A Study on SMO-type Decomposition Methods for Support Vector Machines

Pai-Hsuen Chen, Rong-En Fan, and Chih-Jen Lin
Department of Computer Science, National Taiwan University, Taipei 106, Taiwan
cjlin@csie.ntu.edu.tw

Abstract

Decomposition methods are currently one of the major methods for training support vector machines. They vary mainly according to different working set selections. Existing implementations and analysis usually consider some specific selection rules. This article studies Sequential Minimal Optimization (SMO)-type decomposition methods under a general and flexible way of choosing the two-element working set. Main results include: 1) a simple asymptotic convergence proof, 2) a general explanation of the shrinking and caching techniques, and 3) the linear convergence of the methods. Extensions to some SVM variants are also discussed.

I. Introduction

The support vector machines (SVMs) [1], [6] are useful classification tools. Given training vectors $\mathbf{x}_i \in R^n, i = 1, \ldots, l$, in two classes, and a vector $\mathbf{y} \in R^l$ such that $y_i \in \{1, -1\}$, SVMs require the solution of the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & f(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \\ \text{subject to} \quad & 0 \le \alpha_i \le C, i = 1, \ldots, l, \\ & \mathbf{y}^T \boldsymbol{\alpha} = 0, \end{aligned} \tag{1}$$

where $\mathbf{e}$ is the vector of all ones, $C < \infty$ is the upper bound of all variables, and Q is an l by l symmetric matrix. Training vectors $\mathbf{x}_i$ are mapped into a higher (maybe infinite) dimensional space by the function ϕ, $Q_{ij} \equiv y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) = y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$, and $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function. Then Q is a positive semi-definite (PSD) matrix. Occasionally some researchers use kernel functions not in the form of $\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$, so Q may be indefinite. Here we also consider such cases and, hence, require only Q to be symmetric.

Due to the density of the matrix Q, the memory problem causes that traditional optimization methods cannot be directly applied to solve (1). Currently the decomposition method is one of the major methods to train SVM (e.g. [3], [10], [23], [25]). This method, an iterative procedure, considers only a small subset of variables per iteration. Denoted as B, this subset is called the working set. Since each iteration involves only $|B|$ columns of the matrix Q, the memory problem is solved.

A special decomposition method is the Sequential Minimal Optimization (SMO) [25], which restricts B to only two elements. Then at each iteration one solves a simple two-variable problem without needing optimization software. It is sketched in the following:

Algorithm 1 (SMO-type Decomposition methods)

1) Find $\boldsymbol{\alpha}^1$ as the initial feasible solution. Set $k = 1$.
2) If $\boldsymbol{\alpha}^k$ is an stationary point of (1), stop. Otherwise, find a *two-element* working set $B = \{i, j\} \subset \{1, \ldots, l\}$. Define $N \equiv \{1, \ldots, l\} \backslash B$ and $\boldsymbol{\alpha}^k_B$ and $\boldsymbol{\alpha}^k_N$ as sub-vectors of $\boldsymbol{\alpha}^k$ corresponding to B and N, respectively.
3) Solve the following sub-problem with the variable $\boldsymbol{\alpha}_B$:

$$\begin{aligned} \min_{\boldsymbol{\alpha}_B} \quad & \frac{1}{2}\begin{bmatrix}\boldsymbol{\alpha}_B^T & (\boldsymbol{\alpha}_N^k)^T\end{bmatrix}\begin{bmatrix}Q_{BB} & Q_{BN}\\ Q_{NB} & Q_{NN}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}_B\\ \boldsymbol{\alpha}_N^k\end{bmatrix} - \begin{bmatrix}\mathbf{e}_B^T & \mathbf{e}_N^T\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}_B\\ \boldsymbol{\alpha}_N^k\end{bmatrix} \\ & = \frac{1}{2}\begin{bmatrix}\alpha_i & \alpha_j\end{bmatrix}\begin{bmatrix}Q_{ii} & Q_{ij}\\ Q_{ij} & Q_{jj}\end{bmatrix}\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + (-\mathbf{e}_B + Q_{BN}\boldsymbol{\alpha}_N^k)^T\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + \text{ constant} \\ \text{subject to} \quad & 0 \le \alpha_i, \alpha_j \le C, \\ & y_i\alpha_i + y_j\alpha_j = -\mathbf{y}_N^T\boldsymbol{\alpha}_N^k, \end{aligned} \tag{2}$$

   where $\begin{bmatrix}Q_{BB} & Q_{BN}\\ Q_{NB} & Q_{NN}\end{bmatrix}$ is a permutation of the matrix Q.
   Set $\boldsymbol{\alpha}_B^{k+1}$ to be the optimal point of (2).
4) Set $\boldsymbol{\alpha}_N^{k+1} \equiv \boldsymbol{\alpha}_N^k$. Set $k \leftarrow k + 1$ and goto Step 2.

Note that the set B changes from one iteration to another. To simplify the notation, we use B instead of B^k. Algorithm 1 does not specify how to choose the working set B as there are many possible ways. Regarding the sub-problem (2), if Q is positive semi-definite, then it is convex and can be easily solved. For indefinite Q, the situation is more complicated, and this issue is addressed in Section III.

If a proper working set is used at each iteration, the function value $f(\boldsymbol{\alpha}^k)$ strictly decreases. However, this property does not imply that the sequence $\{\boldsymbol{\alpha}^k\}$ converges to a stationary point of (1). Hence, proving the convergence of decomposition methods is usually a challenging task. Under certain rules for selecting the working set, the asymptotic convergence has been established (e.g. [16], [2], [9], [20]). These selection rules may allow an arbitrary working set size, so they reduce to SMO-type methods when the size is limited to two.

This article provides a comprehensive study on SMO-type decomposition methods. The selection rule of the two-element working set B is under a very general setting. Thus, all results here apply to any SMO-type implementation whose selection rule meets the criteria of this paper. In Section II, we discuss existing working set selections for SMO-type methods and propose a general scheme. Section III proves the asymptotic convergence and gives a comparison to former convergence studies.

Shrinking and caching are two effective techniques to speed up the decomposition methods. Earlier research [18] gives the theoretical foundation of these two techniques, but requires some assumptions. In Section IV, we provide a better and a more general explanation without these assumptions. Convergence rates are another important issue and they indicate how fast the method approaches an optimal solution. We establish the linear convergence of the proposed method in Section V. All the above results hold not only for support vector classification, but also for regression and other variants of SVM. Such extensions are presented in Section VI. Finally, Section VII is the conclusion.

II. Existing and New Working Set Selections for SMO-type Methods

In this Section, we discuss existing working set selections and then propose a general scheme for SMO-type methods.

A. Existing Selections

Currently a popular way to select the working set B is via the "maximal violating pair:"

WSS 1 (Working set selection via the "maximal violating pair")

1) Select

$$\begin{aligned} i &\in \arg \max_{t \in I_{\text{up}}(\boldsymbol{\alpha}^k)} -y_t \nabla f(\boldsymbol{\alpha}^k)_t, \\ j &\in \arg \min_{t \in I_{\text{low}}(\boldsymbol{\alpha}^k)} -y_t \nabla f(\boldsymbol{\alpha}^k)_t, \end{aligned}$$

where

$$\begin{aligned} I_{\text{up}}(\boldsymbol{\alpha}) &\equiv \{t \mid \alpha_t < C, y_t = 1 \text{ or } \alpha_t > 0, y_t = -1\}, \text{ and} \\ I_{\text{low}}(\boldsymbol{\alpha}) &\equiv \{t \mid \alpha_t < C, y_t = -1 \text{ or } \alpha_t > 0, y_t = 1\}. \end{aligned} \tag{3}$$

2) Return $B = \{i, j\}$.

This working set was first proposed in [14] and has been used in many software packages, for example, LIBSVM [3].

WSS 1 can be derived through the Karush-Kuhn-Tucker (KKT) optimality condition of (1): A vector $\boldsymbol{\alpha}$ is a stationary point of (1) if and only if there is a number b and two nonnegative vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ such that

$$\nabla f(\boldsymbol{\alpha}) + b\mathbf{y} = \boldsymbol{\lambda} - \boldsymbol{\mu},$$

$$\lambda_i \alpha_i = 0, \mu_i (C - \boldsymbol{\alpha})_i = 0, \lambda_i \geq 0, \mu_i \geq 0, i = 1, \ldots, l,$$

where $\nabla f(\boldsymbol{\alpha}) \equiv Q\boldsymbol{\alpha} - \mathbf{e}$ is the gradient of $f(\boldsymbol{\alpha})$. This condition can be rewritten as

$$\nabla f(\boldsymbol{\alpha})_i + b y_i \geq 0 \qquad \text{if } \alpha_i < C, \tag{4}$$

$$\nabla f(\boldsymbol{\alpha})_i + b y_i \leq 0 \qquad \text{if } \alpha_i > 0. \tag{5}$$

Since $y_i = \pm 1$, by defining $I_{\rm up}(\boldsymbol{\alpha})$ and $I_{\rm low}(\boldsymbol{\alpha})$ as in (3), and rewriting (4)-(5) to have the range of b, a feasible $\boldsymbol{\alpha}$ is a stationary point of (1) if and only if

$$m(\boldsymbol{\alpha}) \leq M(\boldsymbol{\alpha}), \tag{6}$$

where

$$m(\boldsymbol{\alpha}) \equiv \max_{i \in I_{\rm up}(\boldsymbol{\alpha})} -y_i \nabla f(\boldsymbol{\alpha})_i, \text{ and } M(\boldsymbol{\alpha}) \equiv \min_{i \in I_{\rm low}(\boldsymbol{\alpha})} -y_i \nabla f(\boldsymbol{\alpha})_i.$$

Note that $m(\boldsymbol{\alpha})$ and $M(\boldsymbol{\alpha})$ are well defined except in a rare situation where all $y_i = 1$ (or -1). In this case the zero vector is the only feasible solution of (1), so the decomposition method stops at the first iteration.

Following [14], we define a "violating pair" of the condition (6) as:

Definition 1 (Violating pair) *If $i \in I_{\rm up}(\boldsymbol{\alpha}), j \in I_{\rm low}(\boldsymbol{\alpha})$, and $-y_i \nabla f(\boldsymbol{\alpha})_i > -y_j \nabla f(\boldsymbol{\alpha})_j$, then $\{i, j\}$ is a "violating pair."*

From (6), the indices $\{i, j\}$ which most violate the condition are a natural choice of the working set. They are called a "maximal violating pair" in WSS 1.

Violating pairs play an important role in making the decomposition methods work, as demonstrated by:

Theorem 1 ([9]) *Assume Q is positive semi-definite. The decomposition method has the strict decrease of the objective function value (i.e. $f(\boldsymbol{\alpha}^{k+1}) < f(\boldsymbol{\alpha}^k), \forall k$) if and only if at each iteration B includes at least one violating pair.*

For SMO-type methods, if Q is PSD, this theorem implies that B must be a violating pair. Unfortunately, having a violating pair in B and then the strict decrease of $f(\boldsymbol{\alpha}^k)$ do not guarantee the convergence to a stationary point. An interesting example from [13] is as follows: Given five data $\mathbf{x}^1 = [1, 0, 0]^T, \mathbf{x}^2 = [0, 1, 0]^T, \mathbf{x}^3 = [0, 0, 1]^T, \mathbf{x}^4 = [0.1, 0.1, 0.1]^T$, and $\mathbf{x}^5 = [0, 0, 0]^T$. If $y_1 = \cdots = y_4 = -1, y_5 = 1$, $C = 100$, and the linear kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ is used, then the optimal solution of (1) is $[0, 0, 0, 200/3, 200/3]^T$ and the optimal objective value is $-200/3$.

Starting with $\boldsymbol{\alpha}^0 = [0, 0, 0, 0, 0]^T$, if in the first iteration the working set is $\{1, 5\}$, we have $\boldsymbol{\alpha}^1 = [2, 0, 0, 0, 2]^T$ with the objective value -2. Then if we choose the following working sets at the next three iterations:

$$\{1, 2\} \rightarrow \{2, 3\} \rightarrow \{3, 1\}, \tag{7}$$

the next three $\boldsymbol{\alpha}$ are:

$$[1, 1, 0, 0, 2]^T \rightarrow [1, 0.5, 0.5, 0, 2]^T \rightarrow [0.75, 0.5, 0.75, 0, 2]^T.$$

If we continue as in (7) for choosing the working set, the algorithm requires an infinite number of iterations. In addition, the sequence converges to a non-optimal point $[2/3, 2/3, 2/3, 0, 2]^T$ with the objective value $-10/3$. All working sets used are violating pairs.

A careful look at the above procedure reveals why it fails to converge to the optimal solution. In (7), we have the following for the three consecutive iterations:

$-y_t \nabla f(\boldsymbol{\alpha}^k)_t, t = 1, \ldots, 5$	$-y_i \nabla f(\boldsymbol{\alpha}^k)_i + y_j \nabla f(\boldsymbol{\alpha}^k)_j$	$m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k)$
$[0, 0, -1, -0.8, 1]^T$	1	2
$[0, -0.5, -0.5, -0.8, 1]^T$	0.5	1.8
$[-0.25, -0.5, -0.25, -0.8, 1]^T$	0.25	1.8

Clearly, the selected $B = \{i, j\}$, though a violating pair, does not lead to the reduction of the maximal violation $m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k)$. This discussion shows the importance of the maximal violating pair in the decomposition method. When such a pair is used as the working set (i.e. WSS 1), the convergence has been fully established in [16] and [17]. However, if B is not the maximal violating pair, it is unclear whether SMO-type methods still converge to a stationary point. Motivated from the above analysis, we conjecture that a "sufficiently violated" pair is enough and propose a general working set selection in the following subsection.

B. A General Working Set Selection for SMO-type Decomposition Methods

We propose choosing a "constant-factor" violating pair as the working set. That is, the difference between the two selected indices is larger than a constant fraction of that between the maximal violating pair.

WSS 2 (Working set selection: constant-factor violating pair)

1) Consider a fixed $0 < \sigma \leq 1$ for all iterations.
2) Select any $i \in I_{\text{up}}(\boldsymbol{\alpha}^k), j \in I_{\text{low}}(\boldsymbol{\alpha}^k)$ satisfying

$$-y_i \nabla f(\boldsymbol{\alpha}^k)_i + y_j \nabla f(\boldsymbol{\alpha}^k)_j \geq \sigma(m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k)) > 0. \tag{8}$$

3) Return $B = \{i, j\}$.

Clearly (8) ensures the quality of the selected pair by linking it to the maximal violating pair. We can consider an even more general relationship between the two pairs:

WSS 3 (Working set selection: a generalization of WSS 2)

1) Let $h : R^1 \to R^1$ be any function satisfying
   a) h strictly increases on $x \geq 0$,
   b) $h(x) \leq x, \forall x \geq 0, h(0) = 0$.

2) Select any $i \in I_{\text{up}}(\boldsymbol{\alpha}^k), j \in I_{\text{low}}(\boldsymbol{\alpha}^k)$ satisfying

$$-y_i \nabla f(\boldsymbol{\alpha}^k)_i + y_j \nabla f(\boldsymbol{\alpha}^k)_j \geq h(m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k)) > 0. \tag{9}$$

3) Return $B = \{i, j\}$.

The condition $h(x) \leq x$ ensures that there is at least one pair $\{i, j\}$ satisfying (9). Clearly, $h(x) = \sigma x$ with $0 < \sigma \leq 1$ fulfills all required conditions and WSS 3 reduces to WSS 2. The function h can be in a more complicated form. For example, if

$$h(m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k)) \equiv \min\left(m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k), \sqrt{m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k)}\right), \tag{10}$$

then (10) also satisfies all requirements.

Subsequently, we mainly analyze the SMO-type method using WSS 3 for the working set selection. The only exception is the linear convergence analysis, which considers WSS 2.

III. Asymptotic Convergence

The decomposition method generates a sequence $\{\boldsymbol{\alpha}^k\}$. If it is finite, then a stationary point is obtained. Hence we consider only the case of an infinite sequence. This section establishes the asymptotic convergence of using WSS 3 for the working set selection. First we discuss the sub-problem in Step 3) of Algorithm 1 with indefinite Q. Solving it relates to the convergence proof.

A. Solving the sub-problem

Past convergence analysis usually requires an important property, the function value is sufficiently decreased: There is $\lambda > 0$ such that

$$f(\boldsymbol{\alpha}^{k+1}) \leq f(\boldsymbol{\alpha}^k) - \lambda \|\boldsymbol{\alpha}^{k+1} - \boldsymbol{\alpha}^k\|^2, \text{ for all } k. \tag{11}$$

If Q is positive semi-definite and the working set $\{i, j\}$ is a violating pair, [17] has proved (11). However, it is difficult to obtain the same result if Q is indefinite. Some such kernels are, for example, the sigmoid kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j + r)$ [19] and the edit-distance kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \exp^{-d(\mathbf{x}_i, \mathbf{x}_j)}$, where $d(\mathbf{x}_i, \mathbf{x}_j)$ is the edit distance of two strings $\mathbf{x}_i$ and $\mathbf{x}_j$ [5]. To obtain (11), [24] modified the sub-problem (2) to the following form:

$$\begin{aligned} \min_{\boldsymbol{\alpha}_B} \quad & f\left(\begin{bmatrix} \boldsymbol{\alpha}_B \\ \boldsymbol{\alpha}_N^k \end{bmatrix}\right) + \tau \|\boldsymbol{\alpha}_B - \boldsymbol{\alpha}_B^k\|^2 \\ \text{subject to} \quad & \mathbf{y}_B^T \boldsymbol{\alpha}_B = -\mathbf{y}_N^T \boldsymbol{\alpha}_N^k, \\ & 0 \leq \alpha_i \leq C,\ i \in B, \end{aligned} \tag{12}$$

where $\tau > 0$ is a constant. Clearly, an optimal $\boldsymbol{\alpha}_B^{k+1}$ of (12) leads to

$$f(\boldsymbol{\alpha}^{k+1}) + \tau \|\boldsymbol{\alpha}_B^{k+1} - \boldsymbol{\alpha}_B^k\|^2 \leq f(\boldsymbol{\alpha}^k)$$

and then (11). However, this change also causes differences to most SVM implementations, which consider positive semi-definite Q and use the sub-problem (2). Moreover, (12) may still be non-convex and possess more than one local minimum. Then for implementations, convex and non-convex cases may have to be handled separately. Therefore, we propose another way to solve the sub-problem where

1) The same sub-problem (2) is used when Q is positive definite (PD), and
2) The sub-problem is always convex.

An SMO-type decomposition method with special handling on indefinite Q is in the following algorithm.

Algorithm 2 (Algorithm 1 with specific handling on indefinite Q)

Steps 1), 2), and 4) are the same as those in Algorithm 1.

Step 3') Let $\tau > 0$ be a constant throughout all iterations, and define

$$a \equiv Q_{ii} + Q_{jj} - 2y_i y_j Q_{ij}. \tag{13}$$

a) If $a > 0$, then solve the sub-problem (2). Set $\boldsymbol{\alpha}_B^{k+1}$ to be the optimal point of (2).

b) If $a \le 0$, then solve a modified sub-problem:

$$\begin{aligned} \min_{\alpha_i,\alpha_j} \quad & \frac{1}{2}\begin{bmatrix}\alpha_i & \alpha_j\end{bmatrix}\begin{bmatrix}Q_{ii} & Q_{ij}\\ Q_{ij} & Q_{jj}\end{bmatrix}\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + (-\mathbf{e}_B + Q_{BN}\boldsymbol{\alpha}_N^k)^T\begin{bmatrix}\alpha_i\\ \alpha_j\end{bmatrix} + \\ & \frac{\tau - a}{4}((\alpha_i - \alpha_i^k)^2 + (\alpha_j - \alpha_j^k)^2) \\ \text{subject to} \quad & 0 \le \alpha_i, \alpha_j \le C, \\ & y_i\alpha_i + y_j\alpha_j = -\mathbf{y}_N^T\boldsymbol{\alpha}_N^k. \end{aligned} \tag{14}$$

Set $\boldsymbol{\alpha}_B^{k+1}$ to be the optimal point of (14).

The additional term in (14) has a coefficient $(\tau - a)/4$ related to the matrix Q, but that in (12) does not. We will show later that our modification leads (14) to a convex optimization problem.

Next we discuss how to easily solve the two sub-problems (2) and (14). Consider $\alpha_i \equiv \alpha_i^k + y_i d_i$ and $\alpha_j \equiv \alpha_j^k + y_j d_j$. The sub-problem (2) is equivalent to

$$\begin{aligned} \min_{d_i,d_j} \quad & \frac{1}{2}\begin{bmatrix}d_i & d_j\end{bmatrix}\begin{bmatrix}Q_{ii} & y_i y_j Q_{ij}\\ y_i y_j Q_{ij} & Q_{jj}\end{bmatrix}\begin{bmatrix}d_i\\ d_j\end{bmatrix} + (-\mathbf{e}_B + (Q\boldsymbol{\alpha})_B)^T\begin{bmatrix}y_i d_i\\ y_j d_j\end{bmatrix} \\ \text{subject to} \quad & d_i + d_j = 0, \\ & 0 \le \alpha_i^k + y_i d_i, \alpha_j^k + y_j d_j \le C. \end{aligned} \tag{15}$$

Substituting $d_i = -d_j$ into the objective function leads to a one-variable optimization problem:

$$\begin{aligned} \min_{d_j} \quad & g(d_j) \equiv f(\boldsymbol{\alpha}) - f(\boldsymbol{\alpha}^k) \\ & = \frac{1}{2}(Q_{ii} + Q_{jj} - 2y_i y_j Q_{ij})d_j^2 + (-y_i\nabla f(\boldsymbol{\alpha}^k)_i + y_j\nabla f(\boldsymbol{\alpha}^k)_j)d_j \\ \text{subject to} \quad & L \le d_j \le U, \end{aligned} \tag{16}$$

where L and U are lower and upper bounds of d_j after including information on $d_i : d_i = -d_j$ and $0 \leq \alpha_i^k + y_i d_i \leq C$. As a defined in (13) is the leading coefficient of the quadratic function (16), $a > 0$ if and only if the sub-problem (2) is strictly convex.

We then claim that

$$L < 0 \text{ and } U \geq 0. \tag{17}$$

If $y_i = y_j = 1$, $0 \leq \alpha_i^k + y_i d_i, \alpha_j^k + y_j d_j \leq C$ implies

$$L = \max(-\alpha_j^k, \alpha_i^k - C) \leq d_j \leq \min(C - \alpha_j^k, \alpha_i^k) = U.$$

Clearly, $U \geq 0$. Since $i \in I_{\text{up}}(\boldsymbol{\alpha}^k)$ and $j \in I_{\text{low}}(\boldsymbol{\alpha}^k)$, we have $\alpha_j^k > 0$ and $\alpha_i^k < C$. Thus, $L < 0$. For other values of y_i and y_j, the situations are similar.

When the sub-problem (2) is not strictly convex, we consider (14) via adding an additional term. With $d_i = -d_j$,

$$\frac{\tau - a}{4} \|\boldsymbol{\alpha}_B - \boldsymbol{\alpha}_B^k\|^2 = \frac{\tau - a}{2} d_j^2. \tag{18}$$

Thus, by defining

$$a_1 \equiv \begin{cases} a & \text{if } a > 0, \\ \tau & \text{otherwise,} \end{cases} \tag{19}$$

and

$$a_2 \equiv -y_i \nabla f(\boldsymbol{\alpha}^k)_i + y_j \nabla f(\boldsymbol{\alpha}^k)_j > 0, \tag{20}$$

problems (2) and (14) are essentially the following strictly convex optimization problem:

$$\begin{aligned} \min_{d_j} \quad & \frac{1}{2} a_1 d_j^2 + a_2 d_j \\ \text{subject to} \quad & L \leq d_j \leq U. \end{aligned} \tag{21}$$

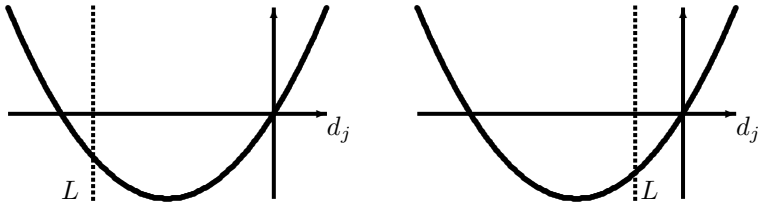
Moreover, $a_1 > 0$, $a_2 > 0$, $L < 0$, and $U \geq 0$. The quadratic objective function is shown in Figure 1 and the optimum is

$$\bar{d}_j = \max(\frac{-a_2}{a_1}, L) < 0. \tag{22}$$

Therefore, once a_1 is defined in (19) according to whether $a > 0$ or not, the two sub-problems can be easily solved by the same method as in (22).

To check the decrease of the function value, the definition of the function g in (16), $\bar{d}_j < 0$ and $a_1 \bar{d}_j + a_2 \geq 0$ from (22), and $\|\boldsymbol{\alpha}^{k+1} - \boldsymbol{\alpha}^k\|^2 = 2\bar{d}_j{}^2$ imply

$$\begin{aligned} f(\boldsymbol{\alpha}^{k+1}) - f(\boldsymbol{\alpha}^k) &= g(\bar{d}_j) \\ &= \frac{1}{2} a_1 \bar{d}_j{}^2 + a_2 \bar{d}_j \\ &= (a_1 \bar{d}_j + a_2) \bar{d}_j - \frac{a_1}{2} \bar{d}_j{}^2 \\ &\leq -\frac{a_1}{2} \bar{d}_j{}^2 \\ &= -\frac{a_1}{4} \|\boldsymbol{\alpha}^{k+1} - \boldsymbol{\alpha}^k\|^2. \end{aligned}$$



(a) $L \le \frac{-a_2}{a_1}$ (b) $L > \frac{-a_2}{a_1}$

Fig. 1. Solving the convex sub-problem (21)

Therefore, by defining

$$\lambda \equiv \frac{1}{4}\min(\tau, \min\{Q_{ii} + Q_{jj} - 2y_i y_j Q_{ij} \mid Q_{ii} + Q_{jj} - 2y_i y_j Q_{ij} > 0\}), \tag{23}$$

we have the following lemma:

Lemma 1 *Assume at each iteration of Algorithm 2 the working set is a violating pair and let* $\{\boldsymbol{\alpha}^k\}$ *be the sequence generated. Then* (11) *holds with* λ *defined in* (23)*.*

The above discussion is related to [19], which examines how to solve (2) when it is non-convex.

Note that $a = 0$ may happen as long as Q is not PD (i.e., Q is PSD or indefinite). Then, (16), becoming a linear function, is convex but not strictly convex. To have (11), an additional term makes it quadratic. Given that the function is already convex, we of course hope that there is no need to modify (2) to (14). The following theorem shows that if Q is PSD and $\tau < 2/C$, the two sub-problems indeed have the same optimal solution. Therefore, in a sense, for PSD Q, the sub-problem (2) is always used.

Theorem 2 *Assume* Q *is PSD, the working set of Algorithm 2 is a violating pair, and* $\tau \le \frac{2}{C}$*. If* $a = 0$*, then the optimal solution of* (14) *is the same as that of* (2)*.*

Proof: From (16) and $a_2 > 0$ in (20), if $a = 0$, (2) has optimal $\bar{d}_j = L$. For (14) and the transformed problem (21), $a = 0$ implies $a_1 = \tau$. Moreover, since Q is PSD,

$$a = 0 = Q_{ii} + Q_{jj} - 2y_i y_j Q_{ij} = \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2,$$

and hence $\phi(\mathbf{x}_i) = \phi(\mathbf{x}_j)$. Therefore, $K_{it} = K_{jt}, \forall t$ implies

$$\begin{aligned} & -y_i \nabla f(\boldsymbol{\alpha}^k)_i + y_j \nabla f(\boldsymbol{\alpha}^k)_j \\ = & -\sum_{t=1}^{l} y_t K_{it} \alpha_t^k + y_i + \sum_{t=1}^{l} y_t K_{jt} \alpha_t^k - y_j \\ = & \; y_i - y_j. \end{aligned}$$

Since $\{i, j\}$ is a violating pair, $y_i - y_j > 0$ indicates that $a_2 = y_i - y_j = 2$. As $\tau \leq \frac{2}{C}$, (22) implies that $\bar{d}_j = L$ is the optimal solution of (14). Thus, we have shown that (2) and (14) have the same optimal point. The derivation of $a_2 = 2$ was first developed in [17]. ■

B. Proof of Asymptotic Convergence

Using (11), we then have the following lemma:

Lemma 2 *Assume the working set at each iteration of Algorithm 2 is a violating pair. If a sub-sequence $\{\boldsymbol{\alpha}^k\}, k \in \mathcal{K}$ converges to $\bar{\boldsymbol{\alpha}}$, then for any given positive integer s, the sequence $\{\boldsymbol{\alpha}^{k+s}\}, k \in \mathcal{K}$ converges to $\bar{\boldsymbol{\alpha}}$ as well.*

Proof: The lemma was first proved in [16, Lemma IV.3]. For completeness we give details here. For the sub-sequence $\{\boldsymbol{\alpha}^{k+1}\}, k \in \mathcal{K}$, from Lemma 1 and the fact that $\{f(\boldsymbol{\alpha}^k)\}$ is a bounded decreasing sequence, we have

$$\begin{aligned} & \lim_{k \in \mathcal{K}, k \to \infty} \|\boldsymbol{\alpha}^{k+1} - \bar{\boldsymbol{\alpha}}\| \\ \leq & \lim_{k \in \mathcal{K}, k \to \infty} (\|\boldsymbol{\alpha}^{k+1} - \boldsymbol{\alpha}^k\| + \|\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}\|) \\ \leq & \lim_{k \in \mathcal{K}, k \to \infty} \left(\sqrt{\frac{1}{\lambda}(f(\boldsymbol{\alpha}^k) - f(\boldsymbol{\alpha}^{k+1}))} + \|\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}\| \right) \\ = & \ 0. \end{aligned}$$

Thus

$$\lim_{k \to \infty, k \in \mathcal{K}} \boldsymbol{\alpha}^{k+1} = \bar{\boldsymbol{\alpha}}.$$

From $\{\boldsymbol{\alpha}^{k+1}\}$ we can prove $\lim_{k \to \infty, k \in \mathcal{K}} \boldsymbol{\alpha}^{k+2} = \bar{\boldsymbol{\alpha}}$ too. Therefore, $\lim_{k \to \infty, k \in \mathcal{K}} \boldsymbol{\alpha}^{k+s} = \bar{\boldsymbol{\alpha}}$ for any given s. ■

Since the feasible region of (1) is compact, there is at least one convergent sub-sequence of $\{\boldsymbol{\alpha}^k\}$. The following theorem shows that the limit point of any convergent sub-sequence is a stationary point of (1).

Theorem 3 *Let $\{\boldsymbol{\alpha}^k\}$ be the infinite sequence generated by the SMO-type method Algorithm 2 using WSS 3. Then any limit point of $\{\boldsymbol{\alpha}^k\}$ is a stationary point of* (1)*.*

Proof: Assume that $\bar{\boldsymbol{\alpha}}$ is the limit point of any convergent sub-sequence $\{\boldsymbol{\alpha}^k\}, k \in \mathcal{K}$. If $\bar{\boldsymbol{\alpha}}$ is not a stationary point of (1), it does not satisfy the optimality condition (6). Thus, there is at least one maximal "violating pair":

$$\bar{i} \in \arg m(\bar{\boldsymbol{\alpha}}) \qquad \bar{j} \in \arg M(\bar{\boldsymbol{\alpha}}) \tag{24}$$

such that

$$\Delta \equiv -y_{\bar{i}} \nabla f(\bar{\boldsymbol{\alpha}})_{\bar{i}} + y_{\bar{j}} \nabla f(\bar{\boldsymbol{\alpha}})_{\bar{j}} > 0. \tag{25}$$

We further define

$$\Delta' \equiv \min\left(\Delta, \frac{1}{2}\min\Big\{|-y_t\nabla f(\bar{\boldsymbol{\alpha}})_t + y_s\nabla f(\bar{\boldsymbol{\alpha}})_s| \,\Big|\, -y_t\nabla f(\bar{\boldsymbol{\alpha}})_t \neq -y_s\nabla f(\bar{\boldsymbol{\alpha}})_s\Big\}\right) > 0. \tag{26}$$

Lemma 2, the continuity of $\nabla f(\boldsymbol{\alpha})$, and $h(\frac{\Delta'}{2}) > 0$ from (26) imply that for any given r, there is $\bar{k} \in \mathcal{K}$ such that for all $k \in \mathcal{K}, k \geq \bar{k}$:

$$\text{For } u = 0, \ldots, r, \qquad -y_{\bar{i}}\nabla f(\boldsymbol{\alpha}^{k+u})_{\bar{i}} + y_{\bar{j}}\nabla f(\boldsymbol{\alpha}^{k+u})_{\bar{j}} > \Delta'. \tag{27}$$

$$\text{If } i \in I_{\text{up}}(\bar{\boldsymbol{\alpha}}), \text{ then } i \in I_{\text{up}}(\boldsymbol{\alpha}^k), \ldots, i \in I_{\text{up}}(\boldsymbol{\alpha}^{k+r}). \tag{28}$$

$$\text{If } i \in I_{\text{low}}(\bar{\boldsymbol{\alpha}}), \text{ then } i \in I_{\text{low}}(\boldsymbol{\alpha}^k), \ldots, i \in I_{\text{low}}(\boldsymbol{\alpha}^{k+r}). \tag{29}$$

$$\begin{aligned}&\text{If } -y_i\nabla f(\bar{\boldsymbol{\alpha}})_i > -y_j\nabla f(\bar{\boldsymbol{\alpha}})_j, \text{ then for } u = 0, \ldots, r,\\ &-y_i\nabla f(\boldsymbol{\alpha}^{k+u})_i > -y_j\nabla f(\boldsymbol{\alpha}^{k+u})_j + \frac{\Delta'}{\sqrt{2}}.\end{aligned} \tag{30}$$

$$\begin{aligned}&\text{If } -y_i\nabla f(\bar{\boldsymbol{\alpha}})_i = -y_j\nabla f(\bar{\boldsymbol{\alpha}})_j, \text{ then for } u = 0, \ldots, r,\\ &|-y_i\nabla f(\boldsymbol{\alpha}^{k+u})_i + y_j\nabla f(\boldsymbol{\alpha}^{k+u})_j| < h(\Delta').\end{aligned} \tag{31}$$

$$\begin{aligned}&\text{For } u = 0, \ldots, r-1,\\ &(\tau - \hat{a})\|\boldsymbol{\alpha}^{k+u+1} - \boldsymbol{\alpha}^{k+u}\| \leq \Delta',\\ &\text{where } \hat{a} \equiv \min\{Q_{ii} + Q_{jj} - 2y_iy_jQ_{ij} \mid Q_{ii} + Q_{jj} - 2y_iy_jQ_{ij} < 0\}.\end{aligned} \tag{32}$$

$$\begin{aligned}&\text{If } -y_i\nabla f(\bar{\boldsymbol{\alpha}})_i > -y_j\nabla f(\bar{\boldsymbol{\alpha}})_j \text{ and } \{i, j\} \text{ is the working set at the } (k+u)\text{th iteration,}\\ &0 \leq u \leq r-1, \text{ then } i \notin I_{\text{up}}(\boldsymbol{\alpha}^{k+u+1}) \text{ or } j \notin I_{\text{low}}(\boldsymbol{\alpha}^{k+u+1}).\end{aligned} \tag{33}$$

The derivation of (27)-(32) is similar, so only the details of (27) are shown. Lemma 2 implies sequences $\{\boldsymbol{\alpha}^k\}, \{\boldsymbol{\alpha}^{k+1}\}, \ldots, \{\boldsymbol{\alpha}^{k+r}\}, k \in \mathcal{K}$ all converge to $\bar{\boldsymbol{\alpha}}$. The continuity of $\nabla f(\boldsymbol{\alpha})$ and (26) imply that for any given $0 \leq u \leq r$ and the corresponding sequence $\{\boldsymbol{\alpha}^{k+u}\}, k \in \mathcal{K}$, there is k_u such that (27) holds for all $k \geq k_u, k \in \mathcal{K}$. As r is finite, by selecting $\bar{k}$ to be the largest of these $k_u, u = 0, \ldots, r$, (27) is valid for all $u = 0, \ldots, r$. Next we derive (33) [16, Lemma IV.4]. Similar to the optimality condition (6) for problem (1), for the sub-problem at the $(k+u)$th iteration, if problem (14) is considered and $\boldsymbol{\alpha}_B$ is a stationary point, then

$$\begin{aligned}&\max_{t \in I_{\text{up}}(\boldsymbol{\alpha}_B)} -y_t\nabla f\left(\begin{bmatrix}\boldsymbol{\alpha}_B\\ \boldsymbol{\alpha}_N^{k+u}\end{bmatrix}\right)_t - \frac{y_t(\tau - a)}{2}(\alpha_t - \alpha_t^k)\\ \leq\ &\min_{t \in I_{\text{low}}(\boldsymbol{\alpha}_B)} -y_t\nabla f\left(\begin{bmatrix}\boldsymbol{\alpha}_B\\ \boldsymbol{\alpha}_N^{k+u}\end{bmatrix}\right)_t - \frac{y_t(\tau - a)}{2}(\alpha_t - \alpha_t^k).\end{aligned}$$

Now $B = \{i, j\}$ and $\boldsymbol{\alpha}_B^{k+u+1}$ is a stationary point of the sub-problem satisfying the above inequality. If

$$i \in I_{\text{up}}(\boldsymbol{\alpha}^{k+u+1}) \text{ and } j \in I_{\text{low}}(\boldsymbol{\alpha}^{k+u+1}),$$

then from (32) and (18)

$$\begin{aligned} & -y_i \nabla f(\boldsymbol{\alpha}^{k+u+1})_i \\ \leq \; & -y_j \nabla f(\boldsymbol{\alpha}^{k+u+1})_j + \frac{y_i(\tau - a)}{2}(\alpha_i^{k+u+1} - \alpha_i^{k+u}) - \frac{y_j(\tau - a)}{2}(\alpha_j^{k+u+1} - \alpha_j^{k+u}) \\ \leq \; & -y_j \nabla f(\boldsymbol{\alpha}^k)_j + \frac{\tau - a}{\sqrt{2}} \|\boldsymbol{\alpha}^{k+u+1} - \boldsymbol{\alpha}^{k+u}\| \\ \leq \; & -y_j \nabla f(\boldsymbol{\alpha}^k)_j + \frac{\Delta'}{\sqrt{2}}. \end{aligned} \tag{34}$$

However, $-y_i \nabla f(\bar{\boldsymbol{\alpha}})_i > -y_j \nabla f(\bar{\boldsymbol{\alpha}})_j$ implies (30) for $\boldsymbol{\alpha}^{k+u+1}$, so there is a contradiction. If $a > 0$ and the sub-problem (2) is considered, (34) has no term $\frac{\Delta'}{\sqrt{2}}$, so immediately we have a contradiction to (30).

For the convenience of writing the proof, we reorder indices of $\bar{\boldsymbol{\alpha}}$ so that

$$-y_1 \nabla f(\bar{\boldsymbol{\alpha}})_1 \leq \cdots \leq -y_l \nabla f(\bar{\boldsymbol{\alpha}})_l. \tag{35}$$

We also define

$$S_1(k) \equiv \sum \{i \mid i \in I_{\text{up}}(\boldsymbol{\alpha}^k)\} \text{ and } S_2(k) \equiv \sum \{l - i \mid i \in I_{\text{low}}(\boldsymbol{\alpha}^k)\}. \tag{36}$$

Clearly,

$$l \leq S_1(k) + S_2(k) \leq l(l-1). \tag{37}$$

If $\{i, j\}$ is selected at the $(k+u)$th iteration ($u = 0, \ldots, r$), then we claim that

$$-y_i \nabla f(\bar{\boldsymbol{\alpha}})_i > -y_j \nabla f(\bar{\boldsymbol{\alpha}})_j. \tag{38}$$

It is impossible that $-y_i \nabla(\bar{\boldsymbol{\alpha}})_i < -y_j \nabla(\bar{\boldsymbol{\alpha}})_j$ as $-y_i \nabla(\boldsymbol{\alpha}^{k+u})_i < -y_j \nabla(\boldsymbol{\alpha}^{k+u})_j$ from (30) then violates (9). If $-y_i \nabla(\bar{\boldsymbol{\alpha}})_i = -y_j \nabla(\bar{\boldsymbol{\alpha}})_j$, then

$$\begin{aligned} -y_i \nabla(\boldsymbol{\alpha}^{k+u})_i + y_j \nabla(\boldsymbol{\alpha}^{k+u})_j \quad & < \quad h(\Delta') < h(-y_{\bar{i}} \nabla f(\boldsymbol{\alpha}^{k+u})_{\bar{i}} + y_{\bar{j}} \nabla f(\boldsymbol{\alpha}^{k+u})_{\bar{j}}) \\ & \leq \quad h(m(\boldsymbol{\alpha}^{k+u}) - M(\boldsymbol{\alpha}^{k+u})). \end{aligned} \tag{39}$$

With the property that h is strictly increasing, the first two inequalities come from (31) and (27), while the last is from $\bar{i} \in I_{\text{up}}(\bar{\boldsymbol{\alpha}}), \bar{j} \in I_{\text{low}}(\bar{\boldsymbol{\alpha}})$, (28), and (29). Clearly, (39) contradicts (9), so (38) is valid.

Next we use a counting procedure to obtain the contradiction of (24) and (25). From the kth to the $(k+1)$st iteration, (38) and then (33) show that

$$i \notin I_{\text{up}}(\boldsymbol{\alpha}^{k+1}) \text{ or } j \notin I_{\text{low}}(\boldsymbol{\alpha}^{k+1}).$$

For the first case, (28) implies $i \notin I_{\text{up}}(\bar{\boldsymbol{\alpha}})$ and hence $i \in I_{\text{low}}(\bar{\boldsymbol{\alpha}})$. From (29) and the selection rule (9), $i \in I_{\text{low}}(\boldsymbol{\alpha}^k) \cap I_{\text{up}}(\boldsymbol{\alpha}^k)$. Thus,

$$i \in I_{\text{low}}(\boldsymbol{\alpha}^k) \cap I_{\text{up}}(\boldsymbol{\alpha}^k) \text{ and } i \notin I_{\text{up}}(\boldsymbol{\alpha}^{k+1}).$$

With $j \in I_{\text{low}}(\boldsymbol{\alpha}^k)$,

$$S_1(k+1) \le S_1(k) - i + j \le S_1(k) - 1, \qquad S_2(k+1) \le S_2(k), \tag{40}$$

where $-i + j \le -1$ comes from (35). Similarly, for the second case,

$$j \in I_{\text{low}}(\boldsymbol{\alpha}^k) \cap I_{\text{up}}(\boldsymbol{\alpha}^k) \text{ and } j \notin I_{\text{low}}(\boldsymbol{\alpha}^{k+1}).$$

With $i \in I_{\text{up}}(\boldsymbol{\alpha}^k)$,

$$S_1(k+1) \le S_1(k), \qquad S_2(k+1) \le S_2(k) - (l-j) + (l-i) \le S_2(k) - 1. \tag{41}$$

From iteration $(k+1)$ to $(k+2)$, we can repeat the same argument. Note that (33) can still be used because (38) holds for working sets selected during iterations k to $k+r$. Using (40) and (41), at $r \equiv l(l-1)$ iterations, $S_1(k) + S_2(k)$ is reduced to zero, a contradiction to (37). Therefore, the assumptions (24) and (25) are wrong and the proof is complete. ■

Moreover, if (1) has a unique optimal solution, then the whole sequence $\{\boldsymbol{\alpha}^k\}$ globally converges. This happens, for example, when Q is PD.

Corollary 1 *If Q is positive definite, $\{\boldsymbol{\alpha}^k\}$ globally converges to the unique minimum of* (1)*.*

Proof: Since Q is positive definite, (1) has a unique solution and we denote it as $\bar{\boldsymbol{\alpha}}$. Assume $\{\boldsymbol{\alpha}^k\}$ does not globally converge to $\bar{\boldsymbol{\alpha}}$. Then there is $\epsilon > 0$ and an infinite subset $\mathcal{K}$ such that $\|\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}\| \ge \epsilon, \forall k \in \mathcal{K}$. Since $\{\boldsymbol{\alpha}^k\}, k \in \mathcal{K}$ are in a compact space, there is a further sub-sequence converging to $\boldsymbol{\alpha}^*$ and $\|\boldsymbol{\alpha}^* - \bar{\boldsymbol{\alpha}}\| \ge \epsilon$. Since $\boldsymbol{\alpha}^*$ is an optimal solution according to Theorem 3, this contradicts that $\bar{\boldsymbol{\alpha}}$ is the unique global minimum. ■

For example, if the RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2}$ is used and all $\mathbf{x}_i \ne \mathbf{x}_j$, then Q is positive definite [22] and we have the result of Corollary 1.

We now discuss the difference between the proof above and earlier convergence work. The work [16] considers a working set selection which allows more than two elements. When the size of the working set is restricted to two, the selection is reduced to WSS 1, the maximal violating pair. This proof in [17] has used a counting procedure by considering two sets related to $\{\bar{i}, \bar{j}\}$, the maximal violating pair at $\bar{\boldsymbol{\alpha}}$:

$$\begin{aligned} I_1(k) &\equiv \{t \mid t \in I_{\text{up}}(\boldsymbol{\alpha}^k), -y_t \nabla f(\bar{\boldsymbol{\alpha}})_t \ge -y_{\bar{j}} \nabla f(\bar{\boldsymbol{\alpha}})_{\bar{j}}\}, \text{ and} \\ I_2(k) &\equiv \{t \mid t \in I_{\text{low}}(\boldsymbol{\alpha}^k), -y_t \nabla f(\bar{\boldsymbol{\alpha}})_t \le -y_{\bar{i}} \nabla f(\bar{\boldsymbol{\alpha}})_{\bar{i}}\}. \end{aligned}$$

Clearly, if WSS 1 is used, $\bar{i} \in I_{\text{up}}(\boldsymbol{\alpha}^k)$ and $\bar{j} \in I_{\text{low}}(\boldsymbol{\alpha}^k)$ imply that the selected $\{i, j\}$ must satisfy $i \in I_1(k)$ and $j \in I_2(k)$. Using (33) to show that $|I_1(k)| + |I_2(k)|$ decreases to zero, we obtain a contradiction to the fact that $|I_1(k)| + |I_2(k)| \ge 2$ from $\bar{i} \in I_{\text{up}}(\boldsymbol{\alpha}^k)$ and $\bar{j} \in I_{\text{low}}(\boldsymbol{\alpha}^k)$. However, now we do not have $i \in I_1(k)$ and $j \in I_2(k)$ any more since our selection may not be the

maximal violating pair. Therefore, a new counting scheme is developed. By arranging $-y_i \nabla f(\bar{\boldsymbol{\alpha}})_i$ in an ascending (descending) order, in (36), we count the sum of their ranks.

IV. Stopping Condition, Shrinking and Caching

In this section we discuss other important properties of our method. In previous sections, Q is any symmetric matrix, but here we further require it to be positive semi-definite. To begin, the following theorem depicts some facts about the problem (1).

Theorem 4 *Assume Q is positive semi-definite.*

1) *If $\bar{\boldsymbol{\alpha}} \neq \hat{\boldsymbol{\alpha}}$ are any two optimal solutions of* (1)*, then*

$$-y_i \nabla f(\bar{\boldsymbol{\alpha}})_i = -y_i \nabla f(\hat{\boldsymbol{\alpha}})_i, i = 1, \ldots, l, \tag{42}$$

and

$$M(\bar{\boldsymbol{\alpha}}) = m(\bar{\boldsymbol{\alpha}}) = M(\hat{\boldsymbol{\alpha}}) = m(\hat{\boldsymbol{\alpha}}). \tag{43}$$

2) *If there is an optimal solution $\bar{\boldsymbol{\alpha}}$ satisfying $m(\bar{\boldsymbol{\alpha}}) < M(\bar{\boldsymbol{\alpha}})$, then $\bar{\boldsymbol{\alpha}}$ is the unique optimal solution of* (1)*.*
3) *The following set is independent of any optimal solution $\bar{\boldsymbol{\alpha}}$:*

$$I \equiv \{i \mid -y_i \nabla f(\bar{\boldsymbol{\alpha}})_i > M(\bar{\boldsymbol{\alpha}}) \text{ or } -y_i \nabla f(\bar{\boldsymbol{\alpha}})_i < m(\bar{\boldsymbol{\alpha}})\}. \tag{44}$$

Moreover, problem (1) *has unique and bounded optimal solutions at α_i, $i \in I$.*

Proof: Since Q is positive semi-definite, (1) is a convex programming problem and $\bar{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\alpha}}$ are both global optima. Then

$$f(\bar{\boldsymbol{\alpha}}) = f(\hat{\boldsymbol{\alpha}}) = f(\lambda \bar{\boldsymbol{\alpha}} + (1-\lambda)\hat{\boldsymbol{\alpha}}), \text{ for all } 0 \leq \lambda \leq 1,$$

implies

$$(\bar{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}})^T Q (\bar{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}) = 0.$$

As Q is PSD, Q can be factorized to LL^T. Thus, $\|L^T(\bar{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}})\| = 0$ and hence $Q\bar{\boldsymbol{\alpha}} = Q\hat{\boldsymbol{\alpha}}$. Then (42) follows.

To prove (43), we will show that

$$m(\hat{\boldsymbol{\alpha}}) \geq M(\bar{\boldsymbol{\alpha}}) \text{ and } m(\bar{\boldsymbol{\alpha}}) \geq M(\hat{\boldsymbol{\alpha}}). \tag{45}$$

With optimality conditions $M(\bar{\boldsymbol{\alpha}}) \geq m(\bar{\boldsymbol{\alpha}})$ and $M(\hat{\boldsymbol{\alpha}}) \geq m(\hat{\boldsymbol{\alpha}})$, (43) holds.

Due to the symmetry, it is sufficient to show the first case of (45). If it is wrong, then

$$m(\hat{\boldsymbol{\alpha}}) < M(\bar{\boldsymbol{\alpha}}). \tag{46}$$

We then investigate different situations by comparing $-y_i \nabla f(\bar{\boldsymbol{\alpha}})_i$ with $M(\bar{\boldsymbol{\alpha}})$ and $m(\hat{\boldsymbol{\alpha}})$. If $-y_i \nabla f(\bar{\boldsymbol{\alpha}})_i \geq M(\bar{\boldsymbol{\alpha}}) > m(\hat{\boldsymbol{\alpha}})$, then $i \notin I_{\text{up}}(\hat{\boldsymbol{\alpha}})$ and

$$\hat{\alpha}_i = \begin{cases} 0 & \text{if } y_i = -1, \\ C & \text{if } y_i = 1. \end{cases} \tag{47}$$

With $0 \leq \bar{\alpha}_i \leq C$,

$$y_i \hat{\alpha}_i - y_i \bar{\alpha}_i \geq 0. \tag{48}$$

If $-y_i \nabla f(\bar{\boldsymbol{\alpha}})_i \leq m(\hat{\boldsymbol{\alpha}}) < M(\bar{\boldsymbol{\alpha}})$, then $i \notin I_{\text{low}}(\bar{\boldsymbol{\alpha}})$ and

$$\bar{\alpha}_i = \begin{cases} C & \text{if } y_i = -1, \\ 0 & \text{if } y_i = 1. \end{cases} \tag{49}$$

Hence, (48) still holds.

Other indices are in the following set

$$S \equiv \{i \mid m(\hat{\boldsymbol{\alpha}}) < -y_i \nabla f(\hat{\boldsymbol{\alpha}})_i = -y_i \nabla f(\bar{\boldsymbol{\alpha}})_i < M(\bar{\boldsymbol{\alpha}})\}.$$

If $i \in S$, then $i \notin I_{\text{up}}(\hat{\boldsymbol{\alpha}})$ and $i \notin I_{\text{low}}(\bar{\boldsymbol{\alpha}})$. Hence (47) and (49) imply

$$y_i \hat{\alpha}_i - y_i \bar{\alpha}_i = C. \tag{50}$$

Thus,

$$\begin{aligned} 0 &= \mathbf{y}^T \hat{\boldsymbol{\alpha}} - \mathbf{y}^T \bar{\boldsymbol{\alpha}} \\ &= \sum_{i: i \notin S} (y_i \hat{\alpha}_i - y_i \bar{\alpha}_i) + C|S|. \end{aligned}$$

Since (48) implies each term in the above summation is non-negative,

$$|S| = 0 \text{ and } \hat{\alpha}_i = \bar{\alpha}_i, \forall i \notin S.$$

Therefore, $\bar{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}$. However, this contradicts the assumption that $\bar{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\alpha}}$ are different optimal solutions. Hence (46) is wrong and we have $m(\hat{\boldsymbol{\alpha}}) \geq M(\bar{\boldsymbol{\alpha}})$ in (45). The proof of (43) is thus complete.

The second result of this theorem and the validity of the set I directly come from (43). Moreover, I is independent of any optimal solution.

For any optimal $\boldsymbol{\alpha}$, if $i \in I$ and $-y_i \nabla f(\boldsymbol{\alpha})_i > M(\boldsymbol{\alpha}) \geq m(\boldsymbol{\alpha})$, then, $i \notin I_{\text{up}}(\boldsymbol{\alpha})$ and α_i is the same as (47). Thus, the optimal α_i is unique and bounded. The situation for $-y_i \nabla f(\boldsymbol{\alpha})_i < m(\boldsymbol{\alpha})$ is similar. ■

Lemma 3 in [12] shows a result similar to (43), but the proof is more complicated. It involves both primal and dual SVM formulations. Using Theorem 4, in the rest of this section we derive more properties of the proposed SMO-type methods.

A. Stopping Condition and Finite Termination

As the decomposition method only asymptotically approaches an optimum, in practice, it terminates after satisfying a stopping condition. For example, we can pre-specify a small stopping tolerance $\epsilon > 0$ and check if

$$m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k) \leq \epsilon \tag{51}$$

is satisfied. Though one can consider other stopping conditions, (51) is commonly used due to its closeness to the optimality condition (6). To justify its validity as a stopping condition, we must make sure that under any given stopping tolerance $\epsilon > 0$, the proposed decomposition method stops in a finite number of iterations. Thus, an infinite loop never happens. To have (51), one can prove a stronger condition:

$$\lim_{k\to\infty} m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k) = 0. \tag{52}$$

This condition is not readily available as from the respective definitions of $m(\boldsymbol{\alpha})$ and $M(\boldsymbol{\alpha})$, it is unclear whether they are continuous functions of $\boldsymbol{\alpha}$. Proving (52) will be the main result of this subsection.

The first study on the stopping condition of SMO-type methods is [11]. They consider a selection rule which involves the stopping tolerance, so the working set is a so-called $\epsilon - violating\, pair$. Since our selection rule is independent of ϵ, their analysis cannot be applied here. Another work [18] proves (52) for WSS 1 under the assumption that the sequence $\{\boldsymbol{\alpha}^k\}$ globally converges. Here, for more general selection WSS 3, we prove (52) with positive semi-definite Q.

Theorem 5 *Assume Q is positive semi-definite and the SMO-type decomposition method Algorithm 2 using WSS 3 generates an infinite sequence $\{\boldsymbol{\alpha}^k\}$. Then*

$$\lim_{k\to\infty} m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k) = 0. \tag{53}$$

Proof: We prove the theorem by contradiction and assume that the result (53) is wrong. Then, there is an infinite set $\bar{\mathcal{K}}$ and a value $\Delta > 0$ such that

$$|m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k)| \geq \Delta, \forall k \in \bar{\mathcal{K}}. \tag{54}$$

Since in the decomposition method $m(\boldsymbol{\alpha}^k) > M(\boldsymbol{\alpha}^k), \forall k$, (54) can be rewritten as

$$m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k) \geq \Delta, \forall k \in \bar{\mathcal{K}}. \tag{55}$$

In this $\bar{\mathcal{K}}$ there is a further sub-sequence $\mathcal{K}$ so that

$$\lim_{k\in\mathcal{K}, k\to\infty} \boldsymbol{\alpha}^k = \bar{\boldsymbol{\alpha}}.$$

As Q is positive semi-definite, from Theorem 4, $\nabla f(\boldsymbol{\alpha}^k)$ globally converges:

$$\lim_{k\to\infty} \nabla f(\boldsymbol{\alpha}^k)_i = \nabla f(\bar{\boldsymbol{\alpha}})_i, i = 1, \ldots, l. \tag{56}$$

We then follow a similar counting strategy in Theorem 3. First we rewrite (55) as

$$m(\boldsymbol{\alpha}^k) \geq M(\boldsymbol{\alpha}^k) + \Delta', \forall k \in \bar{\mathcal{K}}, \tag{57}$$

where

$$\Delta' \equiv \min\left(\Delta, \frac{1}{2}\min\left\{|-y_t\nabla f(\bar{\boldsymbol{\alpha}})_t + y_s\nabla f(\bar{\boldsymbol{\alpha}})_s| \,\middle|\, -y_t\nabla f(\bar{\boldsymbol{\alpha}})_t \neq -y_s\nabla f(\bar{\boldsymbol{\alpha}})_s\right\}\right). \tag{58}$$

We still require (27)-(33), but use (56) and (58) to extend (30) and (31) for all $k \geq \bar{k}$ (i.e., not only $k \in \mathcal{K}$):

$$\text{If } -y_t\nabla f(\bar{\boldsymbol{\alpha}})_t > -y_s\nabla f(\bar{\boldsymbol{\alpha}})_s, \text{ then } -y_t\nabla f(\boldsymbol{\alpha}^k)_t > y_s\nabla f(\boldsymbol{\alpha}^k)_s. \tag{59}$$

$$\text{If } -y_t\nabla f(\bar{\boldsymbol{\alpha}})_t \neq -y_s\nabla f(\bar{\boldsymbol{\alpha}})_s, \text{ then } |-y_t\nabla f(\boldsymbol{\alpha}^k)_t + y_s\nabla f(\boldsymbol{\alpha}^k)_s| > \Delta'. \tag{60}$$

$$\text{If } -y_t\nabla f(\bar{\boldsymbol{\alpha}})_t = -y_s\nabla f(\bar{\boldsymbol{\alpha}})_s, \text{ then } |-y_t\nabla f(\boldsymbol{\alpha}^k)_t + y_s\nabla f(\boldsymbol{\alpha}^k)_s| < h(\Delta'). \tag{61}$$

Then the whole proof follows Theorem 3 except (39), in which we need $m(\boldsymbol{\alpha}^{k+u}) - M(\boldsymbol{\alpha}^{k+u}) \geq \Delta', \forall u = 0, \ldots, r$. This condition does not follow from (57), which holds for only a sub-sequence. Thus, in the following we further prove

$$m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k) \geq \Delta', \forall k \geq \bar{k}. \tag{62}$$

Assume at one $k' \geq \bar{k}$, $0 < m(\boldsymbol{\alpha}^{k'}) - M(\boldsymbol{\alpha}^{k'}) < \Delta'$ and $\{i, j\}$ is the working set of this iteration. As $i \in I_{\text{up}}(\boldsymbol{\alpha}^{k'})$ and $j \in I_{\text{low}}(\boldsymbol{\alpha}^{k'})$ from the selection rule, we have

$$M(\boldsymbol{\alpha}^{k'}) \leq -y_j\nabla f(\boldsymbol{\alpha}^{k'})_j < -y_i\nabla f(\boldsymbol{\alpha}^{k'})_i \leq m(\boldsymbol{\alpha}^{k'}). \tag{63}$$

Then according to (60), $\{i, j\}$ and indices achieving $m(\boldsymbol{\alpha}^{k'})$ and $M(\boldsymbol{\alpha}^{k'})$ have the same value of $-y_t\nabla f(\bar{\boldsymbol{\alpha}})_t$. They are all from the following set:

$$\{t \mid -y_t\nabla f(\bar{\boldsymbol{\alpha}})_t = -y_i\nabla f(\bar{\boldsymbol{\alpha}})_i = -y_j\nabla f(\bar{\boldsymbol{\alpha}})_j\}. \tag{64}$$

For elements not in this set, (59), (60), and (63) imply that

$$\begin{aligned}&\text{If } -y_t\nabla f(\bar{\boldsymbol{\alpha}})_t > -y_i\nabla f(\bar{\boldsymbol{\alpha}})_i, \text{ then}\\ &-y_t\nabla f(\boldsymbol{\alpha}^{k'}) > -y_i\nabla f(\boldsymbol{\alpha}^{k'})_i + \Delta' > m(\boldsymbol{\alpha}^{k'}) \text{ and hence } t \notin I_{\text{up}}(\boldsymbol{\alpha}^{k'}).\end{aligned} \tag{65}$$

Similarly,

$$\text{If } -y_t\nabla f(\bar{\boldsymbol{\alpha}})_t < -y_i\nabla f(\bar{\boldsymbol{\alpha}})_i, \text{ then } t \notin I_{\text{low}}(\boldsymbol{\alpha}^{k'}). \tag{66}$$

As we have explained, the working set is from the set (64), other components remain the same from iteration k' to $k'+1$. Therefore, indices satisfying (65) and (66) have $t \notin I_{\text{up}}(\boldsymbol{\alpha}^{k'+1})$ and $t \notin I_{\text{low}}(\boldsymbol{\alpha}^{k'+1})$, respectively. Furthermore, indices in (65) have larger $-y_t\nabla f(\boldsymbol{\alpha}^{k'+1})_t$ than others according to (59). Thus, their $-y_t\nabla f(\boldsymbol{\alpha}^{k'+1})_t$ are greater than $m(\boldsymbol{\alpha}^{k'+1})$. Similarly, elements in (66) are smaller than $M(\boldsymbol{\alpha}^{k'+1})$. With the fact that $m(\boldsymbol{\alpha}^{k'+1}) > M(\boldsymbol{\alpha}^{k'+1})$, indices which achieve

$m(\boldsymbol{\alpha}^{k'+1})$ and $M(\boldsymbol{\alpha}^{k'+1})$ are again from the set (64). This situation holds for all $k \geq k'$. Using (61) and the condition on h, we have

$$m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k) < h(\Delta') \leq \Delta', \forall k \geq k',$$

a contradiction to (57). Thus, the required condition (62) holds. ■

B. Shrinking and Caching

Shrinking and caching are two effective techniques to make the decomposition method faster. If an α_i^k remains at 0 or C for many iterations, eventually it may stay at the same value. Based on this principle, the shrinking technique [10] reduces the size of the optimization problem without considering some bounded variables. The decomposition method then works on a smaller problem which is less time consuming and requires less memory. In the end we put back the removed components and check if an optimal solution of the original problem is obtained.

Another technique to reduce the training time is caching. Since Q may be too large to be stored, its elements are calculated as they are needed. We can allocate some space (called cache) to store recently use Q_{ij} [10]. If during final iterations only few columns of Q are still needed and the cache contains them, we can save many kernel evaluations. [18, Theorem II.3] has proved that during the final iterations of using WSS 1, only a small subset of variables are still updated. Such a result supports the use of shrinking and caching techniques. However, this proof considers only any convergent subsequence of $\{\boldsymbol{\alpha}^k\}$ or assumes the global convergence. In this subsection, we provide a more general theory without these two assumptions.

Theorem 6 *Assume Q is positive semi-definite and the SMO-type method Algorithm 2 uses WSS 3. Let I be the set defined in* (44)*.*

1) *There is $\bar{k}$ such that after $k \geq \bar{k}$, every $\alpha_i^k, i \in I$ has reached the unique and bounded optimal solution. It remains the same during all subsequent iterations and $i \in I$ is not in the following set:*

$$\{t \mid M(\boldsymbol{\alpha}^k) \leq -y_t \nabla f(\boldsymbol{\alpha}^k)_t \leq m(\boldsymbol{\alpha}^k)\}. \tag{67}$$

2) *If* (1) *has an optimal solution $\bar{\boldsymbol{\alpha}}$ satisfying $m(\bar{\boldsymbol{\alpha}}) < M(\bar{\boldsymbol{\alpha}})$, then $\bar{\boldsymbol{\alpha}}$ is the unique solution and the decomposition method reaches it at a finite number of iterations.*

3) *If $\{\boldsymbol{\alpha}^k\}$ is an infinite sequence, then the following two limits exist and are equal:*

$$\lim_{k\to\infty} m(\boldsymbol{\alpha}^k) = \lim_{k\to\infty} M(\boldsymbol{\alpha}^k) = m(\bar{\boldsymbol{\alpha}}) = M(\bar{\boldsymbol{\alpha}}), \tag{68}$$

where $\bar{\boldsymbol{\alpha}}$ is any optimal solution. Thus, (68) *is independent of any optimal solution.*

Proof:

1) If the result is wrong, there is an index $\bar{i} \in I$ and an infinite set $\hat{\mathcal{K}}$ such that

$$\alpha_{\bar{i}}^k \neq \hat{\alpha}_{\bar{i}}, \forall k \in \hat{\mathcal{K}}, \tag{69}$$

where $\hat{\boldsymbol{\alpha}}_{\bar{i}}$ is the unique optimal component according to Theorem 4. From Theorem 3, there is a further subset $\mathcal{K}$ of $\hat{\mathcal{K}}$ such that

$$\lim_{k \in \mathcal{K}, k \to \infty} \boldsymbol{\alpha}^k = \bar{\boldsymbol{\alpha}} \tag{70}$$

is a stationary point. Moreover, Theorem 4 implies that $\bar{\alpha}_i$, $i \in I$ are unique optimal components. Thus, $\bar{\alpha}_{\bar{i}} = \hat{\alpha}_{\bar{i}}$.

As $\bar{i} \in I$, we first consider one of the two possible situations:

$$-y_{\bar{i}} \nabla f(\bar{\boldsymbol{\alpha}})_{\bar{i}} > M(\bar{\boldsymbol{\alpha}}). \tag{71}$$

Thus, $\bar{i} \notin I_{\text{up}}(\bar{\boldsymbol{\alpha}})$. (69) then implies

$$\bar{i} \in I_{\text{up}}(\boldsymbol{\alpha}^k), \forall k \in \mathcal{K}. \tag{72}$$

For each $j \in \arg M(\bar{\boldsymbol{\alpha}})$, we have $j \in I_{\text{low}}(\bar{\boldsymbol{\alpha}})$. From (70), there is $\bar{k}$ such that

$$j \in I_{\text{low}}(\boldsymbol{\alpha}^k), \forall k \in \mathcal{K}, k \geq \bar{k}. \tag{73}$$

Thus, (72) and (73) imply

$$m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k) \geq -y_{\bar{i}} \nabla f(\boldsymbol{\alpha}^k)_{\bar{i}} + y_j \nabla f(\boldsymbol{\alpha}^k)_j, \forall k \in \mathcal{K}, k \geq \bar{k}. \tag{74}$$

With (70), the continuity of $\nabla f(\boldsymbol{\alpha})$, and (53), taking the limit on both sides of (74) obtains

$$0 \geq -y_{\bar{i}} \nabla f(\bar{\boldsymbol{\alpha}})_{\bar{i}} + y_j \nabla f(\bar{\boldsymbol{\alpha}})_j = -y_{\bar{i}} \nabla f(\bar{\boldsymbol{\alpha}})_{\bar{i}} - M(\bar{\boldsymbol{\alpha}}).$$

This inequality violates (71), so there is a contradiction. For the other situation $-y_{\bar{i}} \nabla f(\bar{\boldsymbol{\alpha}})_{\bar{i}} < m(\bar{\boldsymbol{\alpha}})$, the proof is the same.

The proof that $i \in I$ is not in the set (67) is similar. If the result is wrong, there is an index $\bar{i} \in I$ such that $\forall k \in \mathcal{K}$, $\bar{i}$ is in the set (67). Then (74) holds and causes a contradiction.

2) If the result does not hold, then $\{\boldsymbol{\alpha}^k\}$ is an infinite sequence. From Theorems 3 and 4, $\bar{\boldsymbol{\alpha}}$ is the unique optimal solution and $\{\boldsymbol{\alpha}^k\}$ globally converges to it.

Define

$$I_1 \equiv \{i \mid -y_i \nabla f(\bar{\boldsymbol{\alpha}})_i = M(\bar{\boldsymbol{\alpha}})\},$$

$$I_2 \equiv \{i \mid -y_i \nabla f(\bar{\boldsymbol{\alpha}})_i = m(\bar{\boldsymbol{\alpha}})\}.$$

Using the first result of this theorem, after k is sufficiently large, $\arg m(\boldsymbol{\alpha}^k)$ and $\arg M(\boldsymbol{\alpha}^k)$ must be subsets of $I_1 \cup I_2$. Moreover, using (53), the continuity of $\nabla f(\boldsymbol{\alpha})$, and the property $\lim_{k \to \infty} \boldsymbol{\alpha}^k = \bar{\boldsymbol{\alpha}}$, there is $\bar{k}$ such that for all $k \geq \bar{k}$,

$$\arg m(\boldsymbol{\alpha}^k) \text{ and } \arg M(\boldsymbol{\alpha}^k) \text{ are both subsets of } I_1 \text{ (or } I_2). \tag{75}$$

If at the kth iteration, $\arg m(\boldsymbol{\alpha}^k)$ and $\arg M(\boldsymbol{\alpha}^k)$ are both subsets of I_1, then following the same argument in (63)-(64), we have

$$\text{the working set } B \subset I_1. \tag{76}$$

As the decomposition method maintains feasibility,

$$\sum_{i\in B} y_i\alpha_i^k = \sum_{i\in B} y_i\alpha_i^{k+1}. \tag{77}$$

From (76) and the assumption that $m(\bar{\boldsymbol{\alpha}}) < M(\bar{\boldsymbol{\alpha}})$, every $\bar{\alpha}_i$, $i \in B$ satisfies $i \notin I_{\rm up}(\boldsymbol{\alpha})$. Thus, $\bar{\alpha}_i$, $i \in B$ is the same as (47). This and (77) then imply

$$\begin{aligned}
& \|\boldsymbol{\alpha}^{k+1} - \bar{\boldsymbol{\alpha}}\|_1 \\
= & \sum_{i\notin B} |\alpha_i^{k+1} - \bar{\alpha}_i| + \sum_{i\in B, y_i=1} (C - \alpha_i^{k+1}) + \sum_{i\in B, y_i=-1} (\alpha_i^{k+1} - 0) \\
= & \sum_{i\notin B} |\alpha_i^{k} - \bar{\alpha}_i| + \sum_{i\in B, y_i=1} (C - \alpha_i^{k}) + \sum_{i\in B, y_i=-1} (\alpha_i^{k} - 0) \\
= & \|\boldsymbol{\alpha}^{k} - \bar{\boldsymbol{\alpha}}\|_1.
\end{aligned} \tag{78}$$

If $\arg m(\boldsymbol{\alpha}^k)$ and $\arg M(\boldsymbol{\alpha}^k)$ are both subsets of I_2, (78) still holds. Therefore,

$$0 \neq \|\boldsymbol{\alpha}^{k} - \bar{\boldsymbol{\alpha}}\|_1 = \|\boldsymbol{\alpha}^{k+1} - \bar{\boldsymbol{\alpha}}\|_1 = \cdots,$$

a contradiction to the fact that $\{\boldsymbol{\alpha}^k\}$ converges to $\bar{\boldsymbol{\alpha}}$. Thus, the decomposition method must stop in a finite number of iterations.

3) Since $\{\boldsymbol{\alpha}^k\}$ is an infinite sequence, using the result of 2), problem (1) has no optimal solution $\bar{\boldsymbol{\alpha}}$ satisfying $M(\bar{\boldsymbol{\alpha}}) > m(\bar{\boldsymbol{\alpha}})$. From Theorem 4, we have

$$M(\bar{\boldsymbol{\alpha}}) = m(\bar{\boldsymbol{\alpha}}) = -y_t \nabla f(\bar{\boldsymbol{\alpha}})_t, \forall t \notin I, \tag{79}$$

and this is independent of different optimal solutions $\bar{\boldsymbol{\alpha}}$. From the result of 1), there is $\bar{k}$ such that for all $k \geq \bar{k}$, $i \in I$ is not in the set (67). Thus,

$$I' \equiv \{1, \ldots, l\} \setminus I \tag{80}$$

is a superset of (67) and hence

$$\min_{i\in I'} -y_i \nabla f(\boldsymbol{\alpha}^k)_i \leq M(\boldsymbol{\alpha}^k) < m(\boldsymbol{\alpha}^k) \leq \max_{i\in I'} -y_i \nabla f(\boldsymbol{\alpha}^k)_i. \tag{81}$$

Though $\{\boldsymbol{\alpha}^k\}$ may not be globally convergent, $\{-y_i \nabla f(\boldsymbol{\alpha}^k)_i\}$, $i = 1, \ldots, l$, are according to (42). The limit of both sides of (81) are equal using (79), so (68) follows.

■

Theorem 6 implies that in many iterations, the SMO-type method involves only indices in I'. Thus, caching is very effective. This theorem also illustrates two possible shrinking implementations:

1) Elements not in the set (67) are removed.

2) Any α_i which has stayed at the same bound for a certain number of iterations is removed.

The software LIBSVM [3] considers the former approach, while SVM^{light} [10] uses the latter.

V. Convergence Rate

Though we proved the asymptotic convergence, it is important to investigate how fast the method converges. Under some assumptions, [15] was the first to prove the linear convergence of certain decomposition methods. The work [15] allows the working set to have more than two elements and WSS 1 is a special case. Here we show that when the SMO-type working set selection is extended from WSS 1 to WSS 2, the same analysis holds. Since [15] was not published, we include here all details of the proof.

Note that WSS 3 uses a function h to control the quality of the selected pair. We will see in the proof that it may affect the convergence rate. Proving the linear convergence requires the condition (8), so results established in this section are for WSS 2 but not WSS 3.

First we make a few assumptions.

Assumption 1 *Q is positive definite.*

Then (1) is a strictly convex programming problem and hence has a unique global optimum $\bar{\boldsymbol{\alpha}}$.

By Theorem 6, after large enough iterations working sets are all from the set I' defined in (80). From the optimality condition (6), the scalar $\bar{b}$ satisfies $\bar{b} = m(\bar{\boldsymbol{\alpha}}) = M(\bar{\boldsymbol{\alpha}})$, and the set I' corresponds to elements satisfying

$$\nabla f(\bar{\boldsymbol{\alpha}})_i + \bar{b} y_i = 0. \tag{82}$$

From (4)-(5), another form of the optimality condition, if $\bar{\alpha}_i$ is not at a bound, (82) holds. We further assume that this is the only case that (82) happens.

Assumption 2 (Nondegeneracy) *For the optimal solution $\bar{\boldsymbol{\alpha}}$, $\nabla f(\bar{\boldsymbol{\alpha}})_i + \bar{b} y_i = 0$ if and only if $0 < \bar{\alpha}_i < C$.*

This assumption, commonly used in optimization analysis, implies that indices of all bounded $\bar{\alpha}_i$ are exactly the set I. Therefore, after enough iterations, Theorem 6 and Assumption 2 imply that all bounded variables are fixed and are not included in the working set. By treating these variables as constants, essentially we solve a problem with the following form:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & f(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} + \mathbf{p}^T \boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{y}^T \boldsymbol{\alpha} = \Delta, \end{aligned} \tag{83}$$

where $\mathbf{p}$ is a vector by combining $-\mathbf{e}$ and other terms related to the bounded components. Moreover, $0 < \alpha_i^k < C$ for all i even though we do not explicitly write down inequality constraints in (83). Then

the optimal solution $\bar{\boldsymbol{\alpha}}$ with the corresponding $\bar{b}$ can be obtained by the following linear system:

$$\begin{bmatrix} Q & \mathbf{y} \\ \mathbf{y}^T & 0 \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{\alpha}} \\ \bar{b} \end{bmatrix} = \begin{bmatrix} -\mathbf{p} \\ \Delta \end{bmatrix}. \tag{84}$$

In each iteration, we consider minimizing $f(\boldsymbol{\alpha}_B^k + \mathbf{d})$, where $\mathbf{d}$ is the direction from $\boldsymbol{\alpha}_B^k$ to $\boldsymbol{\alpha}_B^{k+1}$, so the sub-problem (2) is written as

$$\begin{aligned} \min_{\mathbf{d}} \quad & \frac{1}{2}\mathbf{d}^T Q_{BB}\mathbf{d} + \nabla f(\boldsymbol{\alpha}^k)_B^T \mathbf{d} \\ \text{subject to} \quad & \mathbf{y}_B^T \mathbf{d} = 0, \end{aligned} \tag{85}$$

where $\nabla f(\boldsymbol{\alpha}^k) = Q\boldsymbol{\alpha}^k + \mathbf{p}$. If an optimal solution of (85) is $\mathbf{d}^k$, then $\boldsymbol{\alpha}_B^{k+1} = \boldsymbol{\alpha}_B^k + \mathbf{d}^k$ and $\boldsymbol{\alpha}_N^{k+1} = \boldsymbol{\alpha}_N^k$. With the corresponding b^k, this sub-problem is solved by the following equation:

$$\begin{bmatrix} Q_{BB} & \mathbf{y}_B \\ \mathbf{y}_B^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}^k \\ b^k \end{bmatrix} = \begin{bmatrix} -\nabla f(\boldsymbol{\alpha}^k)_B \\ 0 \end{bmatrix}. \tag{86}$$

Using (84),

$$\begin{aligned} Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}) &= Q\boldsymbol{\alpha}^k + \mathbf{p} + \bar{b}\mathbf{y} \\ &= \nabla f(\boldsymbol{\alpha}^k) + \bar{b}\mathbf{y}. \end{aligned} \tag{87}$$

By defining $Y \equiv \text{diag}(\mathbf{y})$ to be a diagonal matrix with elements of $\mathbf{y}$ on the diagonal, and using $y_i = \pm 1$,

$$-YQ(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}) = -Y\nabla f(\boldsymbol{\alpha}^k) - \bar{b}\mathbf{e}. \tag{88}$$

The purpose of checking $Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}})$ is to see how close the current solution is to the optimal one. Then (88) links it to $-Y\nabla f(\boldsymbol{\alpha}^k)$, a vector used for the working set selection. Remember that for finding violating pairs, we first sort $-y_i \nabla f(\boldsymbol{\alpha}^k)_i$ in an ascending order.

The following two theorems are the main results on the linear convergence.

Theorem 7 *Assume the SMO-type decomposition method uses WSS 2 for the working set selection. If problem* (1) *satisfies Assumptions 1 and 2, then there are $c < 1$ and $\bar{k}$ such that for all $k \geq \bar{k}$*

$$(\boldsymbol{\alpha}^{k+1} - \bar{\boldsymbol{\alpha}})^T Q(\boldsymbol{\alpha}^{k+1} - \bar{\boldsymbol{\alpha}}) \leq c(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}})^T Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}). \tag{89}$$

Proof: First, Theorem 6 implies that there is $\bar{k}$ such that after $k \geq \bar{k}$, the problem is reduced to

(83). We then directly calculate the difference between the $(k+1)$st and the kth iterations:

$$
\begin{aligned}
& (\boldsymbol{\alpha}^{k+1}-\bar{\boldsymbol{\alpha}})^T Q(\boldsymbol{\alpha}^{k+1}-\bar{\boldsymbol{\alpha}}) - (\boldsymbol{\alpha}^{k}-\bar{\boldsymbol{\alpha}})^T Q(\boldsymbol{\alpha}^{k}-\bar{\boldsymbol{\alpha}}) && (90)\\
= \;& 2(\mathbf{d}^k)^T(Q(\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}}))_B + (\mathbf{d}^k)^T Q_{BB}\mathbf{d}^k \\
= \;& (\mathbf{d}^k)^T(2(Q(\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}}))_B - \nabla f(\boldsymbol{\alpha}^k)_B - b^k\mathbf{y}_B) && (91)\\
= \;& (\mathbf{d}^k)^T((Q(\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}}))_B + (\bar{b}-b^k)\mathbf{y}_B) && (92)\\
= \;& (\mathbf{d}^k)^T((Q(\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}}))_B + (b^k-\bar{b})\mathbf{y}_B) && (93)\\
= \;& -[-(Q(\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}}))_B + (\bar{b}-b^k)\mathbf{y}_B]^T Q_{BB}^{-1}[-(Q(\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}}))_B + (\bar{b}-b^k)\mathbf{y}_B],
\end{aligned}
$$

where (91) is from (86), (92) is from (87), (93) is by using the fact $\mathbf{y}_B^T\mathbf{d}^k = 0$ from (86), and the last equality is from (86) and (87). If we define

$$\hat{Q} \equiv Y_B Q_{BB}^{-1} Y_B \text{ and } \mathbf{v} \equiv -Y(Q(\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}})), \tag{94}$$

where $Y_B \equiv \text{diag}(\mathbf{y}_B)$, then $\mathbf{v}_B = -Y_B(Q(\boldsymbol{\alpha}^k-\bar{\boldsymbol{\alpha}}))_B$ and (90) becomes

$$-[\mathbf{v}_B + (\bar{b}-b^k)\mathbf{e}_B]^T\hat{Q}[\mathbf{v}_B + (\bar{b}-b^k)\mathbf{e}_B]. \tag{95}$$

From (88), we define

$$
\begin{aligned}
v^1 &\equiv \max_t(v_t) = m(\boldsymbol{\alpha}^k) - \bar{b}, \\
v^l &\equiv \min_t(v_t) = M(\boldsymbol{\alpha}^k) - \bar{b}.
\end{aligned} \tag{96}
$$

Thus, the selection rule (8) of WSS 2 implies

$$|v_i - v_j| \geq \sigma(v^1 - v^l), \tag{97}$$

where $\{i, j\}$ is the working set of the kth iteration.

We denote that $\min(\text{eig}(\cdot))$ and $\max(\text{eig}(\cdot))$ to be the minimal and maximal eigenvalues of a matrix,

respectively. A further calculation of (95) shows

$$
\begin{aligned}
& [\mathbf{v}_B + (\bar{b} - b^k)\mathbf{e}_B]^T \hat{Q} [\mathbf{v}_B + (\bar{b} - b^k)\mathbf{e}_B] \\
\geq\ & \min(\text{eig}(\hat{Q}))[\mathbf{v}_B + (\bar{b} - b^k)\mathbf{e}_B]^T [\mathbf{v}_B + (\bar{b} - b^k)\mathbf{e}_B] \\
\geq\ & \min(\text{eig}(\hat{Q})) \max_{t \in B} (v_t + (\bar{b} - b_k))^2 \\
\geq\ & \min(\text{eig}(\hat{Q}))(\frac{v_i - v_j}{2})^2, \text{ where } \{i, j\} \text{ is working set} && (98) \\
\geq\ & \min(\text{eig}(\hat{Q}))\sigma^2(\frac{v^1 - v^l}{2})^2 && (99) \\
\geq\ & \min(\text{eig}(\hat{Q}))(\frac{\mathbf{y}^T Q^{-1}\mathbf{y}}{2\sum_{i,j} |Q_{ij}^{-1}|})^2 \sigma^2 \max(|v^1|, |v^l|)^2 && (100) \\
\geq\ & \frac{\min(\text{eig}(\hat{Q}))}{l}(\frac{\mathbf{y}^T Q^{-1}\mathbf{y}}{2\sum_{t,s} |Q_{ts}^{-1}|})^2 \sigma^2 (Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}))^T Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}) && (101) \\
\geq\ & \frac{\min(\text{eig}(\hat{Q}))}{l \max(\text{eig}(Q^{-1}))}(\frac{\mathbf{y}^T Q^{-1}\mathbf{y}}{2\sum_{t,s} |Q_{ts}^{-1}|})^2 \sigma^2 (Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}))^T Q^{-1} Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}) \\
\geq\ & \frac{\min(\text{eig}(\hat{Q}))}{l \max(\text{eig}(Q^{-1}))}(\frac{\mathbf{y}^T Q^{-1}\mathbf{y}}{2\sum_{t,s} |Q_{ts}^{-1}|})^2 \sigma^2 (\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}})^T Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}), && (102)
\end{aligned}
$$

where (98) is from Lemma 3, (99) is from (97), (100) is from Lemma 4, and (101) follows from (96). Note that both lemmas are given in Appendix A.

Next we give more details about the derivation of (100): If $v^1 v^l < 0$, then of course

$$|v^1 - v^l| \geq \max(|v^1|, |v^l|).$$

With $y_i = \pm 1$, $\frac{\mathbf{y}^T Q^{-1}\mathbf{y}}{\sum_{t,s} |Q_{ts}^{-1}|} \leq 1$ so (100) follows. In contrast, if $v^1 v^l \geq 0$, we consider $\mathbf{v} = (YQY)(-Y(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}))$ from (94). Since $-\mathbf{e}^T Y(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}) = -\mathbf{y}^T(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}) = 0$, we can apply Lemma 4: With

$$
\begin{aligned}
|(YQY)_{ij}^{-1}| = |Q_{ij}^{-1} y_i y_j| = |Q_{ij}^{-1}| \text{ and} \\
\mathbf{e}^T (YQY)^{-1} \mathbf{e} = \mathbf{y}^T Q^{-1} \mathbf{y},
\end{aligned}
$$

we have

$$
\begin{aligned}
|v^1 - v^l| &\geq \max(|v^1|, |v^l|) - \min(|v^1|, |v^l|) \\
&\geq (\frac{\mathbf{y}^T Q^{-1}\mathbf{y}}{\sum_{t,s} |Q_{ts}^{-1}|}) \max(|v^1|, |v^l|),
\end{aligned}
$$

which implies (100).

Finally we can define a constant c as follows:

$$c \equiv 1 - \min_B \left(\frac{\min(\text{eig}(Q_{BB}^{-1}))}{l \max(\text{eig}(Q^{-1}))}(\frac{\mathbf{y}^T Q^{-1}\mathbf{y}}{2\sum_{t,s} |Q_{ts}^{-1}|})^2 \sigma^2 \right) < 1,$$

where B is any two-element subset of $\{1, \ldots, l\}$. Combining (95) and (102), after $k \geq \bar{k}$, (89) holds.

■

The condition (8) of Algorithm 2 is used in (97) and then (99). If WSS 3 is considered, in (99) we will have a term $h((v^1 - v^l)/2)^2$. Thus, the function h affects the convergence rate. Since $h(x) \le x$, linear rate is the best result using our derivation.

The linear convergence of the objective function is as follows:

Theorem 8 *Under the same assumptions of Theorem 7, there are $c < 1$ and $\bar{k}$ such that for all $k \ge \bar{k}$,*

$$f(\boldsymbol{\alpha}^{k+1}) - f(\bar{\boldsymbol{\alpha}}) \le c(f(\boldsymbol{\alpha}^k) - f(\bar{\boldsymbol{\alpha}})).$$

Proof: We show that for any $k \ge \bar{k}$,

$$f(\boldsymbol{\alpha}^k) - f(\bar{\boldsymbol{\alpha}}) = \frac{1}{2}(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}})^T Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}),$$

and the proof immediately follows from Theorem 7. Using (84),

$$\begin{aligned}
& f(\boldsymbol{\alpha}^k) - f(\bar{\boldsymbol{\alpha}}) \\
= \; & \frac{1}{2}(\boldsymbol{\alpha}^k)^T Q\boldsymbol{\alpha}^k + \mathbf{p}^T\boldsymbol{\alpha}^k - \frac{1}{2}(\bar{\boldsymbol{\alpha}})^T Q\bar{\boldsymbol{\alpha}} - \mathbf{p}^T\bar{\boldsymbol{\alpha}} \\
= \; & \frac{1}{2}(\boldsymbol{\alpha}^k)^T Q\boldsymbol{\alpha}^k + (-Q\bar{\boldsymbol{\alpha}} - \bar{b}\mathbf{y})^T\boldsymbol{\alpha}^k - \frac{1}{2}(\bar{\boldsymbol{\alpha}})^T Q\bar{\boldsymbol{\alpha}} - (-Q\bar{\boldsymbol{\alpha}} - \bar{b}\mathbf{y})^T\bar{\boldsymbol{\alpha}} \\
= \; & \frac{1}{2}(\boldsymbol{\alpha}^k)^T Q\boldsymbol{\alpha}^k - (\bar{\boldsymbol{\alpha}})^T Q\boldsymbol{\alpha}^k + \frac{1}{2}(\bar{\boldsymbol{\alpha}})^T Q\bar{\boldsymbol{\alpha}} \qquad (103) \\
= \; & \frac{1}{2}(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}})^T Q(\boldsymbol{\alpha}^k - \bar{\boldsymbol{\alpha}}).
\end{aligned}$$

Since we always keep the feasibility of $\boldsymbol{\alpha}^k$, (103) comes from $\mathbf{y}^T\boldsymbol{\alpha}^k = \mathbf{y}^T\bar{\boldsymbol{\alpha}}$. ■

VI. Extensions

In this section we show that the same SMO-type methods can be applied to some variants of SVM.

A. Support Vector Regression and One-class SVM

First we extend (1) to the following general form:

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & f(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T Q\boldsymbol{\alpha} + \mathbf{p}^T\boldsymbol{\alpha} \\
\text{subject to} \quad & L_i \le \alpha_i \le U_i, i = 1, \ldots, l, \qquad (104) \\
& \mathbf{y}^T\boldsymbol{\alpha} = \Delta,
\end{aligned}$$

where $-\infty < L_i < U_i < \infty, i = 1, \ldots, l$, are lower and upper bounds, and Q is an l by l symmetric matrix. Clearly, if $L_i = 0, U_i = C,$ and $\mathbf{p} = -\mathbf{e}$, then (104) reduces to (1). The optimality condition is the same as (3) though in the definition of $I_{\rm up}(\boldsymbol{\alpha})$ and $I_{\rm low}(\boldsymbol{\alpha})$, 0 and C must be replaced by L_i and U_i, respectively. Therefore, SMO-type decomposition methods using WSS 2 or 3 can be applied to solve (104). A careful check shows that all results in Sections III-V hold for (104).

Problem (104) covers some SVM formulations such as support vector regression (SVR) [28] and one-class SVM [26]. Next we discuss SVR in detail. Given a set of data points $\{(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_l, z_l)\}$ such that $\mathbf{x}_i \in R^n$ is an input vector and $z_i \in R^1$ is a target output, support vector regression solves the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*} \quad & f(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T K (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} z_i (\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0, \\ & 0 \le \alpha_i, \alpha_i^* \le C, i = 1, \ldots, l, \end{aligned} \tag{105}$$

where $K_{ij} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$, and $\epsilon > 0$ is the width of the ϵ-insensitive tube.

We can rewrite (105) as

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*} \quad & f(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \frac{1}{2} \begin{bmatrix} \boldsymbol{\alpha}^T, \boldsymbol{\alpha}^{*T} \end{bmatrix} \begin{bmatrix} K & -K \\ -K & K \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \end{bmatrix} + \begin{bmatrix} \epsilon \mathbf{e}^T + \mathbf{z}^T, \epsilon \mathbf{e}^T - \mathbf{z}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \end{bmatrix} \\ \text{subject to} \quad & \mathbf{y}^T \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}^* \end{bmatrix} = 0, \\ & 0 \le \alpha_i, \alpha_i^* \le C, i = 1, \ldots, l, \end{aligned} \tag{106}$$

where $\mathbf{y}$ is a $2l$ by 1 vector with $y_i = 1, i = 1, \ldots, l$ and $y_i = -1, i = l+1, \ldots, 2l$. Thus (106) is in the form of (104) and an SMO-type method with WSS 3 can be applied. Moreover, the procedure asymptotically converges and possesses all properties in Section IV. An interesting issue is about Corollary 1, which requires the Hessian matrix $\left[\begin{smallmatrix} K & -K \\ -K & K \end{smallmatrix}\right]$ to be positive definite. This condition never holds as $\left[\begin{smallmatrix} K & -K \\ -K & K \end{smallmatrix}\right]$ is only positive semi-definite. Note that in Corollary 1, the positive definite Hessian is used to have a unique optimal solution. For SVR, [4, Lemma 4] proves that if $\epsilon > 0$ and K is positive definite, then (106) has a unique solution. Thus, for SVR, Corollary 1 can be modified to require only that K is positive definite.

For the linear convergence result in Section V, Assumption 1 does not hold as now $\left[\begin{smallmatrix} K & -K \\ -K & K \end{smallmatrix}\right]$ is not positive definite. However, we will show that similar to Corollary 1, a positive definite K is sufficient. Note that in Section V, the Hessian matrix of (83) is in fact $Q_{I'I'}$ as α_i, $i \in I$ can be removed after large enough iterations, where I and I' are defined as in (44) and (80) except that the set $\{1, \ldots, l\}$ is replaced by $\{1, \ldots, 2l\}$. Then in the linear convergence proof we need $Q_{I'I'}$ to be invertible and this condition holds if Q is positive definite (i.e., Assumption 1). For SVR, this means $\left[\begin{smallmatrix} K & -K \\ -K & K \end{smallmatrix}\right]_{I'I'}$ should be invertible. To prove it, we first claim that for any $1 \le i \le l$,

$$i \text{ and } i + l \text{ are not both in the set } I'. \tag{107}$$

As $\epsilon > 0$ implies

$$\begin{aligned} & -y_i \nabla f(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)_i = -(K(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*))_i + \epsilon + z_i \\ \neq \quad & -(K(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}))_i - \epsilon + z_i = -y_{i+l} \nabla f(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)_{i+l}, \end{aligned} \tag{108}$$

the definition of I' directly leads to (107). We further define

$$\bar{I}' \equiv \text{ an index vector by replacing any } t, l \le t \le 2l \text{ in } I' \text{ with } t - l.$$

From (107), I' and $\bar{I}'$ have the same number of elements and furthermore,

$$\begin{bmatrix} K & -K \\ -K & K \end{bmatrix}_{I'I'} = Y_{I'} K_{\bar{I}'\bar{I}'} Y_{I'}, \tag{109}$$

where $Y_{I'} \equiv \text{diag}(\mathbf{y}_{I'})$ is a diagonal matrix. Clearly, (109) indicates that if K is positive definite, so is $K_{\bar{I}'\bar{I}'}$ and $\left[\begin{smallmatrix} K & -K \\ -K & K \end{smallmatrix}\right]_{I'I'}$. Therefore, by replacing Assumption 1 on the Hessian matrix with that on the kernel matrix, the linear convergence result holds for SVR.

B. Extension to ν-SVM

ν-SVM [27] is another SVM formulation which has a parameter ν instead of C. Its dual form is

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & f(\boldsymbol{\alpha}) = \frac{1}{2} \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} \\ \text{subject to} \quad & \mathbf{y}^T \boldsymbol{\alpha} = 0, \\ & \mathbf{e}^T \boldsymbol{\alpha} = \nu, \\ & 0 \le \alpha_i \le 1/l, i = 1, \ldots, l, \end{aligned} \tag{110}$$

where $\mathbf{e}$ is the vector of all ones. Note that some use $\mathbf{e}^T \boldsymbol{\alpha} \ge \nu$ as a constraint, but [4] and [7] have shown that decision functions are the same. Moreover, [4] proved that (110) is equivalent to problem (1) with certain C. It also discusses the decomposition method for training ν-SVM.

Via the KKT condition, a vector $\bar{\boldsymbol{\alpha}}$ is a stationary point of (110) if and only if there are two scalars ρ and b such that

$$\begin{aligned} \nabla f(\boldsymbol{\alpha})_i - \rho + b y_i \ge 0 \qquad & \text{if } \alpha_i < 1/l, \\ \nabla f(\boldsymbol{\alpha})_i - \rho + b y_i \le 0 \qquad & \text{if } \alpha_i > 0. \end{aligned}$$

By separating the case of $y_i = 1$ and $y_i = -1$, we obtain two conditions on $-\rho + b$ and $-\rho - b$, respectively. Thus, similar to (6), there is the following optimality condition:

$$m_p(\boldsymbol{\alpha}) \le M_p(\boldsymbol{\alpha}) \text{ and } m_n(\boldsymbol{\alpha}) \le M_n(\boldsymbol{\alpha}), \tag{111}$$

where

$$m_p(\boldsymbol{\alpha}) \equiv \max_{i \in I_{\rm up}(\boldsymbol{\alpha}), y_i = 1} -y_i \nabla f(\boldsymbol{\alpha})_i, \qquad M_p(\boldsymbol{\alpha}) \equiv \min_{i \in I_{\rm low}(\boldsymbol{\alpha}), y_i = 1} -y_i \nabla f(\boldsymbol{\alpha})_i, \text{ and}$$

$$m_n(\boldsymbol{\alpha}) \equiv \max_{i \in I_{\rm up}(\boldsymbol{\alpha}), y_i = -1} -y_i \nabla f(\boldsymbol{\alpha})_i, \qquad M_n(\boldsymbol{\alpha}) \equiv \min_{i \in I_{\rm low}(\boldsymbol{\alpha}), y_i = -1} -y_i \nabla f(\boldsymbol{\alpha})_i.$$

One can also define a violating pair as the following:

Definition 2 (Violating pair of (111)**)** *If $i \in I_{\rm up}(\boldsymbol{\alpha}), j \in I_{\rm low}(\boldsymbol{\alpha})$, $y_i = y_j$, and $-y_i \nabla f(\boldsymbol{\alpha})_i > -y_j \nabla f(\boldsymbol{\alpha})_j$, then $\{i, j\}$ is a "violating pair."*

Clearly, the condition $y_i = y_j$ is the main difference from Definition 1. In fact the selected pair $B = \{i, j\}$ must satisfy $y_i = y_j$. If $y_i \neq y_j$, then the two linear equalities result in the sub-problem having only one feasible point $\boldsymbol{\alpha}_B^k$. For the same y_i and y_j, the two equations in the sub-problem are identical, so we could move $\boldsymbol{\alpha}_B^k$ to a better point. In addition, the sub-problem is then in the same form of (2), so the procedure in Section III-A directly works.

All working set selections discussed in Section II can be extended here. We modify WSS 3 as an example:

WSS 4 (Extension of WSS 3 for ν-SVM)

1) Select any $i \in I_{\rm up}(\boldsymbol{\alpha}^k)$, $j \in I_{\rm low}(\boldsymbol{\alpha}^k)$, $y_i = y_j$ satisfying

$$-y_i \nabla f(\boldsymbol{\alpha}^k)_i + y_j \nabla f(\boldsymbol{\alpha}^k)_j \geq h(D(\boldsymbol{\alpha}^k)) > 0, \tag{112}$$

where

$$D(\boldsymbol{\alpha}^k) \equiv \max(m_p(\boldsymbol{\alpha}^k) - M_p(\boldsymbol{\alpha}^k), m_n(\boldsymbol{\alpha}^k) - M_n(\boldsymbol{\alpha}^k)). \tag{113}$$

2) Return $B = \{i, j\}$.

Results in Sections III-IV hold with minor modifications. They are listed in the following without detailed proofs. For easier description, we let

$\boldsymbol{\alpha}^p$ ($\boldsymbol{\alpha}^n$) be the sub-vector of $\boldsymbol{\alpha}$ corresponding to positive (negative) samples.

Theorem 9 *Let $\{\boldsymbol{\alpha}^k\}$ be the infinite sequence generated by the SMO-type decomposition method using WSS 4 for the working set selection. Then any limit point of $\{\boldsymbol{\alpha}^k\}$ is a stationary point of* (110)*.*

Theorem 10 *Assume Q is positive semi-definite.*

1) *If $\bar{\boldsymbol{\alpha}} \neq \hat{\boldsymbol{\alpha}}$ are any two optimal solutions of* (110)*, then*

$$-y_i \nabla f(\bar{\boldsymbol{\alpha}})_i = -y_i \nabla f(\hat{\boldsymbol{\alpha}})_i, i = 1, \ldots, l. \tag{114}$$

If $\bar{\boldsymbol{\alpha}}^p \neq \hat{\boldsymbol{\alpha}}^p$ ($\bar{\boldsymbol{\alpha}}^n \neq \hat{\boldsymbol{\alpha}}^n$), then

$$M_p(\bar{\boldsymbol{\alpha}}) = m_p(\bar{\boldsymbol{\alpha}}) = M_p(\hat{\boldsymbol{\alpha}}) = m_p(\hat{\boldsymbol{\alpha}}) \tag{115}$$

$$(M_n(\bar{\boldsymbol{\alpha}}) = m_n(\bar{\boldsymbol{\alpha}}) = M_n(\hat{\boldsymbol{\alpha}}) = m_n(\hat{\boldsymbol{\alpha}})). \tag{116}$$

2) *If there is an optimal solution $\bar{\boldsymbol{\alpha}}$ satisfying $m_p(\bar{\boldsymbol{\alpha}}) < M_p(\bar{\boldsymbol{\alpha}})$, then $\bar{\boldsymbol{\alpha}}^p$ is unique for* (110)*. Similarly, if $m_n(\bar{\boldsymbol{\alpha}}) < M_n(\bar{\boldsymbol{\alpha}})$, then $\bar{\boldsymbol{\alpha}}^n$ is unique.*

3) *The following sets are independent of any optimal solution $\bar{\boldsymbol{\alpha}}$:*

$$I_p \equiv \{i \mid y_i = 1, -\nabla f(\bar{\boldsymbol{\alpha}})_i > M_p(\bar{\boldsymbol{\alpha}}) \text{ or } -\nabla f(\bar{\boldsymbol{\alpha}})_i < m_p(\bar{\boldsymbol{\alpha}})\}, \tag{117}$$

$$I_n \equiv \{i \mid y_i = -1, \nabla f(\bar{\boldsymbol{\alpha}})_i > M_n(\bar{\boldsymbol{\alpha}}) \text{ or } \nabla f(\bar{\boldsymbol{\alpha}})_i < m_n(\bar{\boldsymbol{\alpha}})\}. \tag{118}$$

Moreover, problem (110) *has unique and bounded optimal solutions at α_i, $i \in I_p \cup I_n$.*

Theorem 11 *Assume Q is positive semi-definite. Let $\{\boldsymbol{\alpha}^k\}$ be the infinite sequence generated by the SMO-type decomposition method using WSS 4. If $\{(\boldsymbol{\alpha}^k)^p\}$ ($\{(\boldsymbol{\alpha}^k)^n\}$) is updated in infinitely many iterations, then*

$$\lim_{k\to\infty} m_p(\boldsymbol{\alpha}^k) - M_p(\boldsymbol{\alpha}^k) = 0 \tag{119}$$

$$(\lim_{k\to\infty} m_n(\boldsymbol{\alpha}^k) - M_n(\boldsymbol{\alpha}^k) = 0). \tag{120}$$

Theorem 12 *Assume Q is positive semi-definite and the SMO-type method Algorithm 2 uses WSS 4. Let I_p and I_n be the sets defined in* (117) *and* (118)*. Define $\mathcal{K}^p$ and $\mathcal{K}^n$ to be iterations in which the working set is from positive and negative samples, respectively.*

1) *There is $\bar{k}$ such that after $k \geq \bar{k}$, every $\alpha_i^k,\ i \in I_p \cup I_n$ has reached the unique and bounded optimal solution. For any $i \in I_p$, there is $\bar{k}$ such that after $k \geq \bar{k}$, $k \in \mathcal{K}^p$, i is not in the following set*

$$\{t \mid y_t = 1, M_p(\boldsymbol{\alpha}^k) \leq -y_t \nabla f(\boldsymbol{\alpha}^k)_t \leq m_p(\boldsymbol{\alpha}^k)\}. \tag{121}$$

The same result holds for the negative part.

2) *If* (110) *has an optimal solution $\bar{\boldsymbol{\alpha}}$ satisfying $m_p(\bar{\boldsymbol{\alpha}}) < M_p(\bar{\boldsymbol{\alpha}})$ ($m_n(\bar{\boldsymbol{\alpha}}) < M_n(\bar{\boldsymbol{\alpha}})$), then $\bar{\boldsymbol{\alpha}}^p$ ($\bar{\boldsymbol{\alpha}}^n$) is unique for* (110) *and the decomposition method reaches it in a finite number of iterations.*

3) *If $\{(\boldsymbol{\alpha}^k)^p\}$ is updated in infinitely many iterations, then the following two limits exist and are equal:*

$$\lim_{k\to\infty} m_p(\boldsymbol{\alpha}^k) = \lim_{k\to\infty} M_p(\boldsymbol{\alpha}^k) = m_p(\bar{\boldsymbol{\alpha}}) = M_p(\bar{\boldsymbol{\alpha}}) \tag{122}$$

where $\bar{\boldsymbol{\alpha}}$ is any optimal solution. The same result holds for $\{(\boldsymbol{\alpha}^k)^n\}$.

VII. Discussion and Conclusions

This paper provides a comprehensive study on SMO-type decomposition methods. Below we discuss some issues for future investigation.

A. Faster Training via a Better Selection Rule

Under the general framework discussed in this paper, we can design various selection rules for practical implementations. Among them, the one using the maximal violating pair has been widely applied in SVM software. Developing better rules from the proposed framework is important. Otherwise, this article has only theoretical values. A working set selection which leads to faster training should:

1) reduce the number of iterations of the decomposition method. In other words, the convergence is faster, and
2) keep the cost of identifying the working set B similar to that of finding the maximal violating pair.

The challenge is that these two goals are often at odds. For example, fewer iterations may not reduce the training time if the cost of working set selections is higher. We proposed some rules with the above properties in [8]. This work shows that the "maximal violating pair" uses only the first-order information of the objective function, and derives a better selection rule using the second-order information. The new rule is a special case of WSS 2. The training time is generally shorter than that by using the "maximal violating pair."

B. Necessary Conditions for the Convergence

Previous studies and the discussion in Section III provide "sufficient conditions" for the convergence of decomposition methods. That is, under given working set selections, we prove the convergence. Investigating the "necessary conditions" is also interesting. When the decomposition method converges, which conditions does its working set selection satisfy?

We may think that WSS 3 is general enough so that every convergent SMO-type method satisfies the condition (9). However, this may not be right. We suspect that even if some iterations select $\{i, j\}$ without enough violation (i.e. $-y_i \nabla f(\boldsymbol{\alpha}^k)_i + y_i \nabla f(\boldsymbol{\alpha}^k)_j < h(m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k))$), the method may still converge if other iterations have used appropriate working sets. Therefore, finding useful necessary conditions may be a challenging task.

It is worth mentioning another working set selection proposed in [21]. This work requires the following condition:

$$\begin{aligned} &\text{There is } N > 0 \text{ such that for all } k, \\ &\text{any violating pair of } \boldsymbol{\alpha}^k \text{ is selected at least once in iterations } k \text{ to } k+N. \end{aligned} \tag{123}$$

Clearly, a cyclic selection of $\{i, j\}$ in every $l(l-1)/2$ iterations satisfies (123):

$$\{1,2\},\{1,3\},\ldots,\{1,l\},\{2,3\},\ldots,\{l-1,l\},$$

where l is the number of data instances. With (123), the convergence proof in Theorem 3 becomes very simple. For the limit point $\bar{\boldsymbol{\alpha}}$ assumed not stationary, its maximal violating pair $\{\bar{i}, \bar{j}\}$ is also a violating pair at iteration k, where $k \in \mathcal{K}, k \geq \bar{k}$. According to (123), this pair $\{\bar{i}, \bar{j}\}$ of $\bar{\boldsymbol{\alpha}}$ must be selected at iteration k', where $k \leq k' \leq k + N$. Then at the $(k'+1)st$ iteration, (28) and (29) imply

$$\bar{i} \in I_{\text{up}}(\boldsymbol{\alpha}^{k'+1}) \text{ and } \bar{j} \in I_{\text{low}}(\boldsymbol{\alpha}^{k'+1}),$$

but (33) indicates that

$$\bar{i} \notin I_{\text{up}}(\boldsymbol{\alpha}^{k'+1}) \text{ or } \bar{j} \notin I_{\text{low}}(\boldsymbol{\alpha}^{k'+1}).$$

Thus, immediately there is a contradiction. In a sense, (123) is a rule "designed" for the convergence proof.

The two conditions (9) and (123) on the working set selection are quite different, so neither is a necessary condition. From the counter-example in Section II we observed that the selected $\{i, j\}$ has a much smaller violation than $m(\boldsymbol{\alpha}^k) - M(\boldsymbol{\alpha}^k)$, and hence proposed WSS 2 and 3. This example also has a violating pair $\{4, 5\}$ never selected, a situation opposite to (123). Thus both WSS 3 and the condition (123) attempt to remedy problems imposed from this counter-example, but they take very different directions. One focuses on issues related to the maximal violating pair, but the other requires that all current violating pairs are selected later in a finite number of iterations. In general, we think the former leads to faster convergence as it more aggressively reduces the violation. However, this also complicates the convergence proof as a counting procedure in Theorem 3 must be involved in order to obtain the contradiction.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Council of Taiwan via the grant NSC 93-2213-E-002-030.

REFERENCES

[1] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

[2] C.-C. Chang, C.-W. Hsu, and C.-J. Lin. The analysis of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 11(4):1003–1008, 2000.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] C.-C. Chang and C.-J. Lin. Training ν-support vector classifiers: Theory and algorithms. *Neural Computation*, 13(9):2119–2147, 2001.

[5] C. Cortes, P. Haffner, and M. Mohri. Positive definite rational kernels. In *Proceedings of the 16th Annual Conference on Learning Theory*, pages 41–56, 2003.

[6] C. Cortes and V. Vapnik. Support-vector network. *Machine Learning*, 20:273–297, 1995.

[7] D. J. Crisp and C. J. C. Burges. A geometric interpretation of ν-SVM classifiers. In S. Solla, T. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, Cambridge, MA, 2000. MIT Press.

[8] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using second order information for training SVM. *Journal of Machine Learning Research*, 6:1889–1918, 2005.

[9] D. Hush and C. Scovel. Polynomial-time decomposition algorithms for support vector machines. *Machine Learning*, 51:51–71, 2003.

[10] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1998. MIT Press.

[11] S. S. Keerthi and E. G. Gilbert. Convergence of a generalized SMO algorithm for SVM classifier design. *Machine Learning*, 46:351–360, 2002.

[12] S. S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, 15(7):1667–1689, 2003.

[13] S. S. Keerthi and C. J. Ong. On the role of the threshold parameter in SVM training algorithms. Technical Report CD-00-09, Department of Mechanical and Production Engineering, National University of Singapore, Singapore, 2000.

[14] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13:637–649, 2001.

[15] C.-J. Lin. Linear convergence of a decomposition method for support vector machines. Technical report, Department of Computer Science, National Taiwan University, 2001.

[16] C.-J. Lin. On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, 12(6):1288–1298, 2001.

[17] C.-J. Lin. Asymptotic convergence of an SMO algorithm without any assumptions. *IEEE Transactions on Neural Networks*, 13(1):248–250, 2002.

[18] C.-J. Lin. A formal analysis of stopping criteria of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 13(5):1045–1052, 2002.

[19] H.-T. Lin and C.-J. Lin. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University, 2003.

[20] N. List and H. U. Simon. A general convergence theorem for the decomposition method. In *Proceedings of the 17th Annual Conference on Learning Theory*, pages 363–377, 2004.

[21] S. Lucidi, L. Palagi, M. Sciandrone, and A. Risi. A convergent decomposition algorithm for support vector machines. *Computational Optimization and Applications*, 2006. To appear.

[22] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[23] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Proceedings of CVPR'97*, pages 130–136, New York, NY, 1997. IEEE.

[24] L. Palagi and M. Sciandrone. On the convergence of a modified version of SVM^{light} algorithm. *Optimization Methods and Software*, 20(2-3):315–332, 2005.

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

[26] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[27] B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[28] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, 1995.

APPENDIX

A. Proof of Two Lemmas Used in Section V

Lemma 3 *If $v_1 \geq \cdots \geq v_l$, then*

$$\max_i(|v_i|) \geq \frac{v_1 - v_l}{2}.$$

Proof: We notice that $\max_i(|v_i|)$ must happen at v_1 or v_l. It is easy to see

$$\frac{v_1 - v_l}{2} \leq \frac{|v_1| + |v_l|}{2} \leq \max(|v_1|, |v_l|).$$

■

Lemma 4 *If Q is invertible, then for any $\mathbf{x}$ such that*

1) $\mathbf{e}^T\mathbf{x} = 0$,
2) $\mathbf{v} \equiv Q\mathbf{x}$, $\max_i((Q\mathbf{x})_i) = v^1 > v^l = \min_i((Q\mathbf{x})_i)$, *and* $v^1 v^l \geq 0$,

we have

$$\min(|v^1|, |v^l|) \leq (1 - \frac{\mathbf{e}^T Q^{-1}\mathbf{e}}{\sum_{i,j} |Q^{-1}_{ij}|}) \max(|v^1|, |v^l|).$$

Proof: Since $v^1 > v^l$ and $v^1 v^l \geq 0$, we have either $v^1 > v^l \geq 0$ or $0 \geq v^1 > v^l$. For the first case, if the result is wrong,

$$v^l > (1 - \frac{\mathbf{e}^T Q^{-1}\mathbf{e}}{\sum_{i,j} |Q^{-1}_{ij}|})v^1,$$

so for $j = 1, \ldots, l$,

$$\begin{aligned} v^1 - v_j &\leq v^1 - v^l \\ &< (\frac{\mathbf{e}^T Q^{-1}\mathbf{e}}{\sum_{i,j} |Q^{-1}_{ij}|})v^1. \end{aligned} \tag{124}$$

With $\mathbf{x} = Q^{-1}\mathbf{v}$ and (124),

$$\begin{aligned} \mathbf{e}^T\mathbf{x} &= \mathbf{e}^T Q^{-1}\mathbf{v} \\ &= \sum_{i,j} Q^{-1}_{ij} v_j \\ &= \sum_{i,j} Q^{-1}_{ij}(v^1 - (v^1 - v_j)) \\ &\geq v^1 \mathbf{e}^T Q^{-1}\mathbf{e} - (v^1 - v^l)\sum_{i,j} |Q^{-1}_{ij}| \\ &> v^1 \left(\mathbf{e}^T Q^{-1}\mathbf{e} - (\frac{\mathbf{e}^T Q^{-1}\mathbf{e}}{\sum_{i,j} |Q^{-1}_{ij}|}) \sum_{i,j} |Q^{-1}_{ij}| \right) \\ &= 0 \end{aligned}$$

causes a contradiction. The case of $0 \geq v^1 > v^l$ is similar. ■