Linear Convergence and Support Vector Identifiation of Sequential Minimal Optimization

Jennifer She
University of British Columbia
Mark Schmidt
University of British Columbia

Abstract

We analyze variants of the classic 2-coordinate sequential minimal optimization (SMO) algorithm for training support vector machines (SVMs) with an unregularized bias. We show these variants achieve a linear convergence rate, improving on previous rates, which were sublinear unless strong-convexity was assumed. We also show that they identify the final set of support vectors in a finite number of iterations, which is the first time a manifold identification property has been shown for any coordinate descent method with non-separable constraints.

1 Motivation

SVMs remain widely-used in many applications, and the SMO dual 2-coordinate ascent method [5] has been a popular method for fitting these models for around 20 years. Unlike modern stochastic dual coordinate ascent (SDCA) methods, SMO allows an unregularized bias which is desirable in applications with unbalanced class labels. This bias complicates the analysis of SDCA methods because it leads to a linear equality constraint across the variables. While previous analyses of SMO show that it converges with a sublinear rate in general, this work shows that this type of algorithm can achieve a linear convergence rate as in SCDA methods. Further, under mild assumptions we show that these variants identify the support vectors in a finite number of iterations.

2 Problem Definition

The general problem we consider is

$$\operatorname*{arg\,min}_{x \in \mathcal{X}} f(x), \tag{1}$$

where $\mathcal{X}$ is a set of the form $\{x \mid l \leq x \leq u, Ax = b\}$ with l and u being upper and lower bounds on the variables, and with the $m \times n$ matrix A and $m \times 1$ vector b defining linear constraints such that $m \leq n$. We assume that the gradient ∇f is Lipschitz continuous, and we require that the problem satisfies the proximal-PL (prox-PL) inequality [1] which is this case can be written as

$$\frac{1}{2}\mathcal{D}_g(x, L) \geq \mu(f(x) - f^*), \tag{2}$$

for all $x \in \mathcal{X}$, for some $\mu > 0$ and with $\mathcal{D}_g(x, \mu) \equiv -2\mu \operatorname{arg\,min}_{y \in \mathcal{X}}\{\langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2\}]$. The dual of the SVM problem with an unregularized bias takes the form (1) for a convex f with a Lipschitz-continuous gradient. Further, Necoara et al. [3] show that this dual satisfies the quadratic growth (QG) property, which combined with convexity implies that the proximal-PL inequality is satisfied [1].

3 Block Coordinate Descent for Linearly Coupled Constraints

Block coordinate descent methods are iterative optimization algorithms, that on each iteration choose a block of variables among a candidate set of blocks B, and then updates this block of variables. Without the linear equality constraints, the constraints would affect each variable individually and linear convergence of block coordinate descent for problem (1) would be implied by the prior work [1]. However, the equality constraints introduce dependencies between all variables that must be maintained during the updates. Our

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

analysis will follow the analysis of Necoara and Patrascu [2] closely, and as with prior work we maintain the equality constraints by constructing the blocks so that the supports of all so-called "elementary vectors" of the null-space of A are contained in the set of candidate blocks B [6], [2]. We first introduce the notion of an elementary vector.

Definition 1. *Let $d, d' \in \mathbb{R}^n$. Then d' is conformal to d if*

$$\mathrm{supp}(d') \subseteq \mathrm{supp}(d) \textit{ and } d'_j d_j \geq 0,\, \forall j = 1, ..., n.$$

Definition 2. *For $A \in \mathbb{R}^{m \times n}$ and $m \leq n$, an elementary vector of $\mathrm{null}(A)$, is a vector $d \in \mathrm{null}(A)$ such that*

$$\forall d' \in \mathrm{null}(A) \textit{ that are conformal to } d,\ \mathrm{supp}(d') = \mathrm{supp}(d).$$

Below we give two properties used in prior works that will prove useful [6], [2].

Lemma 1. *If d is an elementary vector of $\mathrm{null}(A)$, then $|\mathrm{supp}(d)| \leq \mathrm{rank}(A) + 1 \leq m + 1$. Otherwise, d has a conformal realization*

$$d = \textstyle\sum_{j=1}^{r} d_j$$

where $r \geq 1$ and d_j are elementary vectors conformal to d, $j = 1, ..., r$.

Lemma 2. *Let the conformal realization of d be as defined above, then for any coordinate-wise separable and convex function h,*

$$\textstyle\sum_{j=1}^{r} (h(x + d_j) - h(x)) \leq h(x - d) - h(x).$$

For the SVM dual objective (with unregularized bias) Lemma 1 implies that we can use all pairs of variables as the blocks, $B = \{\{i, j\} | i, j \in \{1, 2, ..., n\}, i \neq j\}$, as used by SMO. Under other assumptions, Tseng & Yun have previously shown an asymptotic linear convergence rate when the block to update is chosen according to the greedy Gauss-Southwell-q (GS-q) rule [6]. More recently, Necoara & Patrascu show a non-asymptotic linear convergence rate for randomized block selection when the proximal-PL condition is replaced by the stronger condition of strong-convexity [2], but in general the SVM problem is not strongly-convex. In the next section, we give a simple proof showing that a non-asymptotic linear convergence rate holds under the weaker assumption of the proximal-PL inequality.

3.1 Block Selection and Updates

We focus on the case where we choose the block to update b^k uniformly at random from the set of possible blocks B, but we note that analyzing the greedy GS-q rule only involves a small change. The classic SMO method picks coordinates based on violation of the Karush-Kuhn-Tucker conditions, and makes use of several heuristic that makes analyzing it more difficult. Given the block $b^k \in B$ that we choose on iteration k, we'll consider updates of the form

$$x^{k+1} = \underset{\{y_{b^k} \,\mid\, y \in \mathcal{X}\}}{\arg\min} \left\{ f(x^k) + \langle \nabla f(x^k), y - x^k \rangle + \frac{L}{2} \|y - x^k\|^2_{H^k} \right\}. \tag{3}$$

Here, H^k is a matrix satisfying the generalized inequality $\nabla^2 f(x^k) \preceq H^k \preceq L\mathbb{I}$. In the SMO method (where the objective is quadratic) we set $H^k = \nabla^2 f(x^k)$ which corresponds to a projected-Newton update, while choosing $H^k = L\mathbb{I}$ leads to a projected-gradient update of the chosen block.

4 Convergence Analysis

Our argument closely follows the analysis of Necoara and Patrascu [2]. First we note that it follows from Lipschitz-continuity of the gradient and the well-known descent lemma that

$$\mathbb{E}[f(x^{k+1})] \leq \mathbb{E}[f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2} \|x^{k+1} - x^k\|^2_{H_k}] \tag{4}$$

$$= f(x^k) + \frac{1}{|B|} \sum_{i=1}^{|B|} \left[\min_{\{y_{b_i} \,\mid\, y \in \mathcal{X}\}} \left\{ \langle \nabla f(x^k), y - x^k \rangle + \frac{1}{2} ||y - x^k||^2_{H_k} \right\} \right]. \tag{5}$$

Reparameterizing with $d_{b_i} = y_{b_i} - x^k_{b_i}$ and using the property $H^k \preceq L\mathbb{I}$, we get

$$(5) \leq f(x^k) + \frac{1}{|B|}\sum_{i=1}^{|B|}\left[\min_{\{d_{b_i} \mid x^k_{b_i}+d_{b_i}\in\mathcal{X}\}} \{\langle \nabla f_{b_i}(x^k), d_{b_i}\rangle + \frac{L}{2}\|d_{b_i}\|^2\}\right]. \tag{6}$$

Next, we define specific d'_{b_j} values where $j = i_1, i_2, ..., i_r$ and $r \leq |B|$ as follows. Let

$$d' = \operatorname*{arg\,min}_{\{d \mid x^k+d\in\mathcal{X}\}} \left\{\langle \nabla f(x^k), d\rangle + \frac{L}{2}\|d\|^2\right\}. \tag{7}$$

If d is an elementary vector, then d is in some block b_{i_1}, and we can let $r = 1$ and choose $d'_{b_{i_1}} = d'$. Otherwise, we can choose specific d'_{b_j} values to be the elementary vectors that make up a conformal realization of d' which exists by Lemma 1. If the blocks are not minimal, we additionally combine the elementary vectors with supports in the same block.
In the first case, $x^k + d'_{b_{i_1}} \in \mathcal{X}$ holds by default. In the second case, d'_{b_j} being elementary vectors of $\mathrm{null}(A)$ or summations of them implies $d'_{b_j} \in \mathrm{null}(A)$. In addition $l - x \leq d' \leq u - x$ implies $l - x \leq d'_{b_j} \leq u - x$ due to conformality. Thus, for all $j = i_1, i_2, ..., i_r$, d'_{b_j} satisfies $x^k + d'_{b_j} \in \mathcal{X}$, and using the property that the minimum is less than or equal to a specific instance, we get

$$(6) \leq f(x^k) + \frac{1}{|B|}\left[\langle \nabla f(x^k), \sum_{t=1}^{r} d'_{b_j}\rangle + \sum_{t=1}^{r}\frac{L}{2}\|d'_{b_j}\|^2\right]. \tag{8}$$

Then by Lemma 2, we get

$$(8) \leq f(x^k) + \frac{1}{|B|}\min_{\{d \mid x^k+d\in\mathcal{X}\}} \left\{\langle \nabla f(x^k), d\rangle + \frac{L}{2}\|d\|^2\right\} \tag{9}$$

$$= f(x^k) + \frac{1}{|B|}\min_{y\in\mathcal{X}} \left\{\langle \nabla f(x^k), y - x^k\rangle + \frac{L}{2}\|y - x^k\|^2\right\} \tag{10}$$

$$\leq f(x^k) - \frac{1}{|B|}\frac{\mu}{L}(f(x^k) - f^*), \tag{11}$$

where in the last line we apply the proximal-PL inequality. We can subtract f^* from both sides and rearrange the terms, and apply this recursively to get the linear convergence result that

$$\mathbb{E}[f(x^k)] - f^* \leq \left(1 - \frac{\mu}{|B|L}\right)^k (f(x^0) - f^*), \tag{12}$$

5 Sequential Minimal Optimization

Next, we explain the explicit updates for a given block $b \in B$ for SMO using the notations above. Recall that the blocks are all pairs of coordinates $\{i, j\}$. Without the inequality constraints in (3), we get the minimizers

$$y_i^* = x_i - \frac{1}{H^k_{ii} + H^k_{jj} \pm 2H^k_{ij}}[\nabla_i f(x) \pm \nabla_j f(x)] \tag{13}$$

$$y_j^* = x_j - \frac{1}{H^k_{jj} + H^k_{ii} \pm 2H^k_{ij}}[\nabla_j f(x) \pm \nabla_i f(x)] \tag{14}$$

where $H^k_{ij} = \nabla^2_{ij} f(x^k)$ and $\pm$ is the sign of $-\alpha_i \cdot \alpha_j$, where α_i for $i = 1, 2, ..., n$ are constants $\in \{-1, 1\}$. We can add back the inequality constraints by clipping the minimizers due to convexity. The resulting updates are

$$x_i^{k+1} = \begin{cases} L_i & y_i^* < L_i \\ y_i^* & L_i \leq y_i^* \leq U_i \\ U_i & y_i^* > U_i \end{cases} \qquad x_j^{k+1} = \begin{cases} L_j & y_j^* < L_j \\ y_j^* & L_j \leq y_j^* \leq U_j \\ U_j & y_j^* > U_j \end{cases} \tag{15}$$

where the bounds are

$$L_i = \begin{cases} \max\{0, x_i^k - (c - x_j^k)\} & \alpha_i = \alpha_j \\ \max\{0, x_i^k - x_j^k\} & \alpha_i \neq \alpha_j \end{cases} \qquad U_i = \begin{cases} \min\{c, x_i^k + x_j^k\} & \alpha_i = \alpha_j \\ \min\{c, x_i^k + (c - x_j^k)\} & \alpha_i \neq \alpha_j \end{cases} \tag{16}$$

for x_i^{k+1}, and the bounds for x_j^{k+1} have the indices i and j flipped. Under strong-convexity, we get

$$\nabla^2_{ii} f(x^k) + \nabla^2_{jj} f(x^k) \pm 2\nabla^2_{ij} f(x^k) > 0 \text{ for all } i, j = 1, ..., n, i \neq j$$

so the updates are well formed for $H^k = \nabla^2 f(x^k)$. Without strong-convexity, we can avoid the case where $H^k_{ii} + H^k_{jj} \pm 2H^k_{ij} = 0$ by using $H^k = L\mathbb{I}$ instead, in which case $H^k_{ii} + H^k_{jj} \pm 2H^k_{ij} = 2L$.

5.1 Support Vector Identification

We can also show that SMO can detect the support vectors in a finite number of iterations under mild conditions.The idea of this argument stems from Nutini et al. [4]'s work. We assume that the set of indices that are on the active set for any x^* in the solution set X^* is unique. Let this set of coordinates be Z^*. We introduce the notion of an active set below.

Definition 3. *The active set for SVMs for $x \in \mathcal{X}$ is the set*

$$\mathcal{Z} = \{i \mid x_i = 0 \textit{ or } c\}.$$

In addition, we assume that for all $x^* \in X^*$ and $i \in Z^*$, $\nabla_i f(x^*) \neq 0$, and $|\nabla_i f(x^*)| \neq |\nabla_j f(x^*)|$ for all $j = 1, ..., n$ such that $j \neq i$. Lastly, we assume that either $H^k = L\mathbb{I}$, or $H^k_{ij} = \nabla^2_{ij}(f(x^k)$ and for all $i \neq j$ that $H^k_{ii} + H^k_{jj} \pm 2H^k_{ij} > 0$. From the above analysis, and using the QG property, we get

$$\mathbb{E}[\|x^k - x^k_{\text{proj}}\|^2] \leq \frac{2}{\mu}\mathbb{E}[f(x^k) - f^*] \leq \frac{2}{\mu}\rho^k[f(x^0) - f^*] \tag{17}$$

where x^k_{proj} is the projection of x^k onto the solution set X^* and ρ is a constant, $0 \leq \rho < 1$. Thus, for all $\delta > 0$, there exists a finite number $\bar{k}$ such that $\|x^k - x^k_{\text{proj}}\| \leq \delta, \forall k > \bar{k}$ almost surely.
We choose $\bar{k}'$ and $\delta' = \min\{\zeta, \xi\}$, where

$$\zeta = \min_{\substack{i \in Z, j=1,...,n, \\ x^* \in X^*}} \left\{ \frac{|\nabla_i f(x^*) \pm \nabla_j f(x^*)|}{8L} \right\} \text{ and } \xi = \min_{\substack{j \notin Z^*, \\ x^* \in X^*}} \left\{ \frac{x^*_j}{2}, \frac{c - x^*_j}{2} \right\}.$$

Then we can prove the following.

Lemma 3. *Assume for $i \in Z^*$, $k' > \bar{k}'$, that j is on the active set on iteration k' for all $j \in Z^*$ where $|\nabla_j f(x^*)| > |\nabla_i f(x^*)|$. There exists a finite K', $K' > k'$ such that if we pick i on iteration K',*

$$x^k_i = 0 \textit{ or } c \textit{ for all } k > K'.$$

We omit the proof for space reasons, but the idea is that if i is picked with any $j \notin Z^*$, then i is guaranteed to move onto the active set. Otherwise, if $j \in Z^*$, then i is guaranteed to not move further away from it. We can use this lemma to show that after iteration $\bar{k}'$, there is a finite number of additional iterations until all support vectors are detected by applying it recursively, by induction on the decreasing order of $|\nabla_i f(x^*)|$, for $i \in Z^*$. For $i = \underset{i \in Z^*}{\text{argmax}}\{|\nabla_i f(x^*)|\}$, the assumption in this lemma holds by default, so the lemma applies. Then assuming that this lemma holds for all $j \in Z^*$ such that $|\nabla_j f(x^*)| > |\nabla_i f(x^*)|$ and applying this lemma again for i, this property holds for i and remains to hold for j. As a result, for all $i \in Z^*$, x_i gets set and kept at 0 or c after a finite number of steps almost surely, and we can pick out the indices i that are not on the active set.

6 Discussion

Although our motivation was providing a convergence analysis of the SMO algorithm, our convergence result is much more general and applies in a variety of other settings. For example, our analysis would apply to optimization with a probability simplex constraint. It would likely also cover standard formulations of support vector regression, multi-class SVMs, and margin-based ranking methods.

References

[1] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[2] I. Necoara and A. Patrascu. A random coordinate descent algorithm for optimization problems with composite objective function and linear coupled constraints. *Computational Optimization and Applications*, 57(2):307–337, 2014.

[3] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *arXiv preprint arXiv:1504.06298*, 2015.

[4] J. Nutini, M. Schmidt, and W. Hare. “active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *NIPS Optimization Workshop*, 2017.

[5] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

[6] P. Tseng and S. Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of Optimization Theory and Applications*, 140(3):513, 2009.