilon})$
online EG [2] (for logistic regression)	dual	$\tilde{O}((n + \frac{1}{\lambda}) \log \frac{1}{\epsilon})$
SDCA	primal-dual	$\tilde{O}((n + \frac{1}{\lambda}) \log \frac{1}{\epsilon})$

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

[1] L. Bottou. Stochastic gradient descent examples, Web Page. http://leon.bottou.org/projects/sgd.

[2] M. Collins, A. Globerson, T. Koo, X. Carreras, and P. Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin markov networks. *Journal of Machine Learning Research*, 9:1775–1822, 2008.

[3] C.J. Hsieh, K.W. Chang, C.J. Lin, S.S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *ICML*, pages 408–415, 2008.

[4] D. Hush, P. Kelly, C. Scovel, and I. Steinwart. QP algorithms with guaranteed accuracy and run time for support vector machines. *JMLR*, 7:733–769, 2006.

[5] T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[6] T. Joachims. Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 216–226, 2006.

[7] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Stochastic block-coordinate frank-wolfe optimization for structural svms. *arXiv preprint arXiv:1207.4747*, 2012.

[8] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A Stochastic Gradient Method with an Exponential Convergence Rate for Strongly-Convex Optimization with Finite Training Sets. *arXiv preprint arXiv:1202.6258*, 2012.

[9] Z.Q. Luo and P. Tseng. On the convergence of coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.*, 72:7–35, 1992.

[10] O. Mangasarian and D. Musicant. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10, 1999.

[11] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[12] J. C. Platt. Fast training of Support Vector Machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[13] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *INTERNATIONAL CONFERENCE ON MACHINE LEARNING*, pages 807–814, 2007.

[14] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l_1 regularized loss minimization. In *ICML*, page 117, 2009.

[15] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *arXiv preprint arXiv:1209.1873*, 2012.

[16] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.