The Convergence Rate of Linearly Separable SMO

Jorge López and José R. Dorronsoro

Abstract—It is well known that the dual function value sequence generated by SMO has a linear convergence rate when the kernel matrix is positive definite and sublinear convergence is also known to hold for a general matrix. In this paper we will prove that, when applied to hard–margin, i.e., linearly separable SVM problems, a linear convergence rate holds for the SMO algorithm without any condition on the kernel matrix. Moreover, we will also show linear convergence for the multiplier sequence generated by SMO, the corresponding weight vectors and the KKT gap usually applied to control the number of SMO iterations. This gives a fairly complete picture of the convergence of the various sequences SMO generates. While linear SMO convergence for the general SVM $L_1$ soft margin problem is still open, the approach followed here may lead to such a general result.

I. INTRODUCTION

Given a sample $S = \{(x_i, y_i) : i = 1, \ldots, N\}$ with $y_i = \pm 1$, the standard formulation of SVM for linearly separable problems [1] wants to maximize the hard margin of a separating hyperplane $(W, b)$ by solving the problem

$$\min \frac{1}{2}\|W\|^2 \text{ with } y_i(W \cdot x_i + b) \geq 1, \ i = 1, \ldots, N. \ (1)$$

In practice, however, one solves the simpler dual problem of minimizing

$$\mathcal{D}(\alpha) = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j - \sum_i \alpha_i \ (2)$$

subject to $\alpha_i \geq 0$ for all $i$ and $\sum \alpha_i y_i = 0$. While the dual problem may admit several different solutions, the optimal primal weight $W^*$ is unique. This optimal weight $W^*$ can be then written as $W^* = \sum_i \alpha^*_i y_i x_i$ where $\alpha^*$ is any dual optimal multiplier set.

Most of the algorithms that solve (2) can be traced to Platt’s SMO [2], later refined by Keerthi et al.’s [3]. In particular their Modification 2 proposal is by now the SMO standard approach (see Section II for more details on SMO).

The convergence of SMO is a well studied problem. More precisely, SMO solves the dual problem (2) producing a (possibly not unique) sequence $\alpha^t$ of Lagrange multipliers. The most comprehensive convergence study has been that of Chih-Jen Lin and his coworkers who, in a series of papers, have proved the convergence of SMO for linear and non linearly separable problems under very general conditions. Specifically, asymptotic convergence of SMO when no assumptions are made on the positive definiteness of the kernel matrix is given in [4]. There it is shown that any converging subsequence $\alpha^j$ of the one SMO generates also converges to a dual optimum. As a consequence, convergence of the entire $\mathcal{D}(\alpha^j)$ sequence then follows, for we have $\mathcal{D}(\alpha^j) \to D^*$ and SMO ensures that the $\mathcal{D}(\alpha^j)$ decrease, that is, $\mathcal{D}(\alpha^{j+1}) < \mathcal{D}(\alpha^j)$. In [5] convergence to the single optimum $W^*$ is shown for the full weight sequence $W^t = \sum \alpha^t_i y_i x_i$ derived from any SMO–generated $\alpha^t$ multiplier sequence.

Concerning convergence rates, C.J. Lin and his coworkers have also shown ([4], [6], [7]) that, under some conditions, SMO converges linearly in the sense that there is a $\lambda < 1$ such that $D^{t+1} - D^* \leq \lambda (D^t - D^*)$. More recently, in [8] List and Simon proved sublinear convergence for the $D^t$ over any problem, that becomes linear under the same conditions considered by C.J. Lin et al (see also [9] for more results on convergence rates for SMO and its variants).

As far as we know, no linear convergence rates have been established for the sequence $D^t$ without the positive definiteness assumption. In this work we will show that even without this assumption, this is also true in the hard margin SVM case not only for the $D^t$ sequence but also for any SMO-generated weight sequence $W^t$ as well as for the sequence of the $\Delta^t$ gaps associated to the Karush–Kuhn–Tucker condition that are often used to decide when to stop SMO. Moreover, our arguments will also show that any SMO generated complete multiplier sequence $\alpha^t$ also converges to a dual optimal $\alpha^*$ and does so with a linear convergence rate.

We remark that our arguments are only valid for hard margin, i.e., linearly separable problems. On the other hand, they reveal the geometric working of the various SMO components and give some algorithmic insights on the overall behavior of SMO that could lead to a proof for the general case. In particular, the authors have shown [10] linear convergence for the Mitchell–Demiyanov–Malozemov (MDM) algorithm for the Nearest Point Problem (NPP) over reduced convex hulls along lines quite similar to the ones used here. The MDM algorithm and NPP problems are very much related to SMO and to the general, soft–margin SVM problem.

The paper is organized as follows. In Section II we will briefly review SMO and point to some facts on its convergence that will be used in Section III to give our linear convergence results. The paper ends with a short discussion.

II. THE SMO ALGORITHM AND ITS CONVERGENCE

SMO updates the current multiplier $\alpha^t$ changing just two values $\alpha_L$, $\alpha_U$. More precisely the new weight vector $W^{t+1}$ is given by
\[ W^{t+1} = W^t + \lambda^t y_{L^t} (X_{L^t} - X_{U^t}) = W^t + \lambda^ty_{L^t}Z_{U^t}, \]

for properly selected \(X_{L^t}, X_{U^t}\) and where we write \(Z_{ij} = X_i - X_j\). Writing \(W^t\) as \(W^t = \sum \alpha^t_i y_i X_p\), the weight update in terms of the \(\alpha\) coefficients becomes

\[
\alpha^{t+1}_L = \alpha^t_L + \lambda^t, \quad \alpha^{t+1}_U = \alpha^t_U - \lambda^ty_{U^t}y_{L^t}, \tag{3}
\]

and the others do not change. Forgetting for the time being about the box constraints on the \(\alpha\) and assuming \(L^t, U^t\) selected somehow, an optimal \(\lambda^t\) can be chosen so that the decrease in the SVM dual function is largest, i.e., so that it maximizes

\[
\Phi (\lambda^t) = \mathcal{D} (\alpha^t) - \mathcal{D} (\alpha^{t+1}) = -\lambda^t y_{L^t}W^t \cdot Z_{U^t} - \frac{1}{2} (\lambda^t)^2 \|Z_{U^t}\|^2 - y_{L^t}\lambda^t (y_{L^t} - y_{U^t}), \tag{4}
\]

and the optimal \(\lambda^t\) is given in principle by

\[
\lambda^t = \frac{y_{L^t}W^t \cdot Z_{U^t} - (y_{U^t} - y_{L^t})}{\|Z_{U^t}\|^2} = \frac{\Delta^t}{\|Z_{U^t}\|^2} \equiv y_{L^t}\mu^t, \tag{5}
\]

where we write now \(\Delta^t = W^t \cdot Z_{U^t} - (y_{U^t} - y_{L^t})\) and \(\mu^t = \frac{\Delta^t}{\|Z_{U^t}\|^2}\). If we disregard the denominator, the \(\mathcal{D}\) decrease is approximately largest when \(\Delta^t\) is largest, which can be achieved if we select

\[
I^t = \underset{i \in I_L}{\arg \min} \{W^t \cdot X_i - y_i\};
\]

\[
U^t = \underset{i \in I_U}{\arg \max} \{W^t \cdot X_i - y_i\}. \tag{6}
\]

Here we define the subsets \(I_L\) and \(I_U\) as

\[
I_L = \{i : y_i = 1 \text{ or } y_i = -1, \alpha^t_i > 0\};
\]

\[
I_U = \{i : y_i = 1, \alpha^t_i > 0 \text{ or } y_i = -1\}. \tag{7}
\]

Observe that the \(\alpha\) updates become now

\[
\alpha^{t+1}_L = \alpha^t_L + y_{L^t}\mu^t, \quad \alpha^{t+1}_U = \alpha^t_U - y_{U^t}\mu^t. \tag{8}
\]

Thus, \(\alpha^t_L\) and \(\alpha^t_U\) will decrease if \(y_{L^t} = -1\) or \(y_{U^t} = 1\), an the conditions in \(I_L\) and \(I_U\) ensure that they are > 0 to begin with.

Moreover, since we are solving a constrained problem, we may further have to clip \(\mu\) as \(\mu^t = \min\{\mu, \alpha^t_L\}\) if \(y_{L^t} = -1\) and as \(\mu^t = \min\{\mu, \alpha^t_U\}\) if \(y_{U^t} = 1\). If no such clipping is required, plugging (5) into (4) yields

\[
\mathcal{D} (\alpha^t) - \mathcal{D} (\alpha^{t+1}) = \frac{(\Delta^t)^2}{2 \|Z_{U^t}\|^2}. \tag{9}
\]

If, however, \(\mu\) is clipped at \(\mu^t\), \(\Delta^t \geq \alpha^t_L \|Z_{U^t}\|^2\) must hold and (4) yields

\[
\mathcal{D} (\alpha^t) - \mathcal{D} (\alpha^{t+1}) = \alpha^t_L \Delta^t \geq \frac{\alpha^t_L \Delta^t}{2}, \tag{10}
\]

while we similarly obtain

\[
\mathcal{D} (\alpha^t) - \mathcal{D} (\alpha^{t+1}) \geq \frac{\alpha^t_L \Delta^t}{2}, \tag{11}
\]

We will take advantage of this in our proof of a linear convergence rate for the sequence \(\alpha^t\) of SMO multipliers.

We turn now to discuss SMO convergence and to derive some facts to be used later but first we recall the KKT conditions for SVM. If \(W^*\) is the primal optimum and \(\alpha^*\) a dual optimum, we have \(W^* = \sum y_p\alpha^*_p X_p\) and also

\[
\alpha^*_p(W^* \cdot X_p + b^*) - 1 = 0 \tag{12}
\]

for all \(p\). In particular, \(\alpha^*_p > 0\) implies \(y_p(W^* \cdot X_p + b^*) = 1\) and, conversely, if \(y_p(W^* \cdot X_p + b^*) > 1\) we must have \(\alpha^*_p = 0\). Also, \(\sum \alpha^*_p = \|W^*\|^2\). Moreover we also have

\[
\min_{y_p=1} \{W^* \cdot X_p - y_p\} \leq -b^* = \max_{y_p=1, \alpha^*_p > 0} \{W^* \cdot X_q - y_q\};
\]

\[
\max_{y_p=-1} \{W^* \cdot X_q - y_q\} \leq -b^* = \min_{y_p=-1, \alpha^*_p > 0} \{W^* \cdot X_p - y_p\};
\]

which, if we define

\[
\Delta (\alpha) = \max \{W \cdot X_q - y_q : (y_q = 1, \alpha^*_q > 0) \wedge (y_q = -1)\} - \min \{W \cdot X_p - y_p : (y_p = 1) \wedge (y_p = -1, \alpha^*_p > 0)\}\]

implies for any dual optimal \(\alpha^*, \Delta (\alpha^*) < 0\) while, on the other hand, \(\Delta (\alpha) > 0\) if \(\alpha\) is not optimal. Notice that, by our previous discussion, at each SMO step \(t\) have precisely \(\Delta^t = \Delta (\alpha^t)\), where \(\alpha^t\) determines \(W^t = \sum \alpha^t_i y_i X_p\). Thus, SMO will advance as long as we do not arrive to an optimal \(\alpha^t\).
Going now to the convergence proof, the first step is the bound (see equation (13) in \cite{11})
\[ \|W^t - W^*\|^2 \leq 2(D^t - D^*) . \] (12)
To prove convergence of the $W^t$ it is thus enough to show that $D^t \to D^*$ or, since $D^t$ decreases, that $D^{t_j} \to D^*$ for some subsequence $t_j$, something that we do next.

First, we will argue in Theorem 2 below that there is no iteration $T$ after which only clipped iterations take place. Thus, there must be a subsequence $t_j$ for which clipping does not take place and, therefore, for which the estimate
\[ D^{t_j} - D^{t_j+1} \geq \frac{(\Delta t_j)^2}{2\|Z_{t_j,t_j'}\|^2} \]
holds. But since $D^{t_j}$ decreases, $D^{t_j} - D^{t_j+1} \to 0$ and, thus, $\Delta t_j \to 0$. Since the SMO multiplier sequence lies in a bounded and, hence, compact subset, there is another subsequence $t_{j_k}$ for which the corresponding multipliers $\alpha_{t_{j_k}}$ converge to an $\overline{\alpha}$ for which it can be shown that $\Delta(\overline{\alpha}) \leq 0$. Thus, $\overline{\alpha}$ is a dual optimum and $D(\overline{\alpha}) = D^*$. Therefore, $D^{t_{j_k}} \to D(\overline{\alpha}) = D^*$ and, in fact, we must then have $D^t \to D^*$ for the whole decreasing sequence $D^t$. The bound (12) now yields $W^t \to W^*$.

A well known consequence of the convergence of SMO, and a crucial fact in what follows, is that after some iteration, the only choices for $L, U$ correspond to points for which it must hold that $W^* \cdot X_p + b^* = \pm 1$, i.e., they are in the supporting hyperplanes. More precisely, let us introduce the following notation:

\[
\begin{align*}
\mathcal{O} &= \{ i : y_i(W^* \cdot X_i + b^*) > 1 \}, \\
\mathcal{H} &= \{ i : y_i(W^* \cdot X_i + b^*) = 1 \}.
\end{align*}
\]

Recall that if we arrive at an $\alpha_{t_q}^*$ such that $0 < \alpha_{t_q}^* < 1$, we have $q \in \mathcal{H}$. The just mentioned key ingredient for our linear rate result is the following (see \cite{12} and also \cite{7}).

**Proposition 1:** There is a $T_{\mathcal{H}}$ such that for all $t \geq T_{\mathcal{H}}$ we have $\alpha_{t_p}^* = 0$ for all $p \in \mathcal{O}$, and, moreover, $L^t \in \mathcal{H}$ and $U^t \in \mathcal{H}$.

One of its consequences is that for $t \geq T_{\mathcal{H}}$ it holds that
\[ \|W^t - W^*\|^2 = 2(D^t - D^*) \] (13)
for we have
\[
\begin{align*}
W^t \cdot W^* &= \sum_p \alpha_{t_p}^* y_p W^* \cdot X_p \\
&= \sum_p \alpha_{t_p}^* y_p (W^* \cdot X_p + b^*) \\
&= \sum_{\mathcal{H}} \alpha_{t_p}^*.
\end{align*}
\]

Thus, using the equality $\|W^*\|^2 = \sum \alpha_{t_p}^*$ due to the coincidence of the primal and dual optimum values, we have
\[
\begin{align*}
\|W^t - W^*\|^2 &= \|W^t\|^2 - 2W^t \cdot W^* + \|W^*\|^2 \\
&= \|W^t\|^2 - 2 \sum \alpha_{t_p}^* + 2 \alpha_{t_p}^* - \|W^*\|^2 \\
&= 2 \left( \frac{1}{2}\|W^t\|^2 - \sum \alpha_{t_p}^* - \frac{1}{2}\|W^*\|^2 - \sum \alpha_{t_p}^* \right) \\
&= 2(D^t - D^*).
\end{align*}
\]
which is (13). Therefore, the convergence rates for $W^t$ and $D^t$ are equivalent. In what follows we will use interchangeably the quantities $\|W^t - W^*\|^2$ and $2(D^t - D^*)$ without further mention and write $\\overline{W^t}$ for $W^t - W^*$ and $\overline{D^t}$ for $D^t - D^*$.

Another consequence of Proposition 1 is that $W^* \cdot X_{L_t} + b^* = y_{L_t}$ and $W^* \cdot X_{U_t} + b^* = y_{U_t}$ for all $t \geq T_{\mathcal{H}}$ and, therefore, we have the following formula for $\\Delta_t$:
\[
\\Delta_t = W_t \cdot (X_{U_t} - X_{L_t}) - y_{U_t} + b^* + y_{L_t} - b^* = W_t \cdot (X_{U_t} - X_{L_t}) - W^* \cdot X_{U_t} + W^* \cdot X_{L_t} = (W_t - W^*) \cdot (X_{U_t} - X_{L_t}) \] (15)

### III. Linear Convergence Rate

Our linear convergence rate arguments rely on the following key result.

**Theorem 1:** Assume $t \geq T_{\mathcal{H}}$ and let $\phi^t$ be the angle between $W^t - W^*$ and $Z^t = X_{U_t} - X_{L_t}$, i.e.,
\[ \Delta t = \|W^t - W^*\| \|Z^t\| \cos \phi^t. \] (16)

We then have $\lim \inf \cos \phi^t = \varphi > 0$.

We also need a technical result stating that unclipped iterations must happen with a certain regularity.

**Theorem 2:** There is an integer $L > 0$ such that for all $t \geq T_{\mathcal{H}}$ there is a $\tilde{t}$, with $t < \tilde{t} < t + L$, such that the $\tilde{t}$ iteration is unclipped.

While the ideas leading to Theorems 1 and 2 are quite simple we will defer their proofs to the end of this section for an easier reading. We derive now linear convergence rates for $W^t$ and $\\Delta_t$.

**Theorem 3:** Let $L$ be as in Theorem 2. There is a $\mu$ with $0 < \mu < 1$ such that, for any $t \geq T_{\mathcal{H}} + L$, we have
\[ \|W^t - W^*\|^2 \leq \mu^{t - T_{\mathcal{H}}} \|W^T_{\mathcal{H}} - W^*\|^2. \] (17)

Moreover, for any $t \geq T_{\mathcal{H}} + L$ we have
\[ \Delta t \leq D \|W^T_{\mathcal{H}} - W^*\| \mu^{t - T_{\mathcal{H}}}, \] (18)\]

with $D = \sup_{t} \|X_{U_t} - X_{L_t}\|$.

**Proof.** Notice first that once we show $\lim \inf \cos \phi^t = \varphi > 0$, if the iteration $t$ is unclipped, we have for $t \geq T_{\mathcal{H}}$,
\[ (D^t - D^*) - (D^{t+1} - D^*) = D^t - D^{t+1} \]
\[ = \frac{(W^t - W^*) \cdot (X_{U^t} - X_{L^t})^2}{2\|X_{U^t} - X_{L^t}\|^2} \geq \frac{1 - \varphi^2}{\varphi^2}(D^t - D^*) \]
where we have used (7); therefore,
\[ D^t - D^* \leq (1 - \varphi^2)(D^t - D^*) \]
Thus, if \( \eta = 1 - \varphi^2 \), then we have for any uncropped iteration \( t \) for which \( t \geq T_H \)
\[ \|W^{t+1} - W^*\|^2 = \frac{1}{2}(D^{t+1} - D^*) \leq \eta \|W^t - W^*\|^2 \]
Assume now \( t \geq T_H + L \) and let us denote by \( K \geq 1 \) the integer such that
\[ T_H + KL < t < T_H + (K + 1)L \]
Theorem 2 implies that, for any \( 1 \leq i \leq K \), there is a \( \hat{t}_i \) in the interval \( [T_H + iL, T_H + (i + 1)L] \) whose iteration is uncropped. Therefore,
\[ \|\tilde{W}^{t+1 + L}\|^2 \leq \|\tilde{W}^{t+1}\|^2 \leq \eta \|\tilde{W}^{t+1}\|^2 \leq \eta \|\tilde{W}^{t+1 + L}\|^2 \]
because \( \|\tilde{W}^t\|^2 = D^t = D^t - D^* \) decreases and where \( \eta \) is as in Theorem 1. As a consequence,
\[ \|\tilde{W}^t\|^2 \leq \|\tilde{W}^t\|^2 \leq \eta \|\tilde{W}^t\|^2 \leq \eta \|\tilde{W}^t\|^2 \leq \eta \|\tilde{W}^t\|^2 \leq \eta \|\tilde{W}^t\|^2 \leq \eta \|\tilde{W}^t\|^2 \leq \eta \|\tilde{W}^t\|^2 \leq \eta \|\tilde{W}^t\|^2 \leq \eta \|\tilde{W}^t\|^2 \]
But since we have \( t - T_H < (K + 1)L \), it follows that
\[ K = \frac{t - T_H}{(K + 1)L} \geq \frac{1}{2L} \]
therefore, since \( \eta < 1 \),
\[ \|\tilde{W}^t\|^2 \leq \left(\frac{\eta}{\varphi^2}\right)^{t - T_H} \|\tilde{W}^T\|^2 \leq \left(\frac{\eta}{\varphi^2}\right)^{t - T_H} \|\tilde{W}^T\|^2 \leq \mu^t \|\tilde{W}^T\|^2 \]
with \( \mu = \eta^{1/2L} \). This is just (17) and to deal with \( D^t \), observe that by Theorem 1, only patterns \( X^t \) with \( i \in H \) can be considered for \( U^t \) and \( L^t \) when \( t \geq T_H \). Thus, using (15),
\[ D^t = (W^t - W^*) \cdot (X_{U^t} - X_{L^t}) = \tilde{W}^t : Z_{L^t} \]
\[ \leq \mu \|\tilde{W}^T\| \|Z_{L^t}\| \]
with \( \mu = \sup \|Z_{ij}\| \), as we wanted to prove.

We show next that Theorem 3 implies a linear convergence rate for any one of the possible SMO-generated multiplier sequences \( \alpha^t \).

**Theorem 4:** Any multiplier sequence \( \alpha^t \) generated by the SMO algorithm converges to a dual optimal \( \alpha^o \) and we have
\[ \|\alpha^t - \alpha^o\| \leq \bar{C} \mu^{t/2} \]
for an appropriate \( \bar{C} \) and with \( \mu \) as in Theorem 3.

**Proof:** Assuming \( t \geq T_H \) and applying (10), (15) and (16), we have
\[ \|\tilde{W}^t\|^2 \geq \|\tilde{W}^t\|^2 - \|\tilde{W}^{t+1}\|^2 \geq \frac{1}{\sqrt{2}} \|\alpha^t - \alpha^o\| \|\Delta^t\| \]
where \( \delta = \min \{\|X_p - X_q\|\} \). Therefore,
\[ \|\alpha^{t+1} - \alpha^o\| \leq \kappa \|\tilde{W}^t - \tilde{W}^o\| \]
with \( \kappa = \sqrt{2}/\varphi \delta \) and using Theorem 3, it follows that
\[ \|\alpha^{t+k} - \alpha^o\| \leq \sum_{j=0}^{k-1} \|\alpha^{t+j+1} - \alpha^{t+j}\| \leq \kappa \sum_{j=0}^{k-1} \|\tilde{W}^{t+j}\| \leq \kappa \sum_{j=0}^{k-1} \frac{\|\tilde{W}^T\|^2}{\mu^{(t+j)/2}} \leq C \sum_{j=0}^{\infty} \mu^{(t+j)/2} \mu^{t/2} = \bar{C} \mu^{t/2} \]
where we have written
\[ C = \kappa \frac{\|\tilde{W}^T\|^2}{\mu^{T^2/2}}, \quad \bar{C} = C \sum_{j=0}^{\infty} \mu^{j/2} = \frac{C}{1 - \sqrt{\mu}} \]
Hence, the \( \alpha^t \) are a Cauchy sequence by (21) and since the SMO multiplier domain is bounded and closed, the sequence will converge to some \( \alpha^o \). But then \( \alpha^o \) is optimal for the dual SVM problem for we have \( D(\alpha^o) = \lim D(\alpha^t) = D^* \). Finally, since
\[ \|\alpha^o - \alpha^t\| = \lim_{k \to \infty} \|\alpha^{t+k} - \alpha^t\| \]
the linear rate (19) follows immediately from (21).
We turn now to the proof of Theorems 1 and 2, starting with the latter.

**Proof of Theorem 2.** Observe first that no iteration for which \( y_t = +1, y_{t+1} = -1 \) has to clip \( \alpha_L \) and \( \alpha_U \) as both will increase.

Let now \( N_t \) be the number of 0 multipliers at iteration \( t \). If \( t \) is a clipped iteration, it sends at least one multiplier to the 0 bound and, thus, \( N_t \leq N_{t+1} \). We call \( t \) an *increasing* iteration when \( N_t < N_{t+1} \). Since obviously \( N_t \leq N \), the sample size, the number of increasing iterations in a series of consecutive clipped ons is at most \( N \).

If, however, the iteration \( t \) is clipped but \( N_{t+1} = N_t \), it sends a multiplier to the 0 bound but must take another one out of 0. We argue next that the number of such consecutive iterations is bounded by a certain integer \( Q \). Observe first that if we have \( y_t = -1 \) and \( y_{t+1} = +1 \), both \( \alpha_L \) and \( \alpha_U \) will decrease, and we would be in an increasing iteration. Thus there are only two iteration possibilities left:

1) If we have \( y_t = +1, y_{t+1} = +1 \), then \( \alpha_L \) changes from 0 to \( \alpha_L^+ \), and, therefore, \( \alpha_U \) becomes 0.

2) If we have \( y_t = -1, y_{t+1} = -1 \), then \( \alpha_U \) becomes 0, and, therefore, \( \alpha_L^+ \) changes from 0 to \( \alpha_L^+ \).

It is readily seen that in both cases we are shuffling the values in \( \alpha^t \) to build \( \alpha^{t+1} \). But the number of consecutive distinct shuffles is obviously finite and there must be an integer \( Q \) such that after at most such \( Q \) consecutive iterations, we arrive at a pair \( t_1, t_2 \) with \( t_2 > t_1 \) and for which \( \alpha^{t_2} = \alpha^{t_1} \). But this implies that \( D^{t_2} = D^{t_1} \), contradicting the strict decrease of the \( D^t \). By obvious reasons we will call these iterations *shuffling*.

Now Theorem 2 easily follows, for the longest possible run of consecutive clipping iterations would be composed of a first set of \( Q \) shuffling iterations, then an increasing iteration, then a second set of \( Q \) shuffling iterations, then a second increasing iteration and so on until a last set of \( Q \) shuffling iterations. If the number of increasing iterations is \( P \leq N \), the maximum number of iterations in such a run would be \( (P + 1) \times Q \leq (N + 1) \times Q = L \), as we wanted to prove.

We turn now to the proof of Theorem 1.

**Proof.** We describe first the general strategy to prove \( \varphi > 0 \). If this is not true, we would have \( \lim \inf \cos \phi^t = 0 \) and there would be a subsequence \( t_j \) such that \( \lim \cos \phi^{t_j} = 0 \). Consider the unit vectors

\[
u^{t_j} = \frac{W^{t_j} - W^*}{\|W^{t_j} - W^*\|} \]

and let \( u \) be the limit of some converging subsequence of the \( u^{t_j} \), whose indices we also denote as \( t_j \) for the sake of simplicity. In the remainder of the section we will prove the estimate

\[u \cdot u^{t_j} \leq \frac{1}{\|W^*\|^2} (u \cdot W^*)(u^{t_j} \cdot W^*). \tag{22}\]

It then easily follows that

\[1 = \|u\|^2 = \lim u \cdot u^{t_j} \leq \frac{1}{\|W^*\|^2} (u \cdot W^*)(u^{t_j} \cdot W^*) \tag{23}\]

with \( \phi^* \) the angle between \( u \) and \( W^* \) and, as a consequence, that \( \cos \phi^* = \pm 1 \) and \( u = \pm W^*/\|W^*\| \).

But if this true and \( p, q \) are such that \( \alpha_p^* \alpha_q^* > 0 \) and \( y_p = 1, y_q = -1 \), the corresponding \( X_p, X_q \) must then be in the supporting hyperplanes, and we would have \( W^*(X_p - X_q) = y_p - y_q = 2 \). However, both pairs \( (p, q) \) and \( (q, p) \) are eligible for \( U^t \) and \( L^t \), for all \( t \) large enough and since we will show in Proposition 2 below that \( u \cdot (X_r - X_s) \leq 0 \) for any pair \( r, s \) of indices that are eligible for choosing all the \( L^t, U^t \), we would have \( u \cdot (X_p - X_q) = 0 \) for such a pair \( p, q \). But we arrive then at the contradiction

\[2 = W^* \cdot (X_p - X_q) = \pm \|W^*\| u \cdot (X_p - X_q) = 0, \]

and, therefore, we must have \( \varphi > 0 \).

We go next into the proof of the estimate (22) and begin by proving the following result.

**Proposition 2.** If for some subsequence \( t_{j k} \) the index \( p \) is eligible for \( U^{t_{j k}} \) and the index \( q \) is eligible for \( L^{t_{j k}} \), we have \( u \cdot (X_p - X_q) \leq 0 \).

**Proof.** If, say, \( p \) is eligible for \( U^{t_{j k}} \) and \( q \) is eligible for \( L^{t_{j k}} \), the criterion to choose \( U \) and \( L \) implies for \( t_{j k} \) large enough

\[u^{t_{j k}} \cdot (X_p - X_q) = \frac{W^{t_{j k}} \cdot (X_p - X_q)}{\|W^{t_{j k}}\|} \leq \frac{W^{t_{j k}} \cdot (X_{U^{t_{j k}}} - X_{L^{t_{j k}}})}{\|W^{t_{j k}}\|} = \frac{\|X_{U^{t_{j k}}} - X_{L^{t_{j k}}} \| \cos \phi^{t_{j k}}}{\|X_{U^{t_{j k}}} - X_{L^{t_{j k}}} \| \cos \phi^{t_{j k}}} \leq D \cos \phi^{t_{j k}},\]

with \( D = \sup_t \|X_{U^t} - X_{L^t}\| \). If we assume \( \cos \phi^{t_{j k}} \to 0 \), it follows that

\[u \cdot (X_p - X_q) = \lim u^{t_{j k}} \cdot (X_p - X_q) \leq 0,\]

as we wanted to show.

Consider now the set \( S \) defined as

\[S = \{(p, q) \colon p \text{ eligible for } U^{t_{j k}}, q \text{ eligible for } L^{t_{j k}}, \text{ for some subsequence } t_{j k}\},\]

and let \( S^c \) be its complementary.

**Proposition 3.** There is a \( T_S \) such that if \( t_j \geq T_S \) and \( (p, q) \in S^c \), we either have that \( p \) is not eligible for \( U^{t_j} \) and, hence, \( y_p = +1 \) and \( \alpha_p^t = 0 \), or that \( q \) is not eligible for \( L^{t_j} \) and, hence, \( y_q = -1 \) and \( \alpha_q^t = 0 \).

**Proof.** Assume that, say, \( y_p = +1 \); notice that if \( (p, q) \) is not in \( S \), negating the existence of the eligible subsequence shows that there is then a \( T_p \) such that for all \( t_j \geq T_p, p \) is not eligible for \( U^{t_j} \). Similarly, if \( y_q = -1 \), there is a \( T_q \)...
such that for all $t_j \geq T_q$, $q$ is not eligible for $L^j$. We can simply take now $T_{\mathcal{S}} = \max\{T_p, T_q : (p, q) \in \mathcal{S}\}$ and the assertions $\alpha^j_p = 0$ or $\alpha^j_q = 0$ are immediate.

We have now

$$u \cdot u^j = \frac{1}{\|W^j\|} u \cdot \tilde{W}^j.$$

Let us decompose $W^j$ and $W^*$ as $W^j = W^j_+ - W^j_-$, $W^* = W^*_+ - W^*_-$, with the $W^\pm$ made up only with patterns $X_p$ for which $y_p = \pm 1$, and consider $\tilde{W}^j_+ = W^j_+ - W^*_+$.

Then $u \cdot \tilde{W}^j = u \cdot (\tilde{W}^j_+ - \tilde{W}^j_-)$ and let’s define

$$\sigma^j_+ = \sum_{\delta \neq \pm 1} \alpha^j_p, \quad \sigma^j_\pm = \sum_{\delta = \pm 1} \alpha^j_p,$$

$$\sigma^j* = \sum_{\delta \neq \pm 1} \alpha^j_\pm, \quad \sigma^j* \mp = \sum_{\delta = \pm 1} \alpha^j_\pm.$$

Obviously we have $\sigma^j_+ = \sigma^j$, $2\sigma^j_\pm = \sum \alpha^j_p = \sigma^j$ and $2\sigma^j* = \sigma^j*$. Moreover,

$$u \cdot \tilde{W}^j_+ = \frac{1}{\sigma^j} u \cdot \left\{ \sigma^j* \left( W^j_+ - W^*_+ \right) \right\}$$

$$= \frac{1}{\sigma^j} u \cdot \left\{ \sigma^j* W^j_+ - \sigma^j* W^*_+ \right\}$$

$$= \frac{\sigma^j* \mp - \sigma^j*}{\sigma^j*} u \cdot W^*_+$$

$$= A^+ + B^+.$$

For the first term $A^+$ we have

$$u \cdot \left( \sigma^j* W^j_+ - \sigma^j* W^*_+ \right)$$

$$= \sum_{S_+ \times S_+} \alpha^j_p \alpha^j_q u \cdot (X_p - X_q)$$

$$= \left\{ \sum_{S_+} \sum_{S_+} \right\} \alpha^j_p \alpha^j_q u \cdot (X_p - X_q)$$

$$= A^+_+ + A^+_\mp,$$

where $S_+ = S \cap (\mathcal{I}_+ \times \mathcal{I}_+)$ and the same for $S^*_\mp$. Clearly, by Proposition 2 we have $A^+_\mp \leq 0$ and if $(p, q) \in S^*_\mp$, by Proposition 3 we must have that $\alpha^j_p = 0$, which implies that $A^+_\mp = 0$ and, thus, $A^+ \leq 0$. Now, for the $B^+$ term, we showed in (14) that for $t \geq T_q$ we have $\sigma^t = W^t \cdot W^*$ and, hence, $\sigma^t - \sigma^* = (W^t - W^*) \cdot W^* = W^t \cdot W^*$. Thus, for the $B^+$ term we have

$$\left( \sigma^j_+ - \sigma^j* \right) u \cdot W^*_+ = \frac{1}{2} \left( \sigma^j - \sigma^* \right) u \cdot W^*_+$$

$$= \frac{1}{2} \left( \tilde{W}^j \cdot W^* \right) (u \cdot W^*_+).$$

Putting everything together we arrive at

$$u \cdot \tilde{W}^j_+ \leq \frac{1}{2\sigma^j} \left( \tilde{W}^j \cdot W^* \right) (u \cdot W^*_+)$$

$$= \frac{1}{\|W^*\|^2} \left( \tilde{W}^j \cdot W^* \right) (u \cdot W^*_+),$$

where we have used that $2\sigma^j* = \sigma^* = \|W^*\|^2$. Similarly,

$$u \cdot \tilde{W}^j_- \leq \frac{1}{\sigma^j} u \cdot \left( \sigma^j* W^j_- - \sigma^j* W^*_- \right)$$

$$= \frac{\sigma^j* \mp - \sigma^j*}{\sigma^j*} u \cdot W^*_-$$

$$= A^- + B^-.$$

and, arguing as before, we have now

$$u \cdot \left( \sigma^j* W^j_- - \sigma^j* W^*_- \right)$$

$$= \sum_{S_- \times S_-} \alpha^j_p \alpha^j_q u \cdot (X_p - X_q)$$

$$= \left\{ \sum_{S_-} \sum_{S_-} \right\} \alpha^j_p \alpha^j_q u \cdot (X_p - X_q)$$

$$= A^-_+ + A^-_\mp,$$

where we introduced the set $S^-_\mp$, for which $(p, q) \in S^-_\mp$ iff $(q, p) \in S^-$. By Proposition 2 we have $u \cdot (X_p - X_q) \leq 0$, which implies that $A^-_\mp \leq 0$. Moreover, if $(p, q) \in S^-_\mp$, $(q, p) \in S^+_\mp$, and hence $\alpha^j_p = 0$, i.e., $A^-_\mp = 0$, so that $A^- \geq 0$. Furthermore, just as before, we have for the $B^-$ term that

$$\left( \sigma^j - \sigma^* \right) u \cdot W^*_- = \frac{1}{2} \left( \tilde{W}^j \cdot W^* \right) (u \cdot W^*_-);$$

therefore,

$$u \cdot \tilde{W}^j_- \geq \frac{1}{2\sigma^j} \left( \tilde{W}^j \cdot W^* \right) (u \cdot W^*_-)$$

$$= \frac{1}{\|W^*\|^2} \left( \tilde{W}^j \cdot W^* \right) (u \cdot W^*_+).$$

Putting (24) and (25) together, we have

$$u \cdot u^j = \frac{1}{\|W^j\|} u \cdot \tilde{W}^j,$$

$$\leq \frac{1}{\|W^j\|} u \cdot \left( \tilde{W}^j_+ - \tilde{W}^j_- \right)$$

$$\leq \frac{1}{\|W^j\|} \frac{\|W^j\|^2}{\|W^*\|^2} (u \cdot (W^*_+ - W^*_-))$$

$$= \frac{1}{\|W^*\|^2} \left( u \cdot W^*_+ \right) (u \cdot W^*_+);$$

i.e. the estimate (22), and then we arrive, as we wanted, at
\[ 1 = \|u\|^2 = \lim_{t \to \infty} u^t \cdot u^t = \frac{1}{\sigma^*} (u \cdot W^*)^2 \]

with \(\phi^*\) the angle between \(u\) and \(W^*\). This is just (23) and ends the proof.

IV. DISCUSSION AND CONCLUSIONS

Linear convergence of the \(D_t\) sequence of dual values that SMO generates is a known fact for the general SVM problem when the kernel matrix is definite positive; sublinear convergence is also known for any general positive semidefinite kernel matrix.

For linearly separable, i.e., hard–margin, SVM we have shown in this paper that for any kernel matrix, not only \(D_t\) converges linearly but also do so the multiplier sequence \(\alpha_t\), its associated weight vectors \(W_t\) and the KKT gap \(\Delta_t\). As a consequence of this last fact, it follows that, for linearly separable SVMs and any \(\epsilon > 0\), SMO reaches a KKT gap such that \(\Delta_T \leq \epsilon\) in \(T = \Omega\left(\log\left(\frac{1}{\epsilon}\right)\right)\) iterations, something that theoretically supports the fast convergence observed in practice for SMO.

This still leaves open linear SMO convergence for the general SVM problem. The authors have shown [10] linear convergence for the Mitchel–Dem’yanov–Malozemov (MDM) algorithm for reduced convex hulls along lines quite similar to the ones used here. This is a problem quite related to the general SVM problem and the MDM algorithm is quite similar to SMO as well, suggesting that the approach followed here may lead to extend the linear convergence results given for SMO to the \(L_1\) soft–margin SVM problem with a general kernel matrix. We are working on this and other related questions.

REFERENCES