LA-UR- 07-0621

Approved for public release; distribution is unlimited.

Title:	Gaps in Support Vector Optimization
Author(s):	Ingo Steinwart, z# 191174, CCS-3 Don Hush, z# 115295, CCS-3 Clint Scovel, z#097403, CCS-3 Nicolas List, Ruhr-University Bochum, Germany
Intended for:	20th Conference on Learning Theory San Diego, CA June 13-15, 2007

Los Alamos
NATIONAL LABORATORY
EST.1943

Form 836 (7/06)

Gaps in Support Vector Optimization

Nikolas List[1] (student author), Don Hush[2], Clint Scovel[2], Ingo Steinwart[2]

[1] Lehrstuhl Mathematik und Informatik, Ruhr-University Bochum, Germany
nlist@lmi.rub.de
[2] CCS-3, Informatics Group, Los Alamos National Laboratory,
Los Alamos, New Mexico, USA {dhush,jcs,ingo}@lanl.gov

Abstract. We show that the stopping criteria used in many support vector machine (SVM) algorithms working on the dual can be interpreted as primal optimality bounds which in turn are known to be important for the statistical analysis of SVMs. To this end we revisit the duality theory underlying the derivation of the dual and show that in many interesting cases primal optimality bounds are the same as known dual optimality bounds.

1 Introduction

Given a labeled training set $(x_1, y_1), \ldots, (x_\ell, y_\ell) \in X \times \{-1, 1\}$ on an input space X the standard $L1$-SVM for binary classification introduced by Vapnik et. al in [1] solves an optimization problem of the form

$$\begin{aligned} \arg\min_{(f,b,\xi)} \quad & \mathcal{R}(f, b, \xi) := \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C \sum_{i=1}^{\ell} \xi_i \\ \text{s.t.} \quad & \xi_i \geq 0 \ \text{ and } \ y_i(f(x_i) + b) \geq 1 - \xi_i \ \text{ f.a. } \ i = 1, \ldots, \ell \end{aligned} \tag{1}$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) of a kernel $k : X \times X \to \mathbb{R}$ and $C > 0$ is a free regularization parameter. Instead of solving this problem directly one usually applies standard Lagrange techniques to derive the following dual problem

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^\ell} \quad & W(\alpha) := \frac{1}{2}\langle K\alpha, \alpha \rangle - \alpha \cdot e \\ \text{s.t.} \quad & y \cdot \alpha = 0 \ \text{ and } \ 0 \leq \alpha_i \leq C \ \text{ f.a. } \ i = 1, \ldots, \ell \end{aligned} \tag{2}$$

where $K := (y_i y_j k(x_i, x_j))_{1 \leq i,j \leq \ell}$ is the so-called kernel matrix, $e \in \mathbb{R}^\ell$ is the all ones vector, and $y := (y_1, \ldots, y_\ell)$. Since the kernel is symmetric and positive semi-definite (2) is a standard convex quadratic optimization problem, which is more simple to solve as the primal problem (1).

The motivation for this procedure is usually given by the well known fact from Lagrangian Duality Theory, that for the special convex optimization problems (1) and (2) the strong duality assumption holds (see for example [2, Chapter 5]) in the sense that primal and dual optimal values coincide. Therefore starting from optimal dual solutions one can calculate optimal primal solutions using a simple transformation.

However, due to the usually large and dense kernel matrix it is not easy to solve (2) directly. To address this issue several techniques based on sequentially solving small subproblems have been proposed [14, 7, 15, 13, 5, 11, 21]. Of course, all these methods have in common that they only produce an *approximate* solution to the dual problem (2). However, recall that in order to establish guarantees on the generalization performance of (f, b, ξ) one needs to know that $\mathcal{R}(f, b, \xi)$ approximates the minimum of (1) up to some pre-defined $\varepsilon_P > 0$ (see e.g. [20]). But unfortunately, it is not obvious why the above transformation should produce ε_P-optimal primal points from ε_D-optimal dual points. Consequently, the usual statistical analysis of SVMs does not apply to the learning machines applied in practice. This lack of theoretical guarantees has first been addressed by [6] were the authors showed that ε_D-optimal dual points can be transformed to $O(\sqrt{\varepsilon_D})$-optimal primal points using specific transformations.

In this paper we will show, that certain dual optimality bounds transform directly to primal optimality bounds in the sense of $\varepsilon_P = \varepsilon_D$. Let us note, that there has already been a similar argumentation for the special case of $L1$-SVMs in [18, Sec. 10.1]. The authors there, however, ignore the influence of the offset parameter b which leads to ambiguous formulas in Proposition 10.1. Besides that the approach we describe here is far more general and promises to give a unified approach for analyzing approximate duality.

In addition, we will show, that the above dual optimality bounds coincide with the well known σ-gaps that are used to analyze the convergence behaviour of certain algorithms working on the dual problem (2). Because of this connection, the results of this paper make it possible to combine convergence rates for certain L1-SVM algorithms and oracle inequalities (see e.g. [20]) describing the statistical performance of the resulting classifier.

The rest of this work is organized as follows: In Section 2 we revisit duality theory and introduce certain gap functions. We then illustrate the theory for convex quadratic optimization problems. In Section 3 we apply our findings to L1-SVMs. In particular, we there consider σ-gaps and a stopping criterion for maximal violating pairs algorithms.

2 Gaps in constrained optimization

Let U be a nonempty set and let $\varphi : U \to \mathbb{R}$ and $c_i : U \to \mathbb{R}, i = 1, m$ be real valued functions. Let $c : U \to \mathbb{R}^m$ denote the function with components c_i. Consider the primal constrained optimization problem

$$\sup_{u \in U,\, c(u) \leq 0} \varphi(u) \tag{3}$$

The set $C := \{u \in U \mid c(u) \leq 0\}$ is called *feasibility region* of (3) and each $u \in C$ is called *a (primal) feasible point.* We define the Lagrangian $L : U \times \mathbb{R}^m \to \mathbb{R}$ associated with (3) by

$$L(u, \lambda) := \varphi(u) - \lambda \cdot c(u) \tag{4}$$

and write $(\mathbb{R}^+)^m := \{\lambda \in \mathbb{R}^m : \lambda \geq 0\}$. Note that although it is customary to define the Lagrangian to be ∞ when $\lambda \notin (\mathbb{R}^+)^m$ the definition (4) will be convenient when applying the subdifferential calculus. Now the *dual function* to (3) is defined by

$$\psi(\lambda) := \sup_{u \in U} L(u, \lambda) \tag{5}$$

and for fixed $\lambda \in \mathbb{R}^m$ the maximizers of $L(\cdot, \lambda)$ are denoted by

$$U_\lambda := \arg\max_{u \in U} L(u, \lambda) \ .$$

Note that for any $u \in U_\lambda$ we have $L(u, \lambda) = \psi(\lambda)$ and $L(u, \lambda) \geq L(u', \lambda)$ for all $u' \in U$. Since the latter equation amounts to one of the two inequalities defining a saddle point we refer to any $(u, \lambda) \in U_\lambda \times \mathbb{R}^m$ as a *semi-saddle*. The following lemma attributed to Uzawa from [12, Lemma 5.3.1] provides sufficient conditions for $u \in U$ to be an optimal primal solution:

Lemma 1. *Any $u \in U$ is a primal optimal point if there exists a $\lambda \geq 0$ such that $u \in U_\lambda$,*

$$c(u) \leq 0$$

and

$$\lambda_i c_i(u) = 0 \text{ for all } i = 1, \ldots, m \ .$$

The second condition is the feasibility of u the third one is called complementary slackness.

The next lemma shows that without any assumptions on φ and c the dual function has some remarkable properties.

Lemma 2. *The dual $\psi : \mathbb{R}^{m} \to \mathbb{R} \cup \{+\infty\}$ is convex and for $u \in U_\lambda$ we have $-c(u) \in \partial\psi(\lambda)$, where $\partial\psi(\lambda)$ denotes the subdifferential of ψ at λ.*

Proof. Since ψ is a pointwise supremum of affine functions it is convex. Moreover, $U \neq \emptyset$ implies $\psi(\lambda) = \sup_{u \in U} L(u, \lambda) > -\infty$ for all λ. Finally, for $\lambda' \in \mathbb{R}^{m}$ and $u \in U_\lambda$ we obtain

$$\psi(\lambda') \geq L(u, \lambda') = L(u, \lambda) + \lambda \cdot c(u) - \lambda' \cdot c(u) = \psi(\lambda) - c(u) \cdot (\lambda' - \lambda). \quad \square$$

Given the Lagrangian L of the problem (3) the corresponding *dual problem* is defined by

$$\inf_{\lambda \geq 0} \psi(\lambda). \tag{6}$$

Note that this a convex optimization problem by Lemma 2. We define the feasibility region of the dual to be $(\mathbb{R}^{+})^{m}$ and any $\lambda \geq 0$ is called a *(dual) feasible point*. Now, for any primal feasible u and any dual feasible λ we have $\varphi(u) \leq \varphi(u) - \lambda \cdot c(u) = L(u, \lambda) \leq \psi(\lambda)$ and hence we obtain

$$\psi(\lambda) - \varphi(u) \geq 0, \qquad u \in C,\ \lambda \geq 0\ . \tag{7}$$

Let us write

$$\varphi^{*} := \sup_{u \in U,\, c(u) \leq 0} \varphi(u)$$
$$\psi^{*} := \inf_{\lambda \geq 0} \psi(\lambda)$$

for the values of the primal and dual problem, respectively. Then $\psi^{*} - \varphi^{*}$ is the smallest possible gap in (7) and is called the *duality gap*. However, in this work we also need the gap for *arbitrary* primal-dual pairs, i.e. for not necessarily feasible $u \in U$ and $\lambda \in \mathbb{R}^{m}$ we consider

$$\operatorname{gap}(u, \lambda) := \psi(\lambda) - \varphi(u)\ . \tag{8}$$

The following lemma computes $\operatorname{gap}(u, \lambda)$ for semi-saddles.

Lemma 3. *For all semi-saddles $(u, \lambda) \in U_\lambda \times \mathbb{R}^{m}$ we have*

$$\operatorname{gap}(u, \lambda) = -\lambda \cdot c(u)\ .$$

Proof. We have $\operatorname{gap}(u, \lambda) = \psi(\lambda) - \varphi(u) = L(u, \lambda) - \varphi(u) = -\lambda \cdot c(u)$. $\square$

Now note that for $(u, \lambda) \in (U_\lambda \cap C) \times (\mathbb{R}^+)^m$ we have $c(u) \geq 0$ and $\lambda \geq 0$ and hence Lemma 3 shows that $\mathrm{gap}(u, \lambda) = 0$ is equivalent to the complementary slackness condition of Lemma 1. This fact leads to the following simple and natural optimality bounds:

Definition 1 (Forward Gap). *The forward gap of a feasible $u \in U$ is defined by*

$$\overrightarrow{G}(u) := \inf\{-\lambda \cdot c(u) \mid \lambda \geq 0,\, u \in U_\lambda\}\,. \tag{9}$$

Definition 2 (Backward Gap). *The backward gap of a feasible λ is defined by*

$$\overleftarrow{G}(\lambda) := \inf\big\{-\lambda \cdot c(u) \mid u \in U_\lambda, c(u) \leq 0\big\}\,. \tag{10}$$

Furthermore, for any feasible primal $u \in U$ we define its suboptimality to be

$$\Delta_P(u) := \varphi^* - \varphi(u)$$

and analogously for any feasible dual λ we define its suboptimality to be

$$\Delta_D(\lambda) := \psi(\lambda) - \psi^*\ .$$

The following simple lemma shows that the gaps control suboptimality:

Lemma 4. *Suppose that u and λ are feasible. Then we have*

$$\Delta_P(u) \leq \overrightarrow{G}(u) \quad \textit{and} \quad \Delta_D(\lambda) \leq \overleftarrow{G}(\lambda).$$

Proof. Using (7) we obtain $\Delta_P(u) = \varphi^* - \varphi(u) \leq \psi(\lambda') - \varphi(u) = \mathrm{gap}(u, \lambda')$ for all $\lambda' \geq 0$ satisfying $u \in U_{\lambda'}$. Similarly, for $u' \in U_\lambda \cap C$ we have $\Delta_D(\lambda) = \psi(\lambda) - \psi^* \leq \psi(\lambda) - \varphi(u') = \mathrm{gap}(u', \lambda)$. By Lemma 3 we then obtain the assertion. □

2.1 Forward gap and dual optimality

The forward gap is of particular utility if we have a closed formulation of $\{\lambda \geq 0 : u \in U_\lambda\}$. In this section we illustrate this for the forward gap of the *dual* problem. To that end we write (6) as a maximization problem by changing ψ to $-\psi$. The corresponding Lagrangian is then

$$L^D(\lambda, \mu) = -\psi(\lambda) + \mu \cdot \lambda\ .$$

Since ψ is convex we observe that $\lambda \in U_\mu := \mathrm{argmax}_{\lambda' \in \mathbb{R}^m} L^D(\lambda', \mu)$ if and only if $0 \in \partial_\lambda\big(-L^D(\lambda, \mu)\big) = \partial\psi(\lambda) - \mu$ which occurs if and only if

$\mu \in \partial\psi(\lambda)$. In other words we have $U_\mu = \{\lambda \in \mathbb{R}^m : \mu \in \partial\psi(\lambda)\}$. Since this implies

$$\{\mu \geq 0 \,|\, \lambda \in U_\mu\} = \partial\psi(\lambda) \cap (\mathbb{R}^+)^m$$

we see that the forward gap of (6) can be computed by

$$\overrightarrow{G}(\lambda) = \inf \{\mu \cdot \lambda \mid \mu \in \partial\psi(\lambda),\ \mu \geq 0\} \tag{11}$$

The following two results establish important properties of (11).

Lemma 5. *Given a feasible $\lambda \geq 0$. The minimum value $\overrightarrow{G}(\lambda)$ in (11) is finite and attained.*

Proof. The objective function $\lambda \cdot \mu$ and the constraint set $\{\mu \geq 0 \,|\, \lambda \in U_\mu\}$ have no direction of recession in common. Moreover $\{\mu \geq 0 \,|\, \lambda \in U_\mu\} = \partial\psi(\lambda) \cap (\mathbb{R}^+)^m$ is closed and convex and hence we obtain the assertion by [16, Theorem 27.3]. □

Theorem 1. *If $\lambda \geq 0$ satisfies $\overrightarrow{G}(\lambda) = 0$, then λ is optimal for (6). On the other hand if $\lambda \geq 0$ is optimal for (6) and $\operatorname{ri}(\operatorname{dom}\psi) \cap \operatorname{ri}((\mathbb{R}^+)^m) \neq \emptyset$ then $\overrightarrow{G}(\lambda) = 0$, where $\operatorname{ri} A$ denotes the relative interior of a set A.*

Proof. The first assertion follows directly from Lemma 4. For the second suppose that $\lambda \geq 0$ is optimal for (6). We write (6) as an unconstrained maximization of the function $-\psi(\lambda) - \mathbf{1}_{(\mathbb{R}^+)^m}(\lambda)$ where we note that for $\lambda \geq 0$ we have $\partial \mathbf{1}_{(\mathbb{R}^+)^m}(\lambda) = \{\mu \leq 0 \mid \lambda_i > 0 \Rightarrow \mu_i = 0\}$. Since $\lambda \geq 0$ is optimal it follows that $0 \in \partial(\psi(\lambda) + \mathbf{1}_{(\mathbb{R}^+)^m}(\lambda))$. However, by [16, Thm. 23.8] the assumptions imply that $\partial(\psi(\lambda) + \mathbf{1}_{(\mathbb{R}^+)^m}(\lambda)) = \partial\psi(\lambda) + \partial \mathbf{1}_{(\mathbb{R}^+)^m}(\lambda)$ so that we conclude that there exists a $\mu \in \partial\psi(\lambda)$ such that $\mu \geq 0$ and $\mu_i = 0$ for all i such that $\lambda_i > 0$. This implies $\overrightarrow{G}(\lambda) = 0$. □

2.2 Backward gap and the duality gap

Suppose we have a feasible dual variable λ and we ask for the best possible associated primal $u \in U_\lambda$. A simple calculation reveals, that for each feasible primal u and each feasible dual λ we have

$$\psi^* - \varphi^* \leq \psi(\lambda) - \varphi(u) = \operatorname{gap}(u, \lambda)$$

and therefore the duality gap is obviously a lower bound on the backward gap:

$$\psi^* - \varphi^* \leq \overleftarrow{G}(\lambda) \ .$$

To calculate the backward gap of a given λ recall that Lemma 2 implies that $\{-c(u) \mid u \in U_\lambda\} \subseteq \partial\psi(\lambda)$. Since $\partial\psi(\lambda)$ is convex it then follows that

$$C_\lambda := \{c(u) \mid u \in U_\lambda\} \ .$$

satisfies $-\operatorname{co} C_\lambda \subseteq \partial\psi(\lambda)$. The reverse inclusion will prove to be extremely useful so we recall the following definition from [3, Def. 2.3.1]:

Definition 3. *We say the* filling property *holds for λ, iff*

$$-\operatorname{co} C_\lambda = \partial\psi(\lambda) \ . \tag{12}$$

If in addition C_λ is convex we say, that the strict filling property *holds for λ.*

We will present some conditions under which the strict filling property holds in Section 2.4. We end this section by the following theorem which shows the importance of the (strict) filling property for the connection between the primal and dual problems. Since this theorem is not needed in the following we omit its elementary proof.

Theorem 2. *Assume the filling property holds for a given $\lambda \geq 0$. Then λ is an optimal dual solution iff there exist $s \leq m+1$ feasible primal points $u_1, \ldots, u_s \in U_\lambda$ and $\alpha_1, \ldots, \alpha_s \geq 0$ such that $\sum_{r=1}^s \alpha_r = 1$,*

$$\sum_{r=1}^{s} \alpha_r c(u_r) \leq 0, \quad \text{and} \quad \lambda_i \sum_{r=1}^{s} \alpha_r c_i(u_r) = 0 \quad \text{for all } i = 1, \ldots, m \ .$$

Moreover if the strict filling property holds for an optimal λ, then the duality gap is 0 and the solutions of the primal problem are given by the feasible $u \in U_\lambda$, for which $\operatorname{gap}(u, \lambda) = 0$. Since $(u, \lambda) \in U_\lambda \times (\mathbb{R}^+)^m$ the latter is equivalent to complementary slackness.

2.3 Relation between the gaps

Given a dual feasible point for which only approximate optimality can be guaranteed, our main question in this work is how this can be translated into approximate optimality guarantees for "associated" primal points. Fortunately, using forward and backward gaps as optimality bounds, the answer is quite simple, as the following theorem shows:

Theorem 3. *Let $\lambda \geq 0$ be a dual point for which the strict filling property holds. Then we have*

$$\overleftarrow{G}(\lambda) = \overrightarrow{G}(\lambda) \ .$$

In addition there exists a feasible $\hat{u} \in U_\lambda$ such that $-\lambda \cdot c(\hat{u}) = \overrightarrow{G}(\lambda)$. Moreover, $\hat{u}$ is an optimal solution of

$$\sup \{\varphi(u) \ \mid u \in U_\lambda, c(u) \leq 0\} .$$

Proof. Since the strict filling property implies that the infima in (10) and (11) range over the same set, we obtain equality of the gaps. Lemma 5 and the strict filling property then imply, that there exists feasible $\hat{u} \in U_\lambda$ such that $\overrightarrow{G}(\lambda) = -\lambda \cdot c(\hat{u})$. Moreover, for $(u, \lambda) \in U_\lambda \times \mathbb{R}^m$ Lemma 3 shows $\varphi(u) - \lambda \cdot c(u) = \psi(\lambda)$. Consequently, we see that for fixed $\lambda \geq 0$ maximizing φ is equivalent to minimizing $-\lambda \cdot c(\cdot)$ and therefore $\varphi(\hat{u})$ is also the maximal value φ attains on $\{u \in U_\lambda \mid c(u) \leq 0\}$. □

2.4 Sufficient Conditions for Filling

We now show that for concave quadratic optimization problems the strict filling property holds for any feasible dual point in the effective domain of the dual function (see [19] for more general settings). To that end let U be a Hilbert space, $w \in U$, $d \in \mathbb{R}^m$, $Q : U \to U$ be a nonnegative selfadjoint operator such that $Q : \ker(Q)^\perp \to \ker(Q)^\perp$ has a continuous inverse Q^{-1} and $A : U \to \mathbb{R}^m$ be continuous and linear. Then the convex quadratic problem

$$\sup_{\substack{u \in U \\ Au - d \leq 0}} \ -\frac{1}{2}\langle Qu, u\rangle + \langle w, u\rangle \tag{13}$$

is of the form (3) for $\varphi(u) := -\frac{1}{2}\langle Qu, u\rangle + \langle w, u\rangle$ and $c(u) := Au - d$. The next lemma, which includes the linear programming case, shows that the strict filling property holds:

Lemma 6. *Consider the convex quadratic programming problem (13). Then the strict filling property holds for all λ in the domain of the Lagrangian dual criterion function.*

Proof. The associated Lagrangian is $L(u, \lambda) = -\frac{1}{2}\langle Qu, u\rangle + \langle w, u\rangle - \lambda \cdot (Au - d) = -\frac{1}{2}\langle Qu, u\rangle + \langle w - A^*\lambda, u\rangle + \lambda \cdot d$ and its dual criterion function is defined by (5). If $w - A^*\lambda$ is not orthogonal to $\ker Q$ then it is easy to see that $\psi(\lambda) = \infty$. Now suppose that $w - A^*\lambda \in (\ker Q)^\perp = \operatorname{img} Q$.

Then we can solve $0 = \partial_u L(u, \lambda) = -Qu + w - A^*\lambda$ for u and hence we obtain

$$\begin{aligned} U_\lambda &= Q^{-1}(w - A^*\lambda) + \ker Q\,, \\ \psi(\lambda) &= \frac{1}{2}\langle Q^{-1}(w - A^*\lambda), w - A^*\lambda\rangle + \lambda \cdot d \\ \operatorname{dom}\psi &= \{\lambda \in \mathbb{R}^m \mid w - A^*\lambda \in \operatorname{img} Q\}\,. \end{aligned} \tag{14}$$

The latter formula for $\operatorname{dom}\psi$ implies

$$\partial\psi(\lambda) = AQ^{-1}(A^*\lambda - w) + d + \{\mu \mid A^*\mu \in \operatorname{img} Q\}^\perp$$

for all $\lambda \in \operatorname{dom}\psi$. Moreover, for $\lambda \in \operatorname{dom}\psi$ we also obtain

$$\begin{aligned} -C_\lambda &:= \{\, -c(u) \mid u \in U_\lambda \,\} \\ &= \{\, d - Au \mid u \in Q^{-1}(w - A^*\lambda) + \ker Q\} \\ &= d + AQ^{-1}(A^*\lambda - w) + A\ker Q\,. \end{aligned}$$

From Lemma 2 it suffices to show that $(A\ker Q)^\perp \subset \{\mu | A^*\mu \in \operatorname{img} Q\}$ to complete the proof. To that end suppose that $\mu \perp A\ker Q$. Then we have $\langle A^*\mu, z\rangle = \langle \mu, Az\rangle = 0$ for all $z \in \ker Q$ which implies $A^*\mu \in \operatorname{img} Q$. □

Let us denote the gradient of the dual criterion function (14) restricted to its domain by

$$\nabla\psi(\lambda) := AQ^{-1}(A^*\lambda - w) + d. \tag{15}$$

Using this notation the following corollary follows immediately from (15), the definition of the backward-gap and Theorem 3:

Corollary 1. *Given a dual feasible point $\lambda \in \operatorname{dom}\psi$, $\lambda \geq 0$, we have*

$$G_{QP}(\lambda) := \overleftarrow{G}(\lambda) = \inf_{z \in \ker Q} \{\lambda \cdot (\nabla\psi(\lambda) - Az) \mid \nabla\psi(\lambda) - Az \geq 0\}\,. \tag{16}$$

3 Applications to SVM optimization

In this section we we apply our results to SVMs. We begin by showing, that in this case (16) is a generalization of the well known σ-gap which has been used in [5, 11] both as stopping criterion for the dual problem and as an important quantity in the construction of algorithms which possess convergence rates. We then calculate the forward-gap for $L1$-SVMs in Subsection 3.2. Finally, in Section 3.3 we show that the stopping criteria used in MVP dual algorithms can directly be derived from this gap leading to primal optimality guarantees.

3.1 The σ-gap

Let λ^* denote an optimal solution to the dual problem (6). From the convexity of ψ it then follows that $\psi(\lambda) - \psi(\lambda^*) \leq \partial\psi(\lambda) \cdot (\lambda - \lambda^*)$. Consequently $\sigma(\lambda) := \sup\{\partial\psi(\lambda) \cdot (\lambda - \acute{\lambda}) \mid \acute{\lambda} \in \operatorname{dom}\psi, \acute{\lambda} \geq 0\}$ satisfies $\psi(\lambda) - \psi(\lambda^*) \leq \sigma(\lambda)$ and hence σ can be used as a stopping criteria for the dual. For quadratic convex programs the σ-gap amounts to that defined in [11], namely

$$\sigma(\lambda) = \sup\{\nabla\psi(\lambda) \cdot (\lambda - \mu) \mid \mu \geq 0,\ w - A^*\mu \perp \ker Q\}\ . \tag{17}$$

It was shown in [5] for $L1$-SVMs that iterative schemes which choose a successor λ_{n+1} of λ_n that satisfies $\partial\psi(\lambda_n) \cdot (\lambda_n - \lambda_{n+1}) \geq \tau\sigma(\lambda_n)$ converge to optimal with a rate depending upon τ. This result was improved and extended to general convex quadratic programming problems in [11]. Our next results relate the σ-gap to $G_{QP}(\lambda)$:

Lemma 7. *For any feasible $\lambda \in \operatorname{dom}\psi$ such that $G_{QP}(\lambda) < \infty$ we have*

$$\sigma(\lambda) = G_{QP}(\lambda)\,.$$

Proof. Lemma 6 ensures that the strict filling property holds for any dual point $\lambda \in \operatorname{dom}\psi = \{\lambda \mid w - A^*\lambda \perp \ker Q\}$. Let $P : \mathbb{R}^m \to A\ker Q$ denote the orthogonal projection onto $A\ker Q$. Since the duality gap for linear programming is zero (see for example [3][Cor. 2.3.6]) we have

$$\begin{aligned} G_{QP}(\lambda) &= \inf\{\lambda \cdot (\nabla\psi(\lambda) - Az) \mid z \in \ker Q,\ \nabla\psi(\lambda) - Az \geq 0\} \\ &= -\sup\{\lambda \cdot \eta \mid \eta \in \mathbb{R}^m,\ P\eta = 0,\ \eta \leq \nabla\psi(\lambda)\} + \lambda \cdot \nabla\psi(\lambda) \\ &= -\inf\{\mu \cdot \nabla\psi(\lambda) \mid \mu \geq 0,\ \nu \in \mathbb{R}^m,\ \mu + P\nu = \lambda\} + \lambda \cdot \nabla\psi(\lambda)\,. \end{aligned}$$

Since $(\lambda - \mu) = P\nu$ is equivalent $w - A^*\mu \perp \ker Q$ the right hand is equivalent to the σ-gap defined in (17) and the claim follows. □

The next corollary follows directly from Theorem 3 and Lemma 7:

Corollary 2. *Let λ be feasible such that $w - A^*\lambda \perp \ker Q$ and $\sigma(\lambda) < \infty$. Let $\hat{z}$ optimize the gap $G_{QP}(\lambda)$ defined in (16). Then $\hat{u} := w - A^*\lambda + \hat{z}$ is a $\sigma(\lambda)$-optimal solution of the primal problem, i.e*

$$\Delta_P(\hat{u}) \leq \sigma(\lambda)\ .$$

3.2 L1-SVMs

To represent the $L1$-SVM optimization problem (1) as a quadratic programming problem (13) we write $U := \mathcal{H} \times \mathbb{R} \times \mathbb{R}^\ell$ where $\mathcal{H}$ is the RKHS associated with a kernel k. Recall that the canonical feature map $\Phi : X \to \mathcal{H}$ is given by $\Phi(x) = k(x, \cdot)$, $x \in X$, and that the reproducing property states $f(x) = \langle f, \Phi(x) \rangle$, $f \in \mathcal{H}$, $x \in X$. We further write

$$Q := \begin{pmatrix} \mathrm{Id}_{\mathcal{H}} & 0 & \mathbf{0} \\ 0_{\mathcal{H}} & 0 & \mathbf{0} \\ 0_{\mathcal{H}} & 0 & \mathbf{0} \end{pmatrix}, \; A := \begin{pmatrix} -y_1\Phi(x_1) & -y_1 & -e_1 \\ \vdots & \vdots & \vdots \\ -y_\ell\Phi(x_\ell) & -y_\ell & -e_\ell \\ 0_{\mathcal{H}} & 0 & -e_1 \\ \vdots & \vdots & \vdots \\ 0_{\mathcal{H}} & 0 & -e_\ell \end{pmatrix}, \quad w := -C \begin{pmatrix} 0_{\mathcal{H}} \\ 0 \\ e \end{pmatrix}, \quad d := \begin{pmatrix} -e \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ denotes the zero vector in $\mathbb{R}^\ell$ and e denotes the vector of all 1's in $\mathbb{R}^\ell$. Let us solve (2) using Corollary 1. To that end let us write $\lambda = \binom{\alpha}{\beta} \in \mathbb{R}^{2\ell}$. Then elementary calculations show that the condition $w - A^*\lambda \perp \ker Q$ amounts to

$$y \cdot \alpha = 0 \;\text{ and }\; \alpha + \beta = Ce. \tag{18}$$

For feasible λ satisfying (18) elementary calculations show that

$$\nabla\psi(\lambda) = \begin{bmatrix} \sum_{i=1}^\ell \alpha_i y_1 y_i k(x_i, x_1) - 1 \\ \vdots \\ \sum_{i=1}^\ell \alpha_i y_j y_i k(x_i, x_j) - 1 \\ \vdots \\ \sum_{i=1}^\ell \alpha_i y_\ell y_i k(x_i, x_\ell) - 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \nabla W(\alpha) \\ 0 \end{bmatrix} \in \mathbb{R}^{2\ell},$$

where $W(\alpha)$ is given as in (2). Since $\ker Q$ equals the last two components of $U = \mathcal{H} \times \mathbb{R} \times \mathbb{R}^\ell$ it follows from (16) that the gap $\overleftarrow{G}(\lambda)$ for the L1-SVM is

$$\begin{aligned} \inf_{(b,\xi)} \quad & \alpha \cdot (\nabla W(\alpha) + b \cdot y + \xi) + \beta \cdot \xi \\ \text{s.t.} \quad & \nabla W(\alpha) + b \cdot y + \xi \geq 0, \; \xi \geq 0 \,. \end{aligned}$$

For feasible $\lambda \in \operatorname{dom}\psi$ we have $\alpha_i \geq 0$ and $\beta = C - \alpha_i \geq 0$. Therefore the infimum above for fixed b is obtained by setting each $\xi_i = [-\nabla W(\alpha)_i - by_i]^+$. If we use the equality $\nu = [\nu]^+ - [-\nu]^+$ where $[\nu]^+ := \max(0, \nu)$ we conclude that $\overleftarrow{G}(\lambda) = G(\alpha)$ where

$$G(\alpha) := \inf_{-b \in \mathbb{R}} \left(\sum_{i=1}^{\ell} \alpha_i \left[\nabla W(\alpha)_i - by_i\right]^+ + (C - \alpha_i)\left[by_i - \nabla W(\alpha)_i\right]^+ \right). \tag{19}$$

Note, that $G(\alpha)$ can be computed solely in terms of the dual problem since it is the forward gap on the dual. The connection to the backward gap given in Theorem 3 however leads to a nice consequence:

Corollary 3. *Let $\varepsilon_D > 0$, let $0 \leq \alpha \leq C \cdot \mathbf{e}$ be a vector satisfying $y^\top \alpha = 0$, and let $\hat{b}$ be an optimal solution of (19). Assume that α is ε_D-optimal in the sense that $G(\alpha) = \overrightarrow{G}(\alpha) \leq \varepsilon_D$. Define $\hat{f} := \sum_{i=1}^{\ell} y_i \alpha_i \Phi(x_i)$ and $\hat{\xi}_i := [\hat{b}y_i - \nabla W(\alpha)_i]^+, i = 1, .., \ell$. Then $(\hat{f}, \hat{b}, \hat{\xi})$ is a ε_D-optimal solution of (1), i.e.*

$$\mathcal{R}(\hat{f}, \hat{b}, \hat{\xi}) - \mathcal{R}^* \leq G(\alpha) \ .$$

Recall that [4, Theorem 2] only showed that $(\hat{f}, \hat{b}, \hat{\xi})$ is a $\mathcal{O}(\sqrt{\varepsilon_D})$-optimal primal solution, and consequently the above corollary substantially improves this earlier result.

3.3 Optimality criteria and maximal violating pairs

The most popular SVM algorithms are maximum-violating pair algorithms (MVP), which are implemented for example in SVM^{light} and SMO-type algorithms. Often this selection strategy has been motivated directly from Karush-Kuhn-Tucker (KKT) conditions on the dual [8–10], but there has been no justification in terms of optimality guarantees. Let us first introduce some notation to be able to formulate the stopping criterion used in MVP algorithms. To that end recall the well known top-bottom candidate definition of Joachims and Lin [7, 9]:

$$\begin{aligned} \overline{I_{top}}(\alpha) &:= \{i \mid (\alpha_i < C, y_i = -1) \vee (\alpha_i > 0, y_i = +1)\} \\ \overline{I_{bot}}(\alpha) &:= \{i \mid (\alpha_i < C, y_i = 1) \vee (\alpha_i > 0, y_i = -1)\} \ . \end{aligned} \tag{20}$$

Any pair $(i, j) \in \overline{I_{top}}(\alpha) \times \overline{I_{bot}}(\alpha)$, such that $y_i \nabla W(\alpha)_i > y_j \nabla W(\alpha)_j$ is called a *violating pair*, since it forces at least one of the summands in (19)

corresponding to i or j to be non-zero for any choice of b. For the maximal violating pair define

$$\hat{t} := \max_{i \in \overline{I_{top}}(\alpha)} y_i \nabla W(\alpha)_i \quad \text{and} \quad \hat{b} := \min_{i \in \overline{I_{bot}}(\alpha)} y_i \nabla W(\alpha)_i \,.$$

It is well known, that whenever $\hat{t} \leq \hat{b}$ the dual variable α is optimal. This lead to the heuristic dual stopping criterion $\hat{t} - \hat{b} \leq \varepsilon$. We now show that our results do also provide primal optimality guarantees:

Lemma 8. *Given a final solution $\hat{\alpha}$ of a MVP-algorithm which terminated with accuracy ε, i.e. $\hat{t} - \hat{b} \leq \varepsilon$, then for any $b \in [\hat{b}, \hat{t}]$ the associated primal solution $(\hat{f}, b, \xi(b))$ defined by $\hat{f} := \sum_{i=1}^{\ell} \hat{\alpha}_i \Phi(x_i)$ and $\xi_i(b) := [b y_i - \nabla W(\alpha)_i]^+$ is $C\ell \cdot \varepsilon$ optimal, i.e.*

$$\mathcal{R}(\hat{f}, b, \xi(b)) - \mathcal{R}^* \leq C\ell \cdot \varepsilon \ .$$

Proof. Using the definition (20) the gap $G(\alpha)$ given in (19) can be computed by

$$\inf_{b \in \mathbb{R}} \Bigg(\sum_{\substack{i \in \overline{I_{top}}(\alpha) \\ y_i \nabla W(\alpha)_i > b}} \mu_i^+ \left[y_i \nabla W(\alpha)_i - b \right]^+ + \sum_{\substack{i \in \overline{I_{bot}}(\alpha) \\ y_i \nabla W(\alpha)_i < b}} \mu_i^- \left[b - y_i \nabla W(\alpha)_i \right]^+ \Bigg) \ , \tag{21}$$

where

$$\mu_i^+ = \begin{cases} \alpha_i & \text{if } y_i = +1 \\ C - \alpha_i & \text{else} \end{cases} \qquad \text{and} \qquad \mu_i^- = \begin{cases} C - \alpha_i & \text{if } y_i = +1 \\ \alpha_i & \text{else} \end{cases} \ .$$

Indeed note, that for any $i \in \overline{I_{top}}(\alpha)$ such that $y_i \nabla W(\alpha)_i \leq b$ we either have $i \in \overline{I_{bot}}(\alpha)$ too, and the contribution of index i is counted by the second sum, or i is a top-only candidate, i.e. $\alpha_i = 0$ and $y_i = -1$ or $\alpha_i = C$ and $y_i = 1$. In both cases the contribution of index i to (19), given by

$$\alpha_i \left[\nabla W(\alpha)_i - b y_i \right]^+ + (C - \alpha_i) \left[b y_i - \nabla W(\alpha) \right]^+$$

is zero. Similar arguments hold for bottom-candidates with $y_i \nabla W(\alpha)_i \geq b$.

We now try to bound the terms in (21) for arbitrary $b \in [\hat{b}, \hat{t}]$. Obviously we have

$$\begin{aligned} \left[y_i \nabla W(\alpha)_i - b \right]^+ &\leq \left[\hat{t} - b \right]^+ \leq \left[\hat{t} - \hat{b} \right]^+ && \text{for } i \in \overline{I_{top}}(\alpha) \quad \text{and} \\ \left[b - y_i \nabla W(\alpha)_i \right]^+ &\leq \left[b - \hat{b} \right]^+ \leq \left[\hat{t} - \hat{b} \right]^+ && \text{for } i \in \overline{I_{bot}}(\alpha) \,. \end{aligned}$$

Since the two sums in (21) range over disjoint index-sets and $\mu_i^{+/-} \leq C$ we conclude for any $b \in [\hat{b}, \hat{t}]$, that

$$G(\alpha) \leq C\ell \cdot \left[\hat{t} - \hat{b}\right]^+$$

and the claim follows from Corollary 3. □

Remark 1. If we count the number

$$d := \left| \left\{ i \;\middle|\; (i \in \overline{I_{top}}(\alpha) \wedge y_i \nabla W(\alpha)_i > \hat{b}) \vee (i \in \overline{I_{bot}}(\alpha) \wedge y_i \nabla W(\alpha)_i < \hat{t}) \right\} \right|$$

of indices which could indicate a violation if b is chosen in $[\hat{b}, \hat{t}]$. Then Lemma 8 can be improved so the right hand side is $Cd \cdot \varepsilon$.

4 Conclusion and Open problems

We have presented a general framework for deriving primal optimality guarantees from dual optimality bounds. We improve the results given in [4] insofar as we can directly transform dual in primal optimality guarantees without loosing by an order of $\mathcal{O}(\sqrt{\varepsilon})$. In addition our results are easily extensible to more general cases whenever the strict filling property can be proven. The main advantage in the framework of support vector optimization is however the fact, that important dual optimality bounds which are used in practice could directly be derived from the abstract forward-backward gaps. This closes a main gap in analysis of support vector machine *algorithms* since now optimality guarantees for approximately optimal dual points can be transfered to generalization guarantees for an associated classifier using the results from statistical learning theory.

We point out, that using results from [19], the generalization of tight relation of dual and primal problem even for approximately optimal points should be straight forward but was beyond this work. The question if the strict filling property is also a necessary condition for this relation is however an open question.

We leave it as an objective for future research, whether the deeper knowledge about the optimality bounds presented here can be used to extend known convergence guarantees from quadratic optimization to more general optimization problems.

References

1. B. E. Boser, I. M. Guyon, and V. N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 144–153. ACM Press, 1992.
2. N. Christianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 5th edition, 2003.
3. J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms II*. Springer Verlag, 1993.
4. D. Hush, P. Kelly, C. Scovel, and I. Steinwart. QP Algorithms with Guaranteed Accuracy and Run Time for Support Vector Machines. *Journal of Machine Learning Research*, 7:733–769, May 2006.
5. D. Hush and C. Scovel. Polynomial-time Decomposition Algorithms for Support Vector Machines. *Machine Learning*, 51:51–71, 2003.
6. D. Hush, C. Scovel, and I. Steinwart. Approximate duality. *Journal of Optimization Theory and Applications*, to appear.
7. T. Joachims. Making Large–Scale SVM Learning Practical. In Schölkopf et al. [17], chapter 11, pages 169–184.
8. S. S. Keerthi and E. G. Gilbert. Convergence of a Generalized SMO Algorithm for SVM Classifier Design. *Machine Learning*, 46:351–360, 2002.
9. C.-J. Lin. On the Convergence of the Decomposition Method for Support Vector Machines. *IEEE Transactions on Neural Networks*, 12:1288–1298, 2001.
10. N. List. Convergence of a generalized gradient selection approach for the decomposition method. In *Proceedings of the 15th International Conference on Algorithmic Learning Theory*, pages 338–349, 2004.
11. N. List and H. U. Simon. General Polynomial Time Decomposition Algorithms. In *Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005*, Lecture Notes in Computer Science, Heidelberg, 2005. Springer Verlag.
12. O. Mangasarian. *Nonlinear Programming*. SIAM, Philadelphia, PA, 1994.
13. O. L. Mangasarian and D. R. Musicant. Active Set Support Vector Machine Classification. In T. Lee, T. Diettrich, and V. Tresp, editors, *Neural Information Processing Systems (NIPS) 2000*, pages 577–583. MIT Press, 2001.
14. E. Osuna, R. Freund, and F. Girosi. An Improved Training Algorithm for Support Vector Machines. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, *Neural Networks for Signal Processing VII – Proceedings of the 1997 IEEE Workshop*, pages 276–285, New York, 1997.
15. J. C. Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schölkopf et al. [17], chapter 12, pages 185–208.
16. R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
17. B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods – Support Vector Learning*. MIT Press, Cambridge, MA, 1999.
18. B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, London, 2. edition, 2002.
19. V. Solov'ev. The subdifferential and the directional derivatives of the maximum of a family of convex functions. *Izvestiya: Mathematics*, 62(4):807–832, 1998.
20. I. Steinwart, D. Hush, and C. Scovel. An oracle inequality for clipped regularized risk minimizers. In *Advances in Neural Information Processing Systems 19*, 2007.
21. S. Vishwanathan, A. J. Smola, and M. N. Murty. Simplesvm. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

Abstract:
We show that the stopping
criteria used in many support vector machine (SVM) algorithms working on the dual can be interpreted
as primal optimality bounds which in turn are known to be important for the
statistical analysis of SVMs.
To this end
we revisit the duality theory underlying the
derivation of the dual and show that in many interesting cases
primal optimality bounds are the same as known dual optimality bounds.