Convergence of a Generalized Gradient Selection Approach for the Decomposition Method *

Nikolas List

Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany
Niko.List@gmx.de

Abstract. The decomposition method is currently one of the major methods for solving the convex quadratic optimization problems being associated with support vector machines. For a special case of such problems the convergence of the decomposition method to an optimal solution has been proven based on a working set selection via the gradient of the objective function. In this paper we will show that a generalized version of the gradient selection approach and its associated decomposition algorithm can be used to solve a much broader class of convex quadratic optimization problems.

1 Introduction

In the framework of Support–Vector–Machines (SVM) introduced by Vapnik et al. [1] special cases of convex quadratic optimization problems have to be solved. A popular variant of SVM is the C–Support–Vector–Classification (C–SVC) where we try to classify m given data points with binary labels $y \in \{\pm 1\}^m$. This setting induces the following convex optimization problem:

$$\min_x f_C(x) := \frac{1}{2} x^\top Q x - e^\top x \quad \text{s.t. } 0 \leq x_i \leq C, \forall i = 1, \ldots, m,\ y^\top x = 0 \ , \quad (1)$$

where $x \in \mathbb{R}^m$, C is a real constant and e is the m–dimensional vector of ones. $Q \in \mathbb{R}^{m \times m}$ is a positive semi–definite matrix whose entries depend on the data points and the used kernel[1].

In general, the matrix Q is dense and therefore a huge amount of memory is necessary to store it if traditional optimization algorithms are directly applied to it. A solution to this problem, as proposed by Osuna et al. [4], is to decompose the large problem in smaller ones which are solved iteratively. The key idea is to select a working set B^k in each iteration based on the currently ,,best" feasible x^k. Then the subproblem based on the variables $x_i, i \in B^k$ is solved and the new solution x^{k+1} is updated on the selected indices while it remains unchanged on

* This work was supported in part by the IST Programm of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors views.

[1] The reader interested in more background information concerning SVM is referred to [2, 3].

the complement of B^k. This strategy has been used and refined by many other authors [5–7].

A key problem in this approach is the selection of the working set B^k. One widely used technique is based on the gradient of the objective function at x^k [5][2]. The convergence of such an approach to an optimal solution has been proven by Lin [8][3]. A shortcoming of the proposed selection and the convergence proof of the associated decomposition method in [8] is that they have only been formulated for the special case of (1) or related problems with only one equality constraint. Unfortunately ν-SVM introduced by Schölkopf et al. [10] leads to optimization problems with two such equality constraints (see e. g. [11]). Although there is an extension of the gradient algorithm to the case of ν-SVM [12] no convergence proof for this case has been published.

1.1 Aim of this paper

There exist multiple formulations of SVM which are used in different classification and regression problems as well as quantile estimation and novelty detection[4] which all lead to slightly different optimization problems. Nonetheless all of them can be viewed as special cases of a general convex quadratic optimization problem (2, see below). As the discussion concerning the decomposition method has mainly focused on the single case of C–SVC, it may be worth studying under which circumstances the decomposition method is applicable to solve such general problems and when such a strategy converges to an optimal solution. Recently Simon and List have investigated this topic. They prove a very general convergence theorem for the decomposition method, but for the sake of generality no practical selection algorithm is given [13].

The aim of this paper is to show that the well known method of gradient selection as implemented e. g. in SVM^{light} [5] can be extended to solve a far more general class of quadratic optimization problems. In addition, the achieved theoretical foundation, concerning the convergence and efficiency of this selection method, is preserved by adapting Lin's convergence theorem [8] to the decomposition algorithm associated with the proposed general gradient selection.

We want to point out that not necessarily all cases covered by the given generalization arise in SVM but the class of discussed optimization problems subsumes all the above mentioned versions of SVM. The proposed selection algorithm is therefore useful for the decomposition of all such problems and gives a unified approach to the convergence proof of the decomposition method associated with this selection.

[2] Platts' SMO [6] with the extension of Keerthi et al. [7] can be viewed as a special case of that selection.

[3] For the special case of SMO Keerthi and Gilbert have proven the convergence [9].

[4] [2, 3] both give an overview.

2 Definitions and Notations

2.1 Notations

The following naming conventions will be helpful: If $A \in \mathbb{R}^{n\times m}$ is a matrix, $A_i, i \in \{1,\ldots,m\}$ will denote the i^{th} column. Vectors $x \in \mathbb{R}^m$ will be considered column vectors so that their transpose $x^\top$ is a row vector. We will often deal with a partitioning of $\{1,\ldots,m\}$ in two disjunct sets B and N and in this case the notion A_B will mean the matrix consisting only of columns A_i with $i \in B$. Ignoring permutation of columns we therefore write $A = [\, A_B \; A_N \,]$. The same shall hold for vectors $x \in \mathbb{R}^m$ where x_B denotes the vector consisting only of entries x_i with $i \in B$. We can therefore expand $Ax = d$ to $[\, A_B \; A_N \,] \left(\begin{smallmatrix} x_B \\ x_N \end{smallmatrix}\right) = d$. A matrix $Q \in \mathbb{R}^{m\times m}$ can then be decomposed to four block matrices Q_{BB}, Q_{BN}, Q_{NB} and Q_{NN} accordingly.

Inequalities $l \leq r$ of two vectors $l, r \in \mathbb{R}^m$ will be short for $l_i \leq r_i, \forall i \in \{1,\ldots,m\}$. In addition we will adopt the convention that the maximum over an empty set will be $-\infty$ and the minimum ∞ accordingly.

Throughout the paper, we will be concerned with the following convex quadratic optimization problem $\mathcal{P}$:

$$\min_x f(x) := \frac{1}{2} x^\top Q x + c^\top x \text{ s.t. } l \leq x \leq u,\; Ax = d \tag{2}$$

where Q is a positive semi-definite matrix and $A \in \mathbb{R}^{n\times m}$, $l, u, x, c \in \mathbb{R}^m$ and $d \in \mathbb{R}^n$. The feasibility region of $\mathcal{P}$ will be denoted by $\mathcal{R}(\mathcal{P})$.

2.2 Selections, Subproblems and Decomposition

Let us now define the notions used throughout this paper.

Definition 1 (Selection). *Let $q < m$. A map $B : \mathbb{R}^m \longrightarrow \mathfrak{P}(\{1,\ldots,m\})$ such that $|B(x)| \leq q$ for any x and $\hat{x}$ is optimal wrt. $\mathcal{P}$ iff $B(\hat{x}) = \emptyset$ is called (q–)**significant selection** wrt. $\mathcal{P}$.*[5]

Definition 2 (Subproblem). *For a given set $B \subset \{1,\ldots,m\}$ such that $|B| \leq q$, $N := \{1,\ldots,m\} \backslash B$ and a given $x \in \mathcal{R}(\mathcal{P})$ we define $f_{B,x_N}(x') := \frac{1}{2} x'^\top Q_{BB} x' + (c_B + Q_{BN} x_N)^\top x'$ for every $x' \in \mathbb{R}^q$. The following optimization problem $\mathcal{P}_{B,x_N}$*

$$\min_{x'} f_{B,x_N}(x') \;\; s.t.\; l_B \leq x' \leq u_B,\; A_B x' = d - A_N x_N \tag{3}$$

*will be called the **subproblem** induced by B.*[6]

We are now in the position to define the decomposition method formally:

Algorithm 1 (Decomposition Method). *The following algorithm can be associated with every significant selection wrt. $\mathcal{P}$*

[5] If x is evident from the context we often write B instead of $B(x)$.

[6] Note, that $\mathcal{P}_{B,x_N}$ is a special case of $\mathcal{P}$.

1: ***Initialize:*** $k \leftarrow 0$ *and* $x^0 \in \mathcal{R}(\mathcal{P})$
2: $B \leftarrow B(x^k)$
3: ***while*** $B \neq \emptyset$ ***do***
4: $\quad N \leftarrow \{1, \ldots, m\} \setminus B$
5: $\quad$ *Find* x' *as an optimal solution of* $\mathcal{P}_{B, x^k_N}$
6: $\quad$ *Set* $x^{k+1} \leftarrow \begin{pmatrix} x' \\ x^k_N \end{pmatrix}$
7: $\quad k \leftarrow k+1$, $B \leftarrow B(x^k)$
8: ***end while***

This algorithm shall be called the ***decomposition method*** *of* $\mathcal{P}$ *induced by* B*.*

3 Gradient Selection

The idea of selecting the indices via an ordering according to the gradients of the objective function has been motivated by the idea to select indices which contribute most to the steepest descent in the gradient field (see [5]). We would like to adopt another point of view in which indices violating the KKT–conditions of the problem $\mathcal{P}$ are selected. This idea has been used e. g. in [7] and [11].

To motivate this we will first focus on the well–known problem of C–SVC (Sec. 3.1) and later enhance this strategy to a more general setting (Sec. 3.2).

3.1 *C*-SVC and KKT–violating pairs

In the case of C–SVC a simple reformulation of the KKT–conditions leads to the following optimality criterion: $\hat{x}$ is optimal wrt. (1) iff there exists a $b \in \mathbb{R}$ such that for any $i \in \{1, \ldots, m\}$

$$\left(\hat{x}_i > 0 \Rightarrow \nabla f(\hat{x})_i - b y_i \leq 0\right) \quad \text{and} \quad \left(\hat{x}_i < C \Rightarrow \nabla f(\hat{x})_i - b y_i \geq 0\right) . \tag{4}$$

Following [8] (4) can be rewritten as

$$\left(i \in \overline{I_{bot}}(\hat{x}) \Rightarrow y_i \nabla f(\hat{x})_i \geq b\right) \quad \text{and} \quad \left(i \in \overline{I_{top}}(\hat{x}) \Rightarrow y_i \nabla f(\hat{x})_i \leq b\right) ,$$

where

$$\begin{aligned} \overline{I_{top}}(x) &:= \{i \mid (x_i < C, y_i = -1) \wedge (x_i > 0, y_i = 1)\} \\ \overline{I_{bot}}(x) &:= \{i \mid (x_i > 0, y_i = -1) \wedge (x_i < C, y_i = 1)\} . \end{aligned}$$

The KKT–conditions can therefore be collapsed to a simple inequality[7]: $\hat{x}$ is optimal iff

$$\max_{i \in \overline{I_{top}}(x)} y_i \nabla f(x)_i \leq \min_{i \in \overline{I_{bot}}(x)} y_i \nabla f(x)_i \tag{5}$$

[7] Note, that if one of the sets is empty, it's easy to fulfill the KKT–conditions by choosing an arbitrarily large or small b.

Given a non–optimal feasible x, we can now identify indices $(i, j) \in \overline{I_{top}}(x) \times \overline{I_{bot}}(x)$ that satisfy the following inequality:

$$y_i \nabla f(x)_i > y_j \nabla f(x)_j$$

Such pairs do not admit the selection of a $b \in \mathbb{R}$ according to (4). Following [9], such indices are therefore called KKT–violating pairs .

From this point of view the selection algorithm proposed by Joachims chooses pairs of indices which violate this inequality the most. As this strategy, implemented for example in SVM^{light} [5], selects the candidates in $\overline{I_{top}}(x)$ from the top of the list sorted according to the value of $y_i \nabla f(x)_i$, Lin calls indices in $\overline{I_{top}}(x)$,,top–candidates" (see [8]). Elements from $\overline{I_{bot}}(x)$ are selected from the bottom of this list and are called ,,bottom–candidates". We will adopt this naming convention here.

3.2 General KKT–Pairing

We will now show that a pairing strategy, based on a generalized version of the KKT–conditions in (5), can be extended to a more general class of convex quadratic optimization problems. The exact condition is given in the following definition:

Definition 3. *Let $\mathcal{P}$ be a general convex quadratic optimization problem as given in* (2)*. If any selection of pairwise linearly independent columns $A_i \in \mathbb{R}^n$ of the equality constraint matrix A is linearly independent, we call the problem $\mathcal{P}$* ***decomposable by pairing****.*

Note, that most variants of SVM, including ν–SVM, fulfill this restriction.

We may then define an equivalence relation on the set of indices $i \in \{1, \ldots, m\}$ as follows:

$$i \sim j \Leftrightarrow \exists \lambda_{i,j} \in \mathbb{R} \setminus \{0\} : \lambda_{i,j} A_i = A_j \tag{6}$$

Let $\{i_r \mid r = 1, \ldots, s\}$ be a set of representatives of this relation. The subset of corresponding columns

$$\{a_r := A_{i_r} \mid r = 1, \ldots, s\} \subset \{A_i \mid i = 1, \ldots, m\}$$

therefore represents the columns of A up to scalar multiplication. Additionally, we define $\lambda_i := \lambda_{i,i_r}$ for any $i \in [i_r]$. Thus $\lambda_i A_i = a_r$ if $i \in [i_r]$.

From Definition 3 we draw two simple conclusions for such a set of representatives: As all $\{a_r \mid r = 1, \ldots, s\}$ are pairwise linearly independent by construction, they are linearly independent and it follows that $s = \dim\langle a_r \mid r = 1, \ldots, s\rangle \leq \operatorname{rank} A \leq n$.

To formulate the central theorem of this section we define our generalized notion of ,,top" and ,,bottom" candidates first:

Definition 4. *Let $\{i_1, \ldots, i_s\}$ be a set of representatives for the equivalence relation (6). For any $r \in \{1, \ldots, s\}$ we define:*

$$\begin{aligned}\overline{I_{top,r}}(x) &:= [i_r] \cap \{j \mid (x_j > l_j \wedge \lambda_j > 0) \vee (x_j < u_j \wedge \lambda_j < 0)\} \\ \overline{I_{bot,r}}(x) &:= [i_r] \cap \{j \mid (x_j > l_j \wedge \lambda_j < 0) \vee (x_j < u_j \wedge \lambda_j > 0)\}\end{aligned}$$

*Indices in $\overline{I_{top,r}}(x)$ ($\overline{I_{bot,r}}(x)$) are called **top candidates** (**bottom candidates**). Top candidates for which $x_i = l_i$ or $x_i = u_i$ are called **top–only** candidates. The notion **bottom–only** candidate is defined accordingly. We use the following notation:*

$$\begin{aligned}I_{top,r}(x) &:= \{i \in \overline{I_{top,r}}(x) \mid x_i \in \{l_i, u_i\}\}\ , \\ I_{bot,r}(x) &:= \{i \in \overline{I_{bot,r}}(x) \mid x_i \in \{l_i, u_i\}\}\end{aligned}$$

The following naming convention will be helpful:

$$\overline{I_{top}}(x) := \bigcup_{r \in \{i_1, \ldots, i_s\}} \overline{I_{top,r}}(x)\ .$$

$\overline{I_{bot}}(x)$, $I_{top}(x)$ and $I_{bot}(x)$ are defined accordingly.

We will now generalize (5) to problems $\mathcal{P}$ decomposable by pairing. The KKT–conditions of such a problem $\mathcal{P}$ say that $\hat{x}$ is optimal wrt. $\mathcal{P}$ iff there exists an $h \in \mathbb{R}^n$ such that for any $i \in \{1, \ldots, m\}$

$$\big(\hat{x}_i > l_i \Rightarrow \nabla f(\hat{x})_i - A_i^\top h \leq 0\big) \quad \text{and} \quad \big(\hat{x}_i < u_i \Rightarrow \nabla f(\hat{x})_i - A_i^\top h \geq 0\big)\ .$$

Given a set of representatives $\{i_r \mid r = 1, \ldots, s\}$ and $A_i = \frac{1}{\lambda_i} a_r$ for $i \in [i_r]$ this condition can be written as follows: $\hat{x}$ is optimal wrt. $\mathcal{P}$ iff there exists an $h \in \mathbb{R}^n$ such that for any $i \in \{1, \ldots, m\}$

$$\big(i \in \overline{I_{top,r}}(\hat{x}) \Rightarrow \lambda_i \nabla f(\hat{x})_i \leq a_r^\top h\big) \text{ and } \big(i \in \overline{I_{bot,r}}(\hat{x}) \Rightarrow \lambda_i \nabla f(\hat{x})_i \geq a_r^\top h\big) \quad (7)$$

The following theorem will show that the KKT–conditions of such problems can be collapsed to a simple inequality analogous to (5):

Theorem 1. *If problem $\mathcal{P}$ is decomposable by pairing and a set of representatives $\{i_1, \ldots, i_r\}$ for the equivalence relation (6) is given, the KKT–conditions can be stated as follows:*
$\hat{x}$ is optimal wrt. $\mathcal{P}$ iff for all $r \in \{1, \ldots, s\}$

$$\max_{i \in \overline{I_{top,r}}(\hat{x})} \lambda_i \nabla f(\hat{x})_i \leq \min_{i \in \overline{I_{bot,r}}(\hat{x})} \lambda_i \nabla f(\hat{x})_i \quad (8)$$

Proof. For an optimal $\hat{x}$ equation (8) follows immediately from (7).

To prove the opposite direction we define $h^r_{top} := \max_{i \in \overline{I_{top,r}}(x)} \lambda_i \nabla f(x)_i$ and $h^r_{bot} := \min_{i \in \overline{I_{bot,r}}(x)} \lambda_i \nabla f(x)_i$ for any $r \in \{1, \ldots, s\}$. By assumption $h^r_{top} \leq h^r_{bot}$ and we can choose $\bar{h}^r \in [h^r_{top}, h^r_{bot}]$. As $\mathcal{P}$ is decomposable by pairing, it follows

that $\left(a_1, \ldots, a_s\right)^\top \in \mathbb{R}^{s \times n}$ *represents a surjective linear mapping from* $\mathbb{R}^n$ *to* $\mathbb{R}^s$*. Thus there exists an* $h \in \mathbb{R}^n$ *such that for all* $r \in \{1, \ldots, s\} : a_r^\top h = \bar{h}^r$*. We conclude that*

$$\lambda_i \nabla f(x)_i \leq h^r_{top} \leq \bar{h}^r = a_r^\top h \quad \text{if } i \in \overline{I_{top,r}}(x)$$
$$\lambda_i \nabla f(x)_i \geq h^r_{bot} \geq \bar{h}^r = a_r^\top h \quad \text{if } i \in \overline{I_{bot,r}}(x)$$

holds for all $i \in \{1, \ldots, m\}$*. Therefore the KKT–conditions* (7) *are satisfied and* $\hat{x}$ *is an optimal solution.* □

3.3 Selection Algorithm

Given a set of representatives $\{i_1, \ldots, i_s\}$ and the corresponding λ_i for all $i \in \{1, \ldots, m\}$ we are now able to formalize a generalized gradient selection algorithm. For any $x \in \mathcal{R}(\mathcal{P})$ we call the set

$$\mathcal{C}(x) := \{(i,j) \in \overline{I_{top,r}}(x) \times \overline{I_{bot,r}}(x) \mid \lambda_i \nabla f(x)_i - \lambda_j \nabla f(x)_j > 0, r = 1, \ldots, s\}$$

selection candidates wrt. x.

Algorithm 2. *Given a feasible* x*, we can calculate a* $B(x) \subset \{1, \ldots, m\}$ *for an even* q *as follows:*

1: ***Initialize:*** $\mathcal{C} \leftarrow \mathcal{C}(x)$, $B \leftarrow \emptyset$, $l \leftarrow q$.
2: ***while*** $(l > 0)$ *and* $(\mathcal{C} \neq \emptyset)$ ***do***
3: *Choose* $(i,j) = \operatorname{argmax}_{(i,j) \in \mathcal{C}} \ \lambda_i \nabla f(x)_i - \lambda_j \nabla f(x)_j$
4: $B \leftarrow B \cup \{i, j\}$.
5: $\mathcal{C} \leftarrow \mathcal{C} \setminus \{i, j\}^2$, $l \leftarrow l - 2$.
6: ***end while***
7: ***return*** B

We note that the selection algorithms given in [5, 8, 11, 12, 7] can be viewed as special cases of this algorithm. This holds as well for the extensions to ν–SVM.

4 Convergence of the decomposition method

Theorem 1 implies that the mapping B returned by Algorithm 2 is a significant selection in the sense of Definition 1 and therefore induces a decomposition method. Such a decomposition method converges to an optimal solution of $\mathcal{P}$ as stated in the following theorem:

Theorem 2. *Let* $\mathcal{P}$ *be a convex quadratic optimization problem decomposable by pairing. Then any limit point* $\bar{x}$ *of a sequence* $(x_n)_{n \in \mathbb{N}}$ *of iterative solutions of a decomposition method induced by a general gradient selection according to Algorithm 2 is an optimal solution of* $\mathcal{P}$*.*

The proof is given in the following sections and differs only in some technical details from the one given in [8].

4.1 Technical Lemmata

Let us first note that $\mathcal{R}(\mathcal{P})$ is compact and therefore, for any sequence $(x_n)_{n\in\mathbb{N}}$ of feasible solutions such a limit point $\bar{x}$ exists. Let, in the following, $(x_k)_{k\in\mathcal{K}}$ be a converging subsequence such that $\bar{x} = \lim_{k\in\mathcal{K}, k\to\infty} x_k$.

If we assume that the matrix Q satisfies the equation $\min_I \lambda_{min}(Q_{II}) > 0$ where I ranges over all subsets of $\{1,\ldots,m\}$ such that $|I| \leq q$ we can prove the following Lemma[8]:

Lemma 1. *Let $(x^k)_{k\in\mathcal{K}}$ be a converging subsequence. There exists an $\sigma > 0$ such that*

$$f(x^k) - f(x^{k+1}) \;\geq\; \sigma\|x^{k+1} - x^k\|^2$$

Proof. Let B^k denote $B(x^k)$, $N^k = \{1,\ldots,m\} \setminus B^k$ and $d := x^k - x^{k+1}$. Since f is a quadratic function Taylor-expansion around x^{k+1} yields

$$f(x^k) = f(x^{k+1}) + \nabla f(x^{k+1})^\top d + \frac{1}{2} d^\top Q d \;. \tag{9}$$

As x^{k+1} is an optimal solution of the convex optimization problem $\mathcal{P}_{B^k, x^k_{N^k}}$ we can conclude that the line segment L between x^k and x^{k+1} lies in the feasibility region $\mathcal{R}(\mathcal{P}_{B^k, x^k_{N^k}})$ of the subproblem induced by B^k and therefore $f(x^{k+1}) = \min_{x\in L} f(x)$. Thus, the gradient at x^{k+1} in direction to x^k is ascending, i.e.

$$\nabla f(x^{k+1})^\top d \geq 0 \;. \tag{10}$$

If $\sigma := \frac{1}{2}\min_I \lambda_{min}(Q_{II}) > 0$, the Courant-Fischer Minimax Theorem [15] implies

$$d^\top Q d \geq 2\sigma\|d\|^2 \;. \tag{11}$$

From (9), (10) and (11), the lemma follows. □

Lemma 2. *For any $l \in \mathbb{N}$ the sequence $(x^{k+l})_{k\in\mathcal{K}}$ converges with limit point $\bar{x}$ and, as $\lambda_i \nabla f(x)$ is continuous in x, $(\lambda_i \nabla f(x^{k+l})_i)_{k\in\mathcal{K}}$ converges accordingly with limit point $\lambda_i \nabla f(\bar{x})_i$ for all $i \in \{1,\ldots,m\}$*

Proof. According to Lemma 1, $\big(f(x^k)\big)_{k\in\mathcal{K}}$ is a monotonically decreasing sequence on a compact set and therefore converges and we are thus able to bound $\|x^{k+1} - \bar{x}\|$ for $k \in \mathcal{K}$ as follows:

$$\begin{aligned}\|x^{k+1} - \bar{x}\| \;&\leq \|x^{k+1} - x^k\| + \|x^k - \bar{x}\| \\ &\leq \sqrt{\frac{1}{\sigma}(f(x^k) - f(x^{k+1}))} + \|x^k - \bar{x}\|\end{aligned}$$

As $f(x^k)$ is a cauchy sequence and $x^k \xrightarrow{k\in\mathcal{K}, k\to\infty} \bar{x}$, this term converges to zero for $k \in \mathcal{K}$ and $k \to \infty$. Therefore $(x^{k+1})_{k\in\mathcal{K}}$ converges with limit $\bar{x}$. By induction the claim follows for any $l \in \mathbb{N}$. □

[8] This holds if Q is positive definite. With respect to problems induced by SVM this tends to hold for small q or special kernels like for example RBF-kernels. For the special case of SMO ($q = 2$) Lemma 1 has been proven without this assumption [14].

Lemma 3. *For any $i, j \in \{1, \ldots, m\}$ such that $i \sim j$ and $\lambda_i \nabla f(\bar{x})_i > \lambda_j \nabla f(\bar{x})_j$ and all $l \in \mathbb{N}$ there exists a $k \in \mathcal{K}$ such that for the next l iterations $k' \in \{k, \ldots, k+l\}$ the following holds:*
If i, j are both selected in iteration k', either i becomes bottom–only or j becomes top–only for the next iteration $k'+1$.

Proof. According to Lemma 2 we can find for any given $l \in \mathbb{N}$ a $k \in \mathcal{K}$ such that for any $k' \in \{k, \ldots, k+l\}$ the following inequality holds:

$$\lambda_i \nabla f(x^{k'+1})_i > \lambda_j \nabla f(x^{k'+1})_j \ . \tag{12}$$

Assume that $i, j \in B(x^{k'})$ for any $k' \in \{k, \ldots, k+l\}$. As both indices are in the working set, $x^{k'+1}$ is an optimal solution of $\mathcal{P}_{B(x^{k'}), x^{k'}_{N(k')}}$. For sake of contradiction we assume that i is top candidate and j bottom candidate in iteration $k'+1$ at the same time, i.e. $i \in \overline{I_{top,r}}(x^{k'+1})$ and $j \in \overline{I_{bot,r}}(x^{k'+1})$ for an $r \in \{1, \ldots, s\}$[9]. In this case, as $\mathcal{P}_{B(x^{k'}), x^{k'}_{N(k')}}$ is decomposable by pairing, Theorem 1 implies

$$\lambda_i \nabla f(x^{k'+1})_i \leq \lambda_j \nabla f(x^{k'+1})_j \ .$$

This contradicts to (12) *and the choice of k'.* □

4.2 Convergence Proof

We are now in the position to state the main proof of the convergence theorem. For sake of contradiction, we assume there exists a limit point $\bar{x}$ which is not optimal wrt. $\mathcal{P}$.

In the following, we will concentrate on the set $\mathcal{C}_> := \{r \mid [i_r]^2 \cap C(\bar{x}) \neq \emptyset\} \subset \{1, \ldots, s\}$ of equivalence classes which contribute to the selection candidates $C(\bar{x})$ on the limit point $\bar{x}$. As $\bar{x}$ is not optimal $\mathcal{C}_> \neq \emptyset$. For any such equivalence class we select the most violating pair (ι_r, κ_r) as follows:

$$(\iota_r, \kappa_r) = \operatorname{argmax}_{(i,j) \in \mathcal{C}(\bar{x}) \cap [i_r]^2} \lambda_i \nabla f(\bar{x})_i - \lambda_j \nabla f(\bar{x})_j$$

Based on these pairs, we define the following two sets for any $r \in \mathcal{C}_>$:

$$\begin{aligned} \mathcal{I}_r &:= \{i \in [i_r] \mid \lambda_i \nabla f(\bar{x})_i \geq \lambda_{\iota_r} \nabla f(\bar{x})_{\iota_r}\} \\ \mathcal{K}_r &:= \{i \in [i_r] \mid \lambda_i \nabla f(\bar{x})_i \leq \lambda_{\kappa_r} \nabla f(\bar{x})_{\kappa_r}\} \end{aligned}$$

For any iteration $k \in \mathbb{N}$ an index $i \in \mathcal{I}_r \setminus \{\iota_r\}$ will then be called dominating ι_r in iteration k iff $i \in \overline{I_{top,r}}(x^k)$. An index $j \in \mathcal{K}_r \setminus \{\kappa_r\}$ will be called dominating κ_r iff $j \in \overline{I_{bot,r}}(x^k)$ accordingly. d^k will denote the number of all dominating indices in iteration k. Note that, as no $i \in [i_r]$ can dominate ι_r as well as κ_r in one iteration, d^k is bounded by $m - 2|\mathcal{C}_>|$. We now claim, that there exists a $k \in \mathcal{K}$ such that, for the next $m_{\mathcal{C}} := m - 2|\mathcal{C}_>| + 1$ iterations $k' \in \{k, \ldots, k + m_{\mathcal{C}}\}$,

[9] As $i \sim j$ such an r must exist

the following two conditions hold: 1) $d^{k'} > 0$ and 2) $d^{k'+1} < d^{k'}$ which leads to the aspired contradiction.

The $k \in \mathcal{K}$ we are looking for can be chosen, according to Lemma 2, such that for the next $m_{\mathcal{C}} + 1$ iterations all inequalities on $\bar{x}$ will be preserved, i.e. for $k' \in \{k, \dots, k + m_{\mathcal{C}} + 1\}$ the following holds:

$$\forall (i,j) \in \{1, \dots, m\}^2 : \lambda_i \nabla f(\bar{x})_i < \lambda_j \nabla f(\bar{x})_j \Rightarrow \lambda_i \nabla f(x^{k'})_i < \lambda_j \nabla f(x^k)_j \ ,$$
$$\bar{x}_i > l_i \Rightarrow x^{k'} > l_i \quad \text{and} \quad \bar{x}_i < u_i \Rightarrow x^{k'} < u_i \ .$$

Note that $(\iota_r, \kappa_r) \in \overline{I_{top,r}}(x^{k'}) \times \overline{I_{bot,r}}(x^{k'})$ for any such k' and thus, due to Lemma 3, they cannot be select at the same time. Let us now prove the two conditions introduced earlier for the selected k:
1) As $d^{k'} = 0$ would imply, that for any $r \in \mathcal{C}_>$

$$(\iota_r, \kappa_r) = \operatorname{argmax}_{(i,j) \in \mathcal{C}(x^{k'})} \lambda_i \nabla f(x^{k'})_i - \lambda_j \nabla f(x^{k'})_j$$

at least one pair (ι_r, κ_r) would be selected in the next iteration $k' + 1$ which is, by choice of k, not possible for any $k' \in \{k, \dots, k + m_{\mathcal{C}}\}$. Therefore $d^{k'} > 0$ holds for all such k' as claimed.
2) To prove that $d^{k'}$ will decrease in every iteration $k' \in \{k, \dots, k + m_{\mathcal{C}}\}$ by at least one, we have to consider two aspects: First, there have to be vanishing dominating indices, i.e. a top candidate from $\mathcal{I}_r$ has to become bottom–only or a bottom candidate from $\mathcal{K}_r$ has to become top–only. Second, we have to take care that the number of non-dominating indices, which become dominating in the next iteration, is strictly bounded by the number of vanishing dominating indices.
Note, that the state of an index i, concerning domination, only changes if i is selected, i.e. $i \in B(x^{k'})$. As we will only be concerned with such selected indices, let us define the following four sets:

$$\mathcal{I}_r^+ := \mathcal{I}_r \cap \overline{I_{top,r}}(x^{k'}) \cap B(x^{k'}) \ , \quad \mathcal{I}_r^- := \mathcal{I}_r \setminus \{\iota_r\} \cap I_{bot,r}(x^{k'}) \cap B(x^{k'}) \ ,$$
$$\mathcal{K}_r^+ := \mathcal{K}_r \cap \overline{I_{bot,r}}(x^{k'}) \cap B(x^{k'}) \ , \quad \mathcal{K}_r^- := \mathcal{K}_r \setminus \{\iota_r\} \cap I_{top,r}(x^{k'}) \cap B(x^{k'})$$

$\mathcal{I}_r^+$ contains the selected indices dominating ι_r in the current iteration while $\mathcal{I}_r^-$ contains the selected indices from $\mathcal{I}_r \setminus \{\iota_r\}$ currently not dominating ι_r. $\mathcal{K}_r^+$ and $\mathcal{K}_r^-$ are defined accordingly. Let us first state a simple lemma concerning vanishing dominating indices:

Lemma 4. *In the next iteration all indices from $\mathcal{I}_r^+$ will become bottom–only or all indices from $\mathcal{K}_r^+$ will become top–only.*
In particular, if $\mathcal{I}_r^+ \neq \emptyset$ and $\mathcal{K}_r^+ = \emptyset$ κ_r is selected and all indices from $\mathcal{I}_r^+$ will become bottom–only. The same holds for $\mathcal{K}_r^r \neq \emptyset$ and $\mathcal{I}_r^+ = \emptyset$ accordingly.

Proof. If both sets are empty there's nothing to show. Without loss of generality we assume $\mathcal{I}_r^+ \neq \emptyset$. In this case there have to be selected bottom candidates from $[i_r]$ as the indices are selected pairwise.

If $\mathcal{K}_r^+ = \emptyset$, by choice of k, the first selected bottom candidate has to be κ_r. As κ_r is a bottom candidate in the next iteration as well, the claim follows according to Lemma 3.
If $\mathcal{K}_r^+ \neq \emptyset$, the assumption that a pair $(i,j) \in \mathcal{I}_r^+ \times \mathcal{K}_r^+$ exists, such that i will be top and j will be bottom candidate in the next iteration, contradicts to Lemma 3. □

The next lemma will deal with indices currently non–dominating but, in the next iteration, eventually becoming dominating:

Lemma 5. *If $\mathcal{I}_r^- \neq \emptyset$ the following two conditions hold*[10]*: $|\mathcal{I}_r^-| < |\mathcal{I}_r^+|$ and $\mathcal{K}_r^- = \emptyset$. The same holds for $\mathcal{K}_r^-$ respectively.*

Proof. If $\mathcal{I}_r^- \neq \emptyset$ all selected top candidates dominate ι_r and $\mathcal{K}_r^-$ is therefore empty. As the indices are selected pairwise in every class, it holds that $2|\mathcal{I}_r^+| = |B(x^{k'}) \cap [i_r]| > 0$. In addition, $\kappa_r \in B(x^{k'}) \cap [i_r]$ and it therefore follows that at least one selected bottom candidate is not in $\mathcal{I}_r$ and thus $|\mathcal{I}_r^-| + 1 \leq |\mathcal{I}_r^+|$. □

To finalize the proof, we note that for at least one $r \in \mathcal{C}_>$ there have to be dominating indices, i.e. $\mathcal{I}_r^+ \neq \emptyset$ or $\mathcal{K}_r^+ \neq \emptyset$. Otherwise, by choice of k, (ι_r, κ_r) would be selected for some r.
Thus, according to Lemma 4, the number of vanishing dominating indices is strictly positive and, according to Lemma 5, the number of non–dominating indices, eventually becoming dominating, is strictly smaller. This proves condition 2) and, with condition 1), leads to a contradiction as mentioned above.

Thus the assumption that a limit point $\bar{x}$ is not optimal has to be wrong and the decomposition method, based on the generalized gradient selection, converges for problems decomposable by pairing. □

5 Final remarks and open problems

We have shown that the well–known method of selecting the working set according to the gradient of the objective function can be generalized to a larger class of convex quadratic optimization problems. The complexity of the given extension is equal to Joachims' selection algorithm except for an initialization overhead for the calculation of the classes of indices and the λ_i. Thus implementations like SVM^{light} [5] can easily be extended to solve such problems with little extra cost.

We would like to point out that the given selection algorithm and the extended convergence proof hold for most variants of SVM including the ν-SVM for which the proof of convergence of a decomposition method with a gradient selection strategy (e.g. [12]) has not been published yet.

It would be interesting to eliminate the restriction on the matrix Q (see footnote 8) and to extend the proof in [14] to working sets with more than two indices. A more interesting topic for future research would be the extension of known results concerning speed of convergences (e.g. [16], [17]) to the extended gradient selection approach proposed in this paper.

[10] We briefly note, that such candidates might only exist if ι_r is top–only.

Acknowledgments Thanks to Hans Simon for pointing me to a more elegant formulation of Definition 3 and to a simplification in the main convergence proof. Thanks to Dietrich Braess for pointing out the simpler formulation of the proof of Lemma 1.

References

1. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A Training Algorithm for Optimal Margin Classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory, ACM Press (1992) 144–153
2. Christianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. 5. edn. Cambridge University Press (2003)
3. Schölkopf, B., Smola, A.J.: Learning with Kernels. 2. edn. MIT Press (2002)
4. Osuna, E., Freund, R., Girosi, F.: An Improved Training Algorithm for Support Vector Machines. In Principe, J., Gile, L., Morgan, N., Wilson, E., eds.: Neural Networks for Signal Processing VII – Proceedings of the 1997 IEEE Workshop, New York, IEEE (1997) 276–285
5. Joachims, T.: 11. [18] 169–184
6. Platt, J.C.: 12. [18] 185–208
7. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation **13** (2001) 637–649
8. Lin, C.J.: On the Convergence of the Decomposition Method for Support Vector Machines. IEEE Transactions on Neural Networks **12** (2001) 1288–1298
9. Keerthi, S.S., Gilbert, E.G.: Convergence of a generalized SMO algorithm for SVM classifier design. Machine Learning **46** (2002) 351–360
10. Schölkopf, B., Smola, A.J., Williamson, R., Bartlett, P.: New Support Vector Algorithms. Neural Computation **12** (2000) 1207–1245
11. Chen, P.H., Lin, C.J., Schölkopf, B.: A Tutorial on ν–Support Vector Machines. (http://www.csie.ntu.edu.tw/∼cjlin/papers/nusvmtutorial.pdf)
12. Chang, C.C., Lin, C.C.: Training ν- Support Vector Classifiers: Theory and Algorithms. Neural Computation **10** (2001) 2119–2147
13. Simon, H.U., List, N.: A General Convergence Theorem for the Decomposition Method. In Shawe-Taylor, J., Singer, Y., eds.: Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004. Volume 3120/2004 of Lecture Notes in Computer Science., Heidelberg, Springer Verlag (2004) 363–377
14. Lin, C.J.: Asymptotic Convergence of an SMO Algorithm without any Assumptions. IEEE Transactions on Neural Networks **13** (2002) 248–250
15. Golub, G.H., Loan, C.F.: Matrix Computations. 3. edn. The John Hopkins University Press (1996)
16. Lin, C.C.: Linear Convergence of a Decomposition Method for Support Vector Machines. (http://www.csie.ntu.edu.tw/∼cjlin/papers/linearconv.pdf)
17. Hush, D., Scovel, C.: Polynomial-time Decomposition Algorithms for Support Vector Machines. Machine Learning **51** (2003) 51–71
18. Schölkopf, B., Burges, C.J.C., Smola, A.J., eds.: Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge, MA (1999)